Cell, Volume *134*

Supplemental Data

Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells

Alexander Marson, Stuart S. Levine, Megan F. Cole, Garrett M. Frampton, Tobias Brambrink, Sarah Johnstone, Matthew G. Guenther, Wendy K. Johnston, Marius Wernig, Jamie Newman, J. Mauro Calabrese, Lucas M. Dennis, Thomas L. Volkert, Sumeet Gupta, Jennifer Love, Nancy Hannett, Phillip A. Sharp, David P. Bartel, Rudolf Jaenisch, and Richard A. Young

Supplementary Experimental Procedures and Discussion

Index of Supplementary Tables

Index of Supplementary Figures

Index of Supplementary Files

The following files contain data formatted for upload into the UCSC genome browser (Kent et al., 2002). To upload the files, first copy the files onto a computer with internet access. Then use a web browser to go to http://genome.ucsc.edu/cgi-bin/hgCustom?hgsid=105256378 for mouse and http://genome.ucsc.edu/cgi-bin/hgCustom?hgsid=104842340 for human. In the "Paste URLs or Data" section, select "Browse…" on the right of the screen. Use the pop-up window to select the copied files, then select "Submit". The upload process may take some time.

mouse_miRNA_track.mm8.bed – Map of predicted miRNA genes in mouse. Transcripts with EST or gene evidence are shown as black lines. Presumed transcripts are shown as grey lines. Positions of the mature miRNAs are annotated as thicker lines.
human_miRNA_track.hg17.bed – Map of predicted miRNA genes in human. Transcripts with EST or gene evidence are shown as black lines. Presumed transcripts are shown as grey lines. Positions of the mature miRNAs are annotated as thicker lines.
mES_regulator_ChIPseq.mm8.WIG.gz – ChIP-seq data for Oct4, Sox2, Nanog and Tcf3 in mES cells. Top track for each data set illustrates the normalized number of reads assigned to each 25bp bin. Bars in the second track identify regions of the genome enriched at $p < 10^{-9}$.
mES_chromatin_ChIPseq.mm8.WIG.gz – ChIP-seq data for H3K4me3, H3K79me2, H3K36me3 and Suz12 in mES cells. Top track for each data set illustrates the normalized number of reads assigned to each 25bp bin. Bars in the second track identify regions of the genome enriched at $p < 10^{-9}$.

Supplemental References

Growth Conditions and Quality Control for Human Embryonic Stem Cells

Human embryonic stem (ES) cells were obtained from WiCell (Madison, WI; NIH Code WA09) and grown as described. Cell culture conditions and harvesting have been described previously (Boyer et al., 2005; Lee et al., 2006; Guenther et al., 2007). Quality control for the H9 cells included immunohistochemical analysis of pluripotency markers, alkaline phosphatase activity, teratoma formation, and formation of embryoid bodies and has been previously published as supplemental material(Boyer et al., 2005; Lee et al., 2006).

Growth Conditions for Murine Embryonic Stem Cells and Oct-4 Repressible ZHBTc4 Cells

V6.5 (C57BL/6-129) murine ES cells were grown under typical ES cell culture conditions on irradiated mouse embryonic fibroblasts (MEFs) as previously described (Boyer et al., 2006). Briefly, cells were grown on gelatinized tissue culture plates in Dulbecco's modified Eagle medium supplemented with 15% fetal bovine serum (characterized from Hyclone), 1000 U/ml leukemia inhibitory factor (LIF, Chemicon; ESGRO ESG1106), non-essential amino acids, L-glutamine, Penicillin/Streptomycin and ß-mercaptoethanol. Immunostaining was used to confirm expression of pluripotency markers, SSEA 1 (Developmental Studies Hybridoma Bank) and Oct4 (Santa Cruz, SC-5279). For location analysis, cells were grown for one passage off of MEFs, on gelatinized tissue-culture plates.

Cells harboring a doxycycline-repressible *Oct4* allele (ZHBTc4 cells, Niwa et al., 2000), a gift from A. Smith, were cultured under standard ES cell conditions, described above, on gelatin. Cultures were treated with 2µg/ml doxycycline (SIGMA, D-9891) for 12hrs or 24hrs. For immunostaining, ZHBTc4 cells were treated for 24hrs with doxycycline. Treated cells and non-treated controls were fixed in 4% paraformaldehyde and stained with primary antibodies against Oct4 (Santa Cruz, SC-5279) and Sox2 (R&D Systems, MAB2018,) and secondary Cy3-conjugated antibodies (**Figure S7d**). Oct4 shutdown was confirmed by reverse transcriptase (RT)-PCR of the Oct4 mRNA using the oligos: cacgagtggaaagcaactca and agatggtggtctggctgaac.

Antibodies

Oct4-bound genomic DNA was enriched from whole cell lysate using an epitope specific goat polyclonal antibody purchased from Santa Cruz (sc-8628) and compared to a reference whole cell extract (Boyer et al., 2005). Regions occupied with high confidence for this antibody identified by ChIP-seq in mES cells are listed in **Table S3** and by ChIP-chip on genome-wide tiling arrays in hES cells are on **Table S8**. Oct4 ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file: mES_regulator_ChIPseq.mm8.WIG.gz

Sox2-bound genomic DNA was enriched from whole cell lysate using an affinity purified goat polyclonal antibody purchased from R&D Systems (AF2018) and compared to a reference whole cell extract (Boyer et al., 2005). Regions occupied with high confidence for this antibody identified by ChIP-seq in mES cells are listed in **Table S3**. Sox2 ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file: mES_regulator_ChIPseq.mm8.WIG.gz

Nanog-bound genomic DNA was enriched from whole cell lysate using an affinity purified rabbit polyclonal antibody purchased from Bethyl Labs (bl1662) and compared to a reference whole cell extract (Boyer et al., 2005). Regions bound with high confidence for this antibody are listed in **Table S3**. Nanog ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file: mES_regulator_ChIPseq.mm8.WIG.gz

Tcf3-bound genomic DNA was enriched from whole cell lysate using an epitope specific goat polyclonal antibody purchased from Santa Cruz (sc-8635) and compared to a reference whole cell extract (Cole et al., 2008). Regions occupied with high confidence for this antibody identified by ChIP-seq in mES cells are listed in **Table S3**. Tcf3 ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file: mES_regulator_ChIPseq.mm8.WIG.gz

Suz12-bound genomic DNA was enriched from whole cell lysate using an affinity purified rabbit polyclonal antibody purchased from Abcam (AB12073) and compared to a reference whole cell extract (Lee et al., 2006). Regions bound with high confidence for this antibody are listed in **Table S10**. Suz12 ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file mES_chomatin_ChIPseq.mm8.WIG.gz

H3K4me3-modified nucleosomes were enriched from whole cell lysate using an epitope-specific rabbit polyclonal antibody purchased from Abcam (AB8580) (Santos-Rosa et al., 2002; Guenther et al., 2007). Samples were analyzed using ChIP-seq. Comparison of this data with ChIP-seq published previously (Mikkelsen et al., 2007) showed near identify in profile and bound regions (**Table S5**). H3K4me3 ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file: mES_chomatin_ChIPseq.mm8.WIG.gz

H3K79me2-modified nucleosomes were isolated from mES whole cell lysate using Abcam antibody AB3594 (Guenther et al., 2007). Chromatin immunoprecipitations against H3K36me3 were compared to reference WCE DNA obtained from mES cells. Samples were analyzed using ChIP-seq and were used for visual validation of predicted miRNA promoter association with mature miRNA sequences only **(Figure 2)**. H3K79me2 ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file: mES_chomatin_ChIPseq.mm8.WIG.gz

H3K36me3-modified nucleosomes were isolated from mES whole cell lysate using rabbit polyclonal antibody purchased from Abcam (AB9050) (Guenther et al., 2007). Chromatin immunoprecipitations against H3K36me3 were compared to reference WCE DNA obtained from mES cells. Samples were analyzed using ChIP-seq and were used for visual validation of predicted miRNA promoter association with mature miRNA sequences only **(Figure 2)**. H3K36me3 ChIP-seq data can be visualized on the UCSC browser by uploading supplemental file: mES_chomatin_ChIPseq.mm8.WIG.gz

Chromatin Immunoprecipitation
Protocols describing all materials and methods have been previously described (Lee et al. 2007) and can be downloaded from http://web.wi.mit.edu/young/hES_PRC.

Briefly, we performed independent immunoprecipitations for each analysis. Embryonic stem cells were grown to a final count of $5x10^7$ – $1x10^8$ cells for each location analysis experiment. Cells were chemically crosslinked by the addition of one-tenth volume of fresh 11% formaldehyde solution for 15 minutes at room temperature. Cells were rinsed twice with 1xPBS and harvested using a silicon scraper and flash frozen in liquid nitrogen. Cells were stored at –80°C prior to use.

Cells were resuspended, lysed in lysis buffers and sonicated to solubilize and shear crosslinked DNA. Sonication conditions vary depending on cells, culture conditions, crosslinking and equipment. We used a Misonix Sonicator 3000 and sonicated at approximately 28 watts for 10 x 30 second pulses (90 second pause between pulses). For ChIP of Oct4, Nanog, Tcf3 and Suz12 in murine ES cells, SDS was added to lysate after sonication to a final concentration of 0.1%. Samples were kept on ice at all times.

The resulting whole cell extract was incubated overnight at 4°C with 100 µl of Dynal Protein G magnetic beads that had been pre-incubated with approximately 10 µg of the appropriate antibody. Beads were washed 4-5 times with RIPA buffer and 1 time with TE containing 50 mM NaCl. For ChIP of Oct4, Nanog, Tcf3 and Suz12 in murine ES cells, the following 4 washes for 4 minutes each were used instead of RIPA buffer: 1X low salt (20mM Tris pH 8.1, 150mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), 1X high salt (20mM Tris pH 8.1, 500mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), 1X LiCl (10mM Tris pH 8.1, 250mM LiCl, 1mM EDTA, 1% deoxycholate, 1% NP-40), and 1X TE+ 50mM NaCl. Bound complexes were eluted from the beads by heating at 65°C with occasional vortexing and crosslinking was reversed by overnight incubation at 65°C. Whole cell extract DNA (reserved from the sonication step) was also treated for crosslink reversal.

ChIP-Seq Sample Preparation and Analysis

All protocols for Illumina/Solexa sequence preparation, sequencing and quality control are provided by Illumina (http://www.illumina.com/pages.ilmn?ID=203). A brief summary of the technique and minor protocol modifications are described below.

Sample Preparation

Immunoprecipitated (ChIP) DNA was prepared for sequencing according to a modified version of the Illumina/Solexa Genomic DNA protocol. Fragmented DNA was prepared for ligation of Solexa linkers by repairing the ends and adding a single adenine nucleotide overhang to allow for directional ligation. A 1:100 dilution of the Adaptor Oligo Mix (Illumina) was used in the ligation step. A subsequent PCR step with limited (18) amplification cycles added additional linker sequence to the fragments to prepare them for annealing to the Genome Analyzer flow-cell. After amplification, a narrow range of fragment sizes was selected by separation on a 2% agarose gel and excision of a band between 150-300 bp (representing shear fragments between 50 and 200nt in length and ~100bp of primer sequence). The DNA was purified from the agarose and diluted to 10 nM for loading on the flow cell.

Polony generation on Solexa Flow-Cells

The DNA library (2-4 pM) was applied to the flow-cell (8 samples per flow-cell) using the Cluster Station device from Illumina. The concentration of library applied to the flow-cell was calibrated such that polonies generated in the bridge amplification step originate from single strands of DNA. Multiple rounds of amplification reagents were flowed across the cell in the bridge amplification step to generate polonies of approximately 1,000 strands in 1µm diameter spots. Double stranded polonies were visually checked for density and morphology by staining with a 1:5000 dilution of SYBR Green I (Invitrogen) and visualizing with a microscope under fluorescent illumination. Validated flow-cells were stored at 4^{o}C until sequencing.

Sequencing

Flow-cells were removed from storage and subjected to linearization and annealing of sequencing primer on the Cluster Station. Primed flow-cells were loaded into the Illumina Genome Analyzer 1G. After the first base was incorporated in the Sequencing-by-Synthesis reaction the process was paused for a key quality control checkpoint. A small section of each lane was imaged and the average intensity value for all four bases was compared to minimum thresholds. Flow-cells with low first base intensities were re-primed and if signal was not recovered the flow-cell was aborted. Flow-cells with signal intensities meeting the minimum thresholds were resumed and sequenced for 26 cycles.

Solexa Data Analysis

Images acquired from the Illumina/Solexa sequencer were processed through the bundled Solexa image extraction pipeline which identified polony positions, performed base-calling and generated QC statistics. Sequences were aligned using the bundled ELAND software using murine genome NCBI Build 36 and 37 (UCSC mm8, mm9) as the reference genome. Alignments to build 37 were used for analysis of the *mmu-mir-290-295* cluster only as that cluster is not represented on build 36. Only sequences perfectly and uniquely mapping to the genome were used. A summary of the number of reads used is shown in **Table S1**.

The analysis methods used were derived from previously published methods (Johnson et al., 2007, Mikkelsen et al., 2007). Sequences from all lanes for each chromatin IP were combined, extended 200bp (maximum fragment length accounting for ~100bp of primer sequence), and allocated into 25 bp bins. Genomic bins containing statistically significant ChIP-seq enrichment were identified by comparison to a Poissonian background model, using a p-value threshold of 10^{-9}. A list of the minimum number of counts in a genomic bin required for each sample to meet this threshold are provided in **Table S1** . Additionally, we used an empirical background model obtained from identical Solexa sequencing of DNA from whole cell extract (WCE) from matched cell samples (>5X normalized enrichment across the entire region, see below). A summary of the bound regions and their relation to gene targets can be found in **Tables S2, S3, S5 and S10**.

The p-value threshold was selected to minimize the expected false-positive rate. Assuming background reads are spread randomly throughout the genome, the probability of observing a given number of counts can be modelled as a Poisson process where the expectation can be calculated as the number of mapped reads times the number of bins per read (8) divided by the total number of bins available (we assumed 50% as a very conservative estimate).

The Poissonian background model assumes a random distribution of background reads, however we have observed significant deviations from this expectation in ChIP-seq datasets. These non-random events can be detected as sites of enrichment using control IPs and create a significant number of false positive events for actual ChIP-seq experiments. To remove these regions, we compared genomic bins and regions that meet the statistical threshold for enrichment to an empirical distribution of reads obtained from Solexa sequencing of DNA from whole cell extract (WCE) from matched cell samples. We required that enriched regions have five-fold greater ChIP-seq density in the specific IP sample as compared with the non-specific WCE sample, normalized for the total number of reads. This served to filter out genomic regions that are biased to having a greater than expected background density of ChIP-seq reads. We observed that ~200-500 regions in the genome showed non-specific enrichment in these experiments.

Confirmation of ChIP-seq enrichment by RT-PCR

In order to ascertain the accuracy of our ChIP-seq data, we compared our results to RT(real time)-PCR studies done on Oct4 and Suz12 in mES cells. Previous studies had preformed RT-PCR data for Oct4 binding in mES cells (Loh et al., 2006). Loh and colleagues tested 71 binding events identified by ChIP-PET and confirmed by RT-PCR enrichment for 69 of these regions. Our Solexa data also identifies 69 of these 71 regions as enriched for Oct4 binding **(Figure S1a)**. Loh et al. also used RT-PCR to test enrichment at a set of 10 regions with low ChIP-PET signal (clusters of 2). They found 9 of these regions were not enriched by RT-PCR and our Solexa data found 8 of the 10 regions to not be enriched (the region enriched according the ChIP-PET data from Loh et al. was also identified as enriched in our Solexa data). Globally, 71 of 71 PCR fragments enriched by RT-PCR were confirmed by ChIP-seq while 34 of 39 sequences identified as not-enriched were found to be below our binding threshold. These discrepancies may represent false positives derived from sequence technologies (the majority of these sites were just below threshold in Loh et al.) or may represent false negatives in the RT-PCR experiment.

We also examined a quantitative PCR dataset for a Suz12 ChIP in mES cells (Boyer et al., 2005). This dataset identified 83 regions with Suz12 enrichment and 21 regions with no Suz12 enrichment as measured by RT-PCR. We examined sequence data for the regions amplified by RT-PCR (with the addition of 200 bases on either side of the PCR product to account for chromatin fragment size) to determine the level of agreement between our Suz12 ChIP-seq data and the RT-PCR data from Boyer et al. Of the 83 regions identified as enriched for Suz12 by RT-PCR we found that our Solexa data identified these regions as enriched for 63 of these (76%). Of the 21 regions identified as not enriched for Suz12 by RT-PCR we found that our Solexa data found 18 of these to be not enriched (86%). This analysis reveals that RT-PCR results do largely confirm our ChIP-seq data. Given that RT-PCR itself produces false positive and false negative results we further analyzed the data by incorporating the ChIP-chip data from Boyer et al. Of the 68 regions identified as enriched by both RT-PCR and ChIP-chip, 63 of these (93%) were enriched in our Solexa data. Of the 5 regions not directly enriched above threshold in the Solexa data, 3 of these were very near to regions identified as enriched from the Solexa data and the difference may be due, at least in part, to the much shorter fragment sizes used in ChIP-seq then in either ChIP-chip or RT-PCR. Of the 18 regions identified as not enriched by both RT-PCR and ChIP-chip, all of these (100%) were also determined to not be enriched according to our Solexa data. The results of this analysis are depicted in **Figure S1b.** The strong agreement between our ChIP-sq data and the ChIP-PCR data from Loh et al. and Boyer et al., despite differences in protocols and cell lines, strongly demonstrate the accuracy of our ChIP-seq results using the Solexa sequencing platform.

Comparison of ChIP-seq and ChIP-chip enrichment

We took advantage of our previous genome wide analyses to investigate the similarity between genome-wide location analyses done either using ChIP-seq on the Illumina-Solexa sequencer or using tiling arrays purchased from Agilent Technology. Recent work by our lab has identified the binding sites for Oct4, Nanog and Tcf3 in mES cells using ChIP-chip. These arrays contain over 6 million unique features tiling the non-repeat portion of the mouse genome at an average probe density of 1 oligo / 250 bp. Using these arrays, we identified 11,090 sites as significantly enriched for Oct4, 15,172 sites for Nanog and 13,348 sites for Tcf3. These sites showed extensive overlap with the regions identified by ChIP-seq (**Figure S3a**). Globally, we found that 61% (6,737) of ChIP-chip Oct4 bound sites that were also enriched in the ChIP-seq experiments (similar numbers were found for Nanog and Tcf3).

While the degree of overlap between ChIP-chip and ChIP-seq was relatively high, we wanted to understand the primary sources for the discrepancies between the two techniques. We first focused on the regions that were identified as enriched only in ChIP-seq. Direct examination of these regions revealed that the majority were in regions of the genome that were tiled poorly. In designing the 60bp oligonucleotides used on our genome-wide arrays, we required that each probe had a large degree of “uniqueness” (see supplemental for Boyer et al., 2005 and Lee et al., 2005). This was designed to minimize the degree of cross hybridization and required flexibility in the positioning of the oligonucleotides across the genome. ChIP-seq requires significantly less uniqueness to map reads to the genome and so should be able to detect binding across a much larger fraction of the genome (~70% as reported in Mikklesen et al., 2007). When we examined the Agilent probe density across the ChIP-seq enriched regions, we found a broad range of probe densities, with almost half of all high-confidence targets in regions with less then 3 probes per kb (**Figure S3b**). While the portions of the genome tiled at > 3 probes per kb had strong overlaps, enriched regions of the genome with lower probe densities were much more difficult to identify by ChIP-chip.

Examination of the regions found enriched only by ChIP-chip often found reduced levels of these sites in the ChIP-seq data. Overall, over 87% of ChIP-chip enriched regions showed at least moderate levels of enrichment within 1kb of the center of the Oct4 bound region (>= 4 overlapping mapped reads). Of the remaining regions, a large fraction showed much higher levels of enrichment using antibody against Oct4 then either Sox2 or Nanog, raising the concern that these sites may be false positives or represent transient binding sites that are highly variable by preperation (data not shown).

To further compare our ChIP-chip and ChIP-seq results, we asked how well the level of enrichment found on each platform correlated with the presence of the Oct4 target site. The 16bp Oct4 motif is present at almost all binding sites in our ChIP-seq dataset and has been discovered in previous ChIP-pet (paired-end ditag) experiments (Loh et al., 2006, see the section “DNA Motif Discovery and High-resolution Binding-Site Analysis” below for more detail). At each of the approximately six million probe positions from the ChIP-chip microarrays we determined whether there was a high-scoring motif within a 200 bp window (+/-100 bp) of that position **(Figure S3c)**. Approximately 9% of all probes genome wide are within 100bp of this motif. We then tabulated the ChIP-chip enrichment ratio and ChIP-seq density at each probe position. ~70% of the most enriched ChIP-seq positions were associated with Oct4 motifs. This is somewhat higher then the ~45% of the most enriched ChIP-chip positions. For both ChIP-chip and ChIP-seq, the enrichment level was strongly correlated with the presence of this DNA motif across the full range of enrichment levels.

In addition, we also compared our murine Suz12 ChIP-seq results to those from Suz12 ChIP-chip location analysis performed using promoter microarrays (Boyer et al, 2006). Here we limited our analysis to regions of the genome covered on the arrays. Of the 1,802 sites identified as enriched by ChIP-chip, 54% (978) were confirmed by ChIP-seq. Those sites that were not confirmed tended to be short with relatively few enriched microarray probes. When we focused on the larger domains of Suz12 that are found at developmental regulators (Lee 2005), we found that of the 697 Suz12 enriched regions from the ChIP-seq experiments greater than 1kb in size and covered by the microarrays, 89% (617) were also called enriched in the ChIP-chip experiment.

Finally, we compared our data to ChIP-seq data published by Mikkelsen et al. (2007). Previous studies on array platforms have often shown large differences between labs (see Boyer et al., 2005, Loh et al., 2006 for an example). We were curious if the protocol simplifications created by using a sequencing approach would allow for more reproducible data. Using the histone modification H3K4me3 as a test we found that there was an extremely high overlap between these two experiments. Of 19,632 H3K4me3 enriched regions that were identified, 96% (18,849) were overlapped in the comparison dataset.

Identifications of regions enriched for Oct4/Sox2/Nanog/Tcf3

The identification of enriched regions in ChIP-chip and ChIP-seq experiments is typically done using threshold for making a binary determination of enriched or not enriched. Unfortunately, there is not actually a clear delineation between truly bound and unbound regions. Instead, enrichment is a continuum and the threshold is set to minimize false positives (high-confidence sites). This typically requires that thresholds be set at a level that allows a high false-negative rate (~30% for ChIP-chip, Lee et al). When multiple factors are compared, focusing only on the intersection of the different data sets compounds this effect, leading to higher false negative rates and the loss of many critical target genes.

Oct4, Sox2, Nanog and Tcf3 co-occupy promoters throughout the genome (Cole, **Figure 1)** and cluster analysis of enriched sites reveals apparent co-enrichment for all 4 factors at >90% of sites (Frampton & Young, unpublished data). However, the overlap for any two factors at the cut-off for high-confidence enrichment is only about two thirds **(Figure S2, Tables S2 and S3)**. Therefore many of these sites must have enrichment that is below the high-confidence threshold for at least some of the participating factors. Variability in the enrichment observed for each factor at different binding sites is common in the data **(Figures 1b, 3, and S3)**.

To determine a threshold of binding for multiple factors, we used two complementary methods to examine high-confidence targets of the four regulators. First, the classes of genes enriched by different numbers of factors at high-confidence were compared to the known classes of targets based on gene ontology **(Figure S2b,** http://gostat.wehi.edu.au/cgi-bin/goStat.pl,Beissbarth and Speed, 2004**)**. The highest confidence targets (those with high levels of immuno-enrichment observed for all for factors) preferentially encoded factors involved in DNA binding, regulation of transcription and development as has been previously shown (Boyer et al., 2005). These gene ontology categories continued to be overrepresented among high-confidence targets of either 3 of the 4 factors or 2 of 4 the factors, albeit at lower levels, but were barely enriched among high confidence targets of only one factor.

As a second test, we examined how different numbers of overlapping high-confidence targets affected the overlap with our previous genome-wide studies using ChIP-chip. Because not all regions of the genome are tiled with equal density on the microarrays used for ChIP-chip, we first determined the minimum probe density required to confirm binding detected by ChIP-seq (**Figure S3**). At most genes with high probe density, the ChIP-seq and ChIP-chip data were very highly correlated. However, regions of the genome with microarray coverage of less than three probes per kilobase were generally unreliable in detecting this enrichment. These regions, which had low probe coverage on the microarrays, represent approximately 1/3 of all sites co-enriched for the four factors by ChIP-seq. In regions where probe density was greater than three probe per kilobase the fraction of ChIP-seq sites confirmed by ChIP-chip experiments increased with additional factors co-binding with a large fall off below 2 factors (data not shown). Based on these two analyses, we elected to choose targets occupied at high-confidence by 2 or more of the 4 factors tested for further analysis in this manuscript. **(Figures 1a and S2a (red line)).**

While a majority of the miRNA promoters identified as occupied by Oct4/Sox2/Nanog/Tcf3 are not occupied by all four factors at high-confidence, it is interesting to note that all of the miRNA genes that share highly similar seeds to miR-302 are occupied at high confidence by all four factors (miR-302 cluster, miR-290 cluster and miR-106a cluster), similar to the promoters of core transcriptional regulators of ES cells. By comparison, promoters also occupied by Suz12 almost never showed high-confidence binding for all four factors (**Table S4**, see *mmu-mir-9-2* in **Figure 3**). Similar effects were observed for protein-coding genes in mES cells (Lee et al., 2006). Whether this is caused by reduced epitope availability in PcG bound regions or reflects reduced protein binding is unclear.

DNA Motif Discovery and High-resolution Binding-Site Analysis

DNA motif discovery was performed on the genomic regions that were enriched at high-confidence by anti-Oct4 chromatin immunoprecipitation. In order to obtain maximum resolution, a modified version of the ChIP-seq read mapping algorithm was used. Genomic bins were reduced in size from 25 bp to 10 bp. Furthermore, a read extension that placed greater weight towards the middle of the 200 bp extension was used. This model placed 1/3 count in the 8 bins from 0-40 and 160-200 bp, 2/3 counts in the 8 bins from 40-80 and 120-160 bp and 1 count in the 4 bins from 80-120 bp. This allowed increased precision for determination of the peak of ChIP-seq density in each Oct4 bound region. One-hundred bp of genomic sequence, centered at the 500 largest peaks of Oct4 ChIP-seq density, were submitted to the motif disocvery tool MEME (Bailey and Elkan, 1995; Bailey et al., 2006) to search for over-represented DNA motifs. A single sixteen basepair motif was discovered by the MEME algorithm (**Table S4, Figure S4i**). This motif was significantly ($p<10^{-100}$) over-represented in the Oct4 bound input sequences and occurred in 445 of the 500 one-hundred bp sequences.

As a default, MEME uses the individual nucleotide frequencies within input sequences to model expected motif frequencies. This simple model might result discovery of motifs which are enriched because of non-random di-, tri-, etc. nucleotide frequencies. Consequently, three different sets of control sequences of identical length were used to ensure the specificity of the motif discovery results. First, the sequences immediately flanking each input sequence were used as control sequences. Second, randomly selected sequences having the same distribution of distances from transcription start sites as the Oct4 input sequences were used as control sequences. Third, sequences from completely random genomic regions were used as control sequences. Each of these sets of control seqeunces were also examined using MEME. For each of these controls, the motif discovered from actual Oct4 bound sequences was not identified in the control sequences.The motif discovery process was repeated using different numbers and lengths of sequences, but the same motif was discovered for a wide array of input sequences.

When motif discovery was repeated with the top 500 Sox2, Nanog, and Tcf3 binding peaks, the same motif was identified. Overall, the motif occurs within 100 bp of the peak of ChIP-seq density at more than 90% of the top regions enriched in each experiment, while occuring in the same span at 24-28% of control regions and within 25 bp of the ChIP-seq peak at more than 80% of regions versus 9-11% of control regions.

We next attempted to determine the precise sites in the genome bound by Oct4, Sox2, Nanog, and Tcf3 at basepair resolution using composite analysis of the bound regions for each factor. In particular, we examined if the different factors tended to associate with specific sequences within the assymetric DNA motif. A set of ~2,000 of the highest confidence bound regions was determined for each factor based on a count threshold two-fold higher than the threshold for high-confidence regions shown in **Table S1** (Poisson: $p < 10^{-9}$). Regions without a motif within 50bp of the peak of ChIP-seq enrichment, typically ~10% of regions, were removed from this analysis. The distance from the first base of the central motif in each bound region to the 5' end of all reads within 250bp was tabulated, seperating reads mapping to the same strand as the motif separate from reads mapping to the oppositie strand. The difference in ChIP-seq read frequency between reads mapping to the same strand as the motif and the reads mapping to the oppositite strand was calculated at every basepair within the 500 bp window **Figure S4**. We made the assumption that the precise peak of the ChIP-seq distribution was the point at which this strand bias was equal to zero.

To determine the precise position where the strand bias was equal to zero, we modelled the strand bias for each transcription factor with a simple function. We chose a function with 4 parameters (A,B, C, and M), one of which (M) was the point at which the curve crosses the x-axis.

(1)

$$f(x) = A \times \arctan\left(\frac{x-M}{B}\right) \times e^{-\left|\frac{x-M}{B}\right|^{C}}$$

Least squares curve fitting was performed using GNUplot (http://www.gnuplot.info/) with approximated initial conditions (A = -1000, B = 100, C = 2, M = 10). The variablity in M was detemined by bootsrapping (n=25) using a random set of half of the ChIP-seq reads in each dataset and is shown in **Figure S4.**

Identification of miRNA start sites in Human and Mouse Genomes

To better understand the regulation of miRNAs, we sought to identify the sites of transcription initiation for all miRNAs in both human and mouse, at least to low resolution (~1kb). Most methods used to identify promoters require active transcription of the miRNA and isolation of rare primary miRNA transcripts. We decided to use an approach based on *in vivo* chromatin signature of promoters. This approach has two principle advantages. First, the required data has been published by a variety of laboratories and is readily accessible and second, it does not require the productive transcription of the miRNA primary transcript.

Recent results using genome-wide location analysis of H3K4me3 indicate that between 60 and 80% of all protein-coding genes in any cell population have promoters enriched in methylated nucleosomes, even where the gene is not detected by typical transcription profiling (Guenther et al., 2007). Importantly, over 90% of the H3K4me3 enriched regions in these cells map to known or predicted promoters, suggesting that H3K4me3 can be used as a proxy for sites of active initiation. Our strategy to identify miRNA promoters, therefore, uses H3K4me3 enriched sites from as many sources as possible as a collection of promoters. In human, H3K4me3 sites were identified in ES cells (H9), hepatocytes, a pro-B cell line (REH cells) (Guenther et al., 2007) and T cells (Barski et al., 2007). Mouse H3K4me3 sites were identified from ES cells (V6.5), neural precursors, and embryonic fibroblasts (Mikkelsen et al., 2007). In total, we identified 34,793 high-confidence H3K4me3 enriched regions in human and 34,096 high-confidence regions enriched in mouse, collectively present at ~75% of all protein-coding genes.

The list of miRNAs identified in the miRNA atlas (Landgraf et al., 2007) were used as the basis for our identification. The total list consists of 496 miRNAs in human, 382 miRNAs in mouse. ~65% of the murine miRNAs can be found in both species. For each of these miRNAs, possible start sites were derived from both all H3K4me3 enriched regions within 250kb upstream of the miRNA as well as all known start sites for any miRNAs that were identified as being within known transcripts from RefSeq(Pruitt et al., 2005) Mammalian Gene Collection (MGC) (Gerhard et al., 2004) Ensembl (Hubbard et al., 2005), or University of California Santa Cruz (UCSC) Known Genes (genome.ucsc.edu)(Kent et al., 2002)for which EntrezGene (http://www.ncbi.nlm.nih.gov/entrez/) gene IDs had been generated. Where an annotated start site was found to overlap an H3K4me3 enriched region, the known start was used in place of the enriched region.

A scoring system was derived empirically to select the most likely start sites for each miRNA. Each possible site was given a bonus if it was either the start of a known transcript that spanned the miRNA or of an EST that spanned the miRNA. Scores were reduced if the H3K4me3 enriched region was assignable instead to a transcript or EST that did not overlap the miRNA. Additional positive scores were given to enriched sites within 5kb of the miRNA, while additional negative scores were given based on the number of intervening H3K4me3 sites between the test region and the miRNA. Finally, each enriched region was tested for conservation between human and mouse using the UCSC liftover program (Hinrichs et al., 2006). If two test regions overlapped, they were considered to be conserved (21%). In the cases where human and mouse disagreed on the quality of a site, if the site had an EST or gene overlapping the miRNA, that site was given a high score in both species. Alternatively, if one species had a non-overlapping site, that site was considered to be an unlikely promoter in both species. Finally, for miRNAs where a likely promoter was identified in only one species, we manually checked the homologous region of the other genome to search for regions enriched for H3K4me3-modified nucleosomes that may have fallen below the high-confidence threshold. Start sites were considered to be likely if the total score was $\geq$ 0 (**Figure S5 and S6**). In total, we identified likely start sites for ~85% of all miRNAs in both species (**Tables S6 and S7**). Predicted miRNA genes can be visualised on the

UCSC browser by uploading the supplemental files:mouse_miRNA_track.mm8.bed and human_miRNA_track.hg17.bed

Several lines of evidence suggest the high quality of these predictions. First, previous studies have found that miRNAs within 50kb of each other are likely to be co-regulated (Lagos-Quintana et al., 2001; Lau et al., 2001). While the nature of these clusters was not presupposed in our analysis, nearly all miRNAs within a cluster end up identifying the same promoter region (see **Figures 2, 3, and 5**). The only exceptions to this are found in the large clusters of repeat derived miRNAs found in chromosome 12 of mouse and chromosome 14 in human where a single H3K4me3 enriched region splits the clusters. Second, consistent with the frequent association of CpG islands with the transcriptional start sites for protein-coding genes, ~50% of the miRNA promoters identified here overlap CpG islands (**Tables S6 and S7**). Finally, for miRNAs that were active in ES cells, histone modifications associated with elongation were able to “connect” the mature miRNAs to the predicted transcription start site (**Figure 2**).

To further ascertain the accuracy of our promoter predictions, we compared our predicted start sites to those identified in recent studies. Predictions were tested against *mmu-mir-34b* / *mmu-mir-34c* (Corney et al., 2007), *hsa-mir-34a* (Chang et al., 2007) *mmu-mir-101a*, *mmu-mir-202*, *mmu-mir-22*, *mmu-mir-124a-1*, *mmu-mir-433* (Fukao et al., 2007), *mmu-mir-290-295*, *hsa-mir-371-373* (Houbivay et al., 2005) and *hsa-mir-17/18a/19a/20a/19b-1/92a-1* (O’Donnell et al., 2005). Additional miRNA promoters in these manuscripts were not predicted strongly by the above algorithm. For these 23 miRNAs, H3K4me3 sites were identified within 1kb of all but two of the sites. *mmu-mir-202* was predicted about 20kb upstream of the annotated start site, but may reflect an H3K4me3 site absent from the tissues sampled. *mmu-mir-433* is in the middle of a large cluster of miRNAs on mouse chromosome 12. The annotated TSS lies within the cluster between *mir-433* and *mir-431* suggesting the promoter may be incorrect. Overall, the accuracy of the promoter predictions is believed to be ~80% (8/10). Additional H3K4me3 data sets and EST data should allow for improved accuracy in predicting and validating these initiation sites.

Among the predicted miRNA promoters validated by previous studies is the start site for the *mir-290-295/371-373* polycistron (Houbaviy et al., 2005). This miRNA cluster includes the most abundant miRNAs in murine embryonic stem cells and the seed sequences specified by multiple mature miRNAs in this cluster form the basis of the network illustrations in **Figures 6 and 7**. In their study, Houbaviy and colleagues tested the ES cell specificity of this promoter using a heterologous reporter assay including a construct spanning from -2kb to +5kb surrounding the start site of *mir-290-295*. This region is notable in that, while it excludes the largest peaks of Oct4/Sox2/Nanog/Tcf3, it does contain a smaller (yet significantly enriched) region located over the promoter (small peak at the promoter in **Figure 3a**). This promoter proximal construct showed 5-10x higher maximal expression in

ES cells relative to more differentiated cells. Expression of this construct was dependent on a small portion of the construct that included the TATAA box and a proximal site of Oct4/Sox2/Nanog/Tcf3 occupancy.

ChIP-chip Sample Preperation and Analysis

Immunoprecipitated DNA and whole cell extract DNA were purified by treatment with RNAse A, proteinase K and multiple phenol:chloroform:isoamyl alcohol extractions. Purified DNA was blunted and ligated to linker and amplified using a two-stage PCR protocol. Amplified DNA was labeled and purified using Bioprime random primer labeling kits (Invitrogen): immunoenriched DNA was labeled with Cy5 fluorophore, whole cell extract DNA was labelled with Cy3 fluorophore.

Labeled DNA was mixed (~5 µg each of immunoenriched and whole cell extract DNA) and hybridized to arrays in Agilent hybridization chambers for up to 40 hours at 40°C. Arrays were then washed and scanned.

Slides were scanned using an Agilent DNA microarray scanner BA. PMT settings were set manually to normalize bulk signal in the Cy3 and Cy5 channel. For efficient batch processing of scans, we used Genepix (version 6.0) software. Scans were automatically aligned and then manually examined for abnormal features. Intensity data were then extracted in batch.

44k Human Whole Genome Array

The human promoter array was purchased from Agilent Technology (www.agilent.com). The array consists of 115 slides each containing ~44,000 60mer oligos designed to cover the non-repeat portion of the human genome. The design of these arrays are discussed in detail elsewhere (Lee et al., 2006).

Data Normalization and Analysis

We used GenePix software (Axon) to obtain background-subtracted intensity values for each fluorophore for every feature on the whole genome arrays. Among the Agilent controls is a set of negative control spots that contain 60-mer sequences that do not cross-hybridize to human genomic DNA. We calculated the median intensity of these negative control spots in each channel and then subtracted this number from the intensities of all other features.

To correct for different amounts of each sample of DNA hybridized to the chip, the negative control-subtracted median intensity value of control oligonucleotides from the Cy3-enriched DNA channel was then divided by the median of the control oligonucleotides from the Cy5-enriched DNA channel. This yielded a normalization factor that was applied to each intensity in the Cy5 DNA channel.

Next, we calculated the log of the ratio of intensity in the Cy3-enriched channel to intensity in the Cy5 channel for each probe and used a whole chip error model

(Hughes et al., 2000) to calculate confidence values for each spot on each array (single probe p-value). This error model functions by converting the intensity information in both channels to an X score which is dependent on both the absolute value of intensities and background noise in each channel using an f-score calculated as described (Boyer et al., 2005) for promoter regions or using a score of 0.3 for tiled arrays. When available, replicate data were combined, using the X scores and ratios of individual replicates to weight each replicate's contribution to a combined X score and ratio. The X scores for the combined replicate are assumed to be normally distributed which allows for calculation of a p-value for the enrichment ratio seen at each feature. P-values were also calculated based on a second model assuming that, for any range of signal intensities, IP:control ratios below 1 represent noise (as the immunoprecipitation should only result in enrichment of specific signals) and the distribution of noise among ratios above 1 is the reflection of the distribution of noise among ratios below 1.

High-Confidence Enrichment

To automatically determine bound regions in the datasets, we developed an algorithm to incorporate information from neighboring probes. For each 60-mer, we calculated the average X score of the 60-mer and its two immediate neighbours. If a feature was flagged as abnormal during scanning, we assumed it gave a neutral contribution to the average X score. Similarly, if an adjacent feature was beyond a reasonable distance from the probe (1000 bp), we assumed it gave a neutral contribution to the average X score. The distance threshold of 1000 bp was determined based on the maximum size of labelled DNA fragments put into the hybridization. Since the maximum fragment size was approximately 550 bp, we reasoned that probes separated by 1000 or more bp would not be able to contribute reliable information about a binding event halfway between them.

This set of averaged values gave us a new distribution that was subsequently used to calculate p-values of average X (probe set p-values). If the probe set p-value was less than 0.001, the three probes were marked as potentially bound.

As most probes were spaced within the resolution limit of chromatin immunoprecipitation, we next required that multiple probes in the probe set provide evidence of a binding event. Candidate bound probe sets were required to pass one of two additional filters: two of the three probes in a probe set must each have single probe p-values < 0.005 or the centre probe in the probe set has a single probe p-value < 0.001 and one of the flanking probes has a single point p-value < 0.1. These two filters cover situations where a binding event occurs midway between two probes and each weakly detects the event or where a binding event occurs very close to one probe and is very weakly detected by a neighboring probe. Individual probe sets that passed these criteria and were spaced closely together were collapsed into bound regions if the center probes of the probe sets were within 1000 bp of each other.

Comparing Enriched Regions to Known Genes and miRNAs

Enriched regions were compared relative to transcript start and stop coordinates of known genes compiled from four different databases: RefSeq (Pruitt et al., 2005), Mammalian Gene Collection (MGC) (Gerhard et al., 2004), Ensembl (Hubbard et al., 2005), and University of California Santa Cruz (UCSC) Known Genes (genome.ucsc.edu) (Kent et al., 2002). All human coordinate information was downloaded in January 2005 from the UCSC Genome Browser (hg17, NCBI build 35). Mouse data was downloaded in June of 2007 (mm8, NCBI build 36).
To convert bound transcription start sites to more useful gene names, we used conversion tables downloaded from UCSC and Ensembl to automatically assign EntrezGene (http://www.ncbi.nlm.nih.gov/entrez/) gene IDs and symbols to the RefSeq, MGC, Ensembl, UCSC Known Gene. Comparisons of Oct4, Sox2, Nanog, Tcf3, H3K4me3 and Suz12 to annotated regions of the genomes can be found in **Tables S3, S5, S8 and S10**

For miRNAs start sites, two separate windows were used to evaluate overlaps. For chromatin marks and non-sequence specific proteins, miRNA promoters were considered bound if they were within 1kb of an enriched sequence. For sequence specific factors such as Oct4, we used a more relaxed region of 8kb surrounding the promoter, consistent with previous work we have published (Boyer et al., 2005). A full list of the high confidence start sites bound to promoters can be found in **Tables S6 and S7.**

Growth Conditions for Neural Precursors, Mouse Embryonic Fibroblasts, Induced Pluripotent Stem Cells.

To generate neural precursor cells, ES cells were differentiated along the neural lineage using standard protocols. V6.5 ES cells were differentiated into neural progenitor cells (NPCs) through embryoid body formation for 4 days and selection in ITSFn media for 5–7 days, and maintained in FGF2 and EGF2 (R&D Systems) (Okabe et al., 1996).

Mouse embryonic fibroblasts were prepared from DR-4 strain mice as previously described (Tucker et al., 1997). Cells were cultured in Dulbecco’s modified Eagle medium supplemented with 10% cosmic calf serum, β-mercaptoethanol, non-essential amino acids, L-glutamine and penniclin/streptomycin.

Murine induced pluripotent stem cells (iPS) were generated as described in Wernig et al., 2007. iPS cells were cultured under the same conditions as mES cells.

Analysis of Mature miRNA Frequency by Solexa Sequencing

Short RNA cloning

A method of cloning the 18-30nt transcripts previously described (Lau et al., 2001) was modified to allow for Solexa (Illumina) sequencing (manuscript submitted). Single-stranded cDNA libraries of short transcripts were generated using size selected RNA. RNA extraction was performed using Trizol, followed by RNeasy purification (Qiagen).

5µg of RNA was size selected and gel purified. 3' Adaptor (pTCGTATGCCGTCTTCTGTTG [idT]) was ligated to RNA with T4 RNA ligase and also, separately with RNA Ligase (Rnl2(1-249)k->Q). Ligation products were gel purified and mixed. 5' adaptor (GUUCAGAGUUCUACAGUCCGACGAUC) was ligated with T4 RNA Ligase.

RT-PCR (Superscript II, Invitrogen) was performed with 5' primer (CAAGCAGAAGACGGCATA). Splicing of overlapping ends PCR (SOEPCR) was performed (Phusion, NEB) with 5' primer and 3' PCR primer (AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA), generating cDNA with extended 3' adaptor sequence. PCR product (40 µl) was denatured (85°C, 10 min, formamide loading dye), and the differently sized strands were purified on a 90% formamide, 8% acrylamide gel, yielding single-stranded DNA suitable Solexa sequencing.

The single-stranded DNA samples were resuspended in 10mM Tris (EB buffer)/0.1% Tween and then used as indicated in the standard Solexa sequencing protocol (Illumina). Each library was run on one lane of the Solexa sequencer.

Polony generation on Solexa Flow-Cells

The DNA library (2-4 pM) was applied to the flow-cell (8 samples per flow-cell) using the Cluster Station device from Illumina. The concentration of library applied to the flow-cell was calibrated such that polonies generated in the bridge amplification step originate from single strands of DNA. Multiple rounds of amplification reagents were flowed across the cell in the bridge amplification step to generate polonies of approximately 1,000 strands in 1µm diameter spots. Double stranded polonies were visually checked for density and morphology by staining with a 1:5000 dilution of SYBR Green I (Invitrogen) and visualizing with a microscope under fluorescent illumination. Validated flow-cells were stored at 4^{o}C until sequencing.

Sequencing and Analysis

Flow-cells were removed from storage and subjected to linearization and annealing of sequencing primer on the Cluster Station. Primed flow-cells were loaded into the Illumina Genome Analyzer 1G. After the first base was incorporated in the Sequencing-by-Synthesis reaction the process was paused for a key quality control checkpoint. A small section of each lane was imaged and the average intensity value for all four bases was compared to minimum thresholds. Flow-cells with low first base intensities were re-primed and if signal was not recovered the flow-cell was aborted. Flow-cells with signal intensities meeting the minimum thresholds were resumed and sequenced for 36 cycles. Images acquired from the Illumina/Solexa sequencer were processed through the bundled Solexa image extraction pipeline which identified polony positions, performed base-calling and generated QC statistics. Sequences were then assigned to a miRNA if they perfectly matched at least the first 20bp of the mature miRNA sequences downloaded from targetScan (http://www.targetscan.org/). Mature miRNA frequencies were then normalized to each other by determining the expected frequency in mapped reads/million. A full list of the miRNAs detected can be found in **Table S9**

In each cell type examined, a small subset of mature miRNA transcripts predominated (**Figure 4a**). Members of the miR-290-295 cluster, which encodes multiple miRNAs with the same seed sequence, constituted approximately two thirds of all mature miRNA transcripts in murine ES cells. Let-7 family members constituted roughly one quarter and one half of miRNAs in MEFs and NPCs, respectively. The miR-290-295 cluster, which dominated the expression profile of ES cells, but was scarce in both MEFs and NPCs, is occupied at its promoters by Oct4, Sox2, Nanog and Tcf3 (**Figure 3a**), consistent with the hypothesis that these factors are important for maintaining the expression of the miR-290-295 miRNA cluster in ES cells.

Read Normalization

Rigorous comparison of the time points in the Oct4 depletion experiment required normalization of the total number of miRNAs in each cell. Normalization was preformed using Northern analysis of the samples using multiple probes. First the total RNA in each lane was normalized using both a probe against tRNA and a probe againt the U6-RNA. This normalization was then applied to a series of individual mature miRNAs, including mmu-miR-16, -19b, -21, -291, and –293 (data not shown). On average, these samples showed the very similar results as when the Solexa small RNA reads were normalized to total miRNA content with a relatively large error of +/- 20%. We therefore assumed that the approximation that the total miRNA content of the cells was largely unchanged within 24 hours and normalized the final data presented in **Table S9** using the total numbers of miRNAs sequenced.

Northern blots of mature miRNAs

Northern blots were performed as in Houbaviy et al. (2003), except hybridization was carried out in Oligo Hyb (Ambion) at 37 C. Probe sequences were DNA oligos complementary to annotated miRBase miRNAs, Probes used in normalization were U6 snoRNA: 5'-GGGCCATGCTAATCTTCTCTGT-3' (Houbaviy et al. 2005) and tRNA-gln: 5'-TGGAGGTTCCACCGAGAT-3'.

miRNA Microarray Expression Analysis

RNA from murine embryonic stem cells (mES, V6.5), mouse embryonic fibroblasts (MEFs) and murine induced pluripotent (iPS) cells was extracted with RNeasy (Qiagen) reagents. 5 µg total RNA from treated and control samples were labeled with Hy3™ and Hy5™ fluorescent label, using the miRCURY™ LNA Array labeling kit (Exiqon, Denmark) following the procedure described by the manufacturer. The labeled samples were mixed pair-wise and hybridized to the miRNA arrays printed using miRCURY™ LNA oligoset version 8.1 (Exiqon, Denmark). Each miRNA was printed in duplicate, on codelink slides (GE), using GeneMachines Omnigrid 100. The hybridization was performed at 60C overnight using the Agilent Hybridization system - SurHyb, after which the slides were washed using the miRCURY™ LNA washing buffer kit (Exiqon, Denmark) following the procedure described by the manufacturer. The slides were then scanned using Axon 4000B scanner and the image analysis was performed using Genepix Pro 6.0.

Data from experiments with mES cells, MEFs and iPS cells were then combined. Median signal intensities for all microarray probes were background subtracted and tabulated. The data were then quantile normalized by assigning each probe the average signal intensity for all probes of the same intensity rank across the six experiments. Signal intensities were then floored at one unit and log transformed. Control probes were removed from further analysis.

We next looked to identify miRNA probes that were differentially enriched in mES cell and MEF samples and to compare them to data from iPS cells. Statistically significant differential expression for the two samples was calculated using the online NIA Array Analysis Tool (http://lgsun.grc.nia.nih.gov/ANOVA/). Probes from 3 MEF samples and 2 mES cell samples were tested for differential expression using the following settings:

Threshold z-value to remove outliers: 10,000
Error Model: Max(Average,Bayesian)
Error variance averaging window: 100
Proportion of highest error variances to be removed: 0.01
Bayesian degrees of freedom: 5
FDR threshold: 0.10

Of 1008 probes, 230 were determined to be differentially expressed between the MEF and mES samples.

For clustering and heat map display, expression data were Z-score normalized. Centroid linkage, Spearman rank correlation distance, and hierarchical clustering of genes and arrays was performed using Gene Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv). Heatmaps were generated using Java Treeview (http://jtreeview.sourceforge.net/) with color saturation at 0.6 standard deviations. Complete miRNA microarray expression data, differential expression analysis results, and clustergram data are provided (**Table S11**).

Tissue-Specificity of miRNAs

To determine the global tissue-specificity for miRNAs we used data from the recent publication of the miRNA atlas (Landgraf et al., 2007). Specificity scores were taken from Table S34 Node 0 from Landgraf et al. (2007). Of the 45 distinct mature miRNAs with specificity scores >1 that are not bound only by Oct4/Sox2/Nanog/Tcf3, 16 were identified as Suz12 targets. These 16 represent over 40% of the distinct mature miRNAs whose promoters are occupied by Suz12 ($p < 5x10^{-4}$ for specificity scores > 1.3)

Functional regulation of miRNAs by Oct4 and Tcf3 in mES cells

The occupancy of the promoters of the ES cell specific miRNAs by Oct4/Sox2/Nanog/Tcf3 implicates these critical factors as possible regulators of miRNA transcription. However, binding data alone cannot test the importance of the regulatory circuitry at any given gene. Gene regulation is complex and factors associated with the gene are likely to be redundant and may play different roles at different genes. In order to understand the role of members of the core regulatory circuitry at the miRNA genes, we have performed perturbation experiments on Oct4 and Tcf3.

miRNA regulation by Oct4

Oct4 is a critical regulator of ES cell pluripotency and disruption of Oct4 leads to rapid differentiation of the ES cells (Niwa et al., 2000). In order to understand the roll of Oct4 in regulating ES cell miRNAs, we utilized a system where both endogenous copies of Oct4 had been disrupted and Oct4 is supplied by an exogenous transgene under the control of a doxycyline regulated promoter (Niwa et al., 2000 and **Figure S7a**). Oct4 mRNA is rapidly lost in these cells upon doxycycline induction (**Figure S7b**). This system allowed us to examine the cells at early timepoints (12 and 24 hours) when the cells remain ES-like morphologically and still express Sox2 (**Figure S7c**), however we must assume this state is only transient as these cells are presumably in the early steps of differentiating.

We used two separate approaches to ascertain the role of Oct4 on miRNAs in these cells. First we directly tested the levels of selected primary transcripts of the miRNA promoters occupied by Oct4/Sox2/Nanog/Tcf3. Real time PCR primers were designed within the ~200nt immediately upstream of the tested miRNA hairpins or in the middle of *mir-290-295* polycistron, but outside of any hairpin regions (**Figure 3d** and see below for sequences). Primers were used to test samples at 0, 12 and 24 hours post-exposure to doxycycline (**Figure S8b**). Within 24 hours, the levels of all five miRNAs tested were reduced significantly ($p < 0.001$). The reduction of pri-*let-7g* is particularly interesting since the production of mature let-7g, found at high levels in differentiated cell types, is inhibited by the Oct4 target Lin28 (see **Figure 6**).

As a second approach, we examined the levels of the mature miRNAs by quantitative sequencing of short RNAs. This approach has the advantage of covering all miRNAS expressed in mES cells but is limited by the long half-life of mature miRNAs (believed to be as long as 24 hours D. Bartel, personal communication). Using the Solexa sequencer, we were able to identify 5.8 million reads across the six samples that matched known miRNAs. These samples were normalized based on Northern blot analysis (see above, data not shown). Using this methodology, we were able to observe changes in the relative abundance of Oct4/Sox2/Nanog/Tcf3 occupied miRNAs, with a general tendency for these miRNAs to decrease in relative abundance. Changes in miRNA expression were subtle, as expected, but some miRNAs were down regulated as much as 4-fold (**Figure S7d, Table S9**).

miRNA regulation by Tcf3

Tcf3 is a terminal component of the canonical Wnt pathway in ES cell has been integrated into the core circuitry regulating ES cells. Recent reports have indicated that Tcf3 depletion causes impaired differentiation in ES cells and upregulation of pluripotency genes, including Oct4, Sox2 and Nanog (Cole et al., Genes and Dev 2008; Tam et al., Stem Cells 2008; Yi et al., Stem Cells 2008). Genes encoding several key pluripotency factors were observed to increase in expression, albeit only mildly, but other genes decreased in expression or remained expressed at the same level. The different regulatory effects at different target genes may depend on the proteins associated with Tcf3 at the each promoter.

In order to understand the role of Tcf3 in regulating miRNAs in ES cells, we performed knockdown experiments on V6.5 ES cells using lentiviral shRNAs as described in Cole et al. Independent knockdowns resulted in over 70% depletion of endogenous Tcf3 (**Figure S8a**). Unlike Oct4, Tcf3 depleted ES cells are stable, allowing longer time points in these experiments. As described above, we investigated the effect of Tcf3 depletion on selected primary transcripts **(Figure S8b,c)**. Depletion of Tcf3 resulted in small but significant increases in the levels of the primary transcript for the *mir-302* and *mir-290-295* clusters by 72 hours post-infection. In addition, we see the primary transcript for the non-ES cell specific *mir-106-363* cluster decreases in steady state levels. Unlike the results we observe for Oct4, the primary transcript for *let-7g* does not show significant changes in expression.

Tcf3 knockdown

Tcf3 knockdown experiments were performed essentially as in Cole et al., 2008 with minor modifications. Lentivirus was produced according to Open Biosystems *Trans*-lentiviral shRNA Packaging System (TLP4614). The shRNA constructs targeting murine *Tcf3* were designed using an siRNA rules-based algorithm consisting of sequence, specificity, and position scoring for optimal hairpins that consist of a 21-base stem and a 6-base loop (RMM4534-NM-009332). A knockdown control virus targeting EGFP was produced from vector obtained from the RNAi Consortium. V6.5 mES cells were plated at ~30% confluence on the day of infection. Cells were seeded in mES media with 6 µg/mL polybrene (Sigma, H9268-10G) and *Tcf3* knockdown or control virus was immediately added. After 24 h, infection media was removed and replaced with mES media with 2 µg/mL Puromycin (Sigma, P8833). RNA was harvested at 72 h after infection.

Knockdown efficiency was measured using real-time PCR to measure levels of *Tcf3* mRNA (**Figure S8a**).

qRT-PCR of primary miRNAs

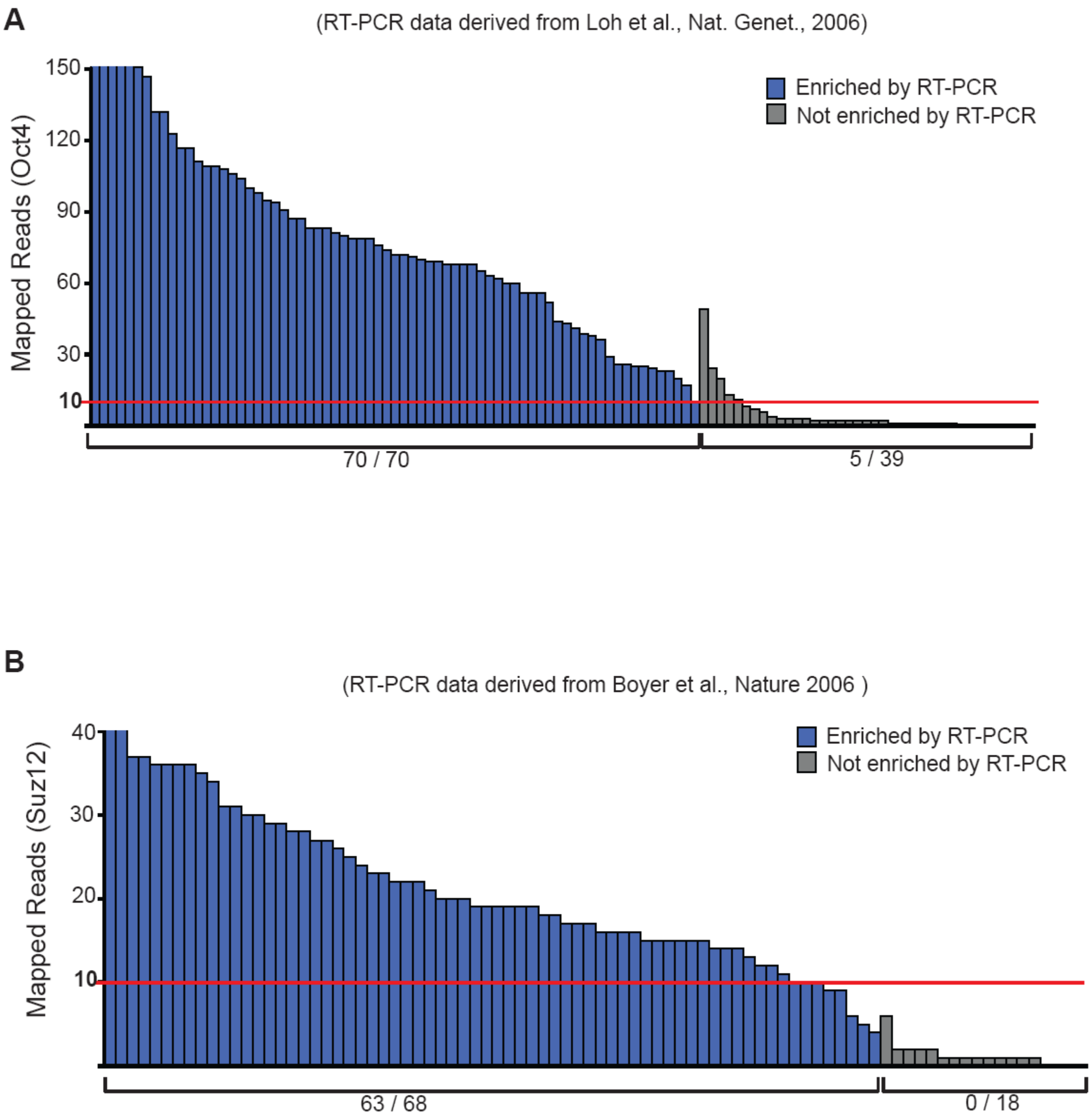
qPCR primers were designed using the standard specifications of PrimerExpress (Applied Biosystems) or Primer3 (Rozen and Skaletsky, 2000) for real time primer design. Primers were then used in SybrGreen quantitative PCR assays on the Appied Biosystems 7500 Real Time PCR system. Expression levels were calculated relative to Gapdh mRNA levels, which were quantified with in parallel with by Taqman analysis. The effects here were observed with high concentrations of cDNA template.

	Forward Primer	**Reverse Primer**
pri-*let-7g*	TGGCGGGTGCAGCTTT	AGGACGCACTTGAGGAAGGA
pri-*mir-302*-cluster	TTCACCCTCCGAGGACAGAA	ACAGACATAAGCTTTACCTCCTTTACCT
pri-*mir-290-295*	ACCTGGCTCCTAGCACAAACA	GGGCTATTGTAAAGCCAAAAGGTA
pri-*mir-106a-363*-cluster	AGGCATACTCCAGGAGTGTAACCT	CTTGTTTTAGAAATAGTAACTCACAGTTCACTT
pri-*mir-708*	TGCCTTCGTTCCCCTAACC	GGAGAAGTCAGGCTTGAGGAATT

Identification of Oct4/Sox2/Nanog/Tcf3 occupied Feed-Forward Loops

To identify feed forward loops, we examined the recent data set identifying functional targets of the miR-290-295 cluster (Sinkkonen et al., 2008). In their study, Sinkkonen et al. identified miR-290-295 targets by both looking at mRNAs that increase in level in a Dicer -/- cell line and overlap that data set with mRNAs that decrease in expression when miR-290-295 mimic siRNA is added back to the cells. Because the promoter of the *mir-290-295* polycistron is occupied by Oct4/Sox2/Nanog/Tcf3, any targets of the miRNA cluster that are also occupied by the 4 factors would represent feed forward targets. Of the 245 miR-290-295 cluster targets identified in the Sinkkonen et al. study (2008), promoters for 64 are occupied by Oct4/Sox2/Nanog/Tcf3. This is approximately 50% more interactions then would be expected by random (binomial p-value < $1x10^{-4}$).

Interestingly, only a small minority of these genes are also occupied by significant quantities of the PRC2 subunit Suz12. Of the 64 targets whose promoters are occupied by Oct4/Sox2/Nanog/Tcf3, only 5 are occupied by domains of Suz12 binding >500bp (larger region sizes have been correlated with gene silencing, Lee et al., 2006). This may be because PcG bound proteins are not functional targets of mir-290-295 in mES cells. Alternatively these proteins are not expressed in ES cells following Dicer deletion and are thus excluded from the target list (Sinkkonen et al., 2008), but may be targets at other stages of development. In the later case, the miRNAs may serve as a redundant silencing mechanism, along with Polycomb group complexes, to help prevent even low levels of expression of the developmental regulators in ES cells.



(RT-PCR data derived from Loh et al., Nat. Genet., 2006)

(RT-PCR data derived from Boyer et al., Nature 2006)

Figure S1. Comparison of ChIP-seq and RT-PCR data for Oct4 and Suz12. a. ChIP-seq reads for Oct4 enrichment were compared to RT-PCR results previously described in Loh et al. 71 probes that were identified as enriched are shown in blue and 39 regions identified as not-enriched are shown in gray. The maximum number of ChIP-seq reads assigned within the region is shown on the vertical axis. Red line denoted the threshold of binding with $p < 10^{-9}$. Ambiguous RT-PCR results were excluded. **b.** ChIP-seq reads for Suz12 enrichement were compared to RT-PCR results previously described in Boyer et al. as in **a.** ChIP-seq data within 200bp of 68 probes identified as enriched by RT-PCR and confirmed by ChIP-chip are shown in blue and 18 probes identified as not-enriched are shown in gray.

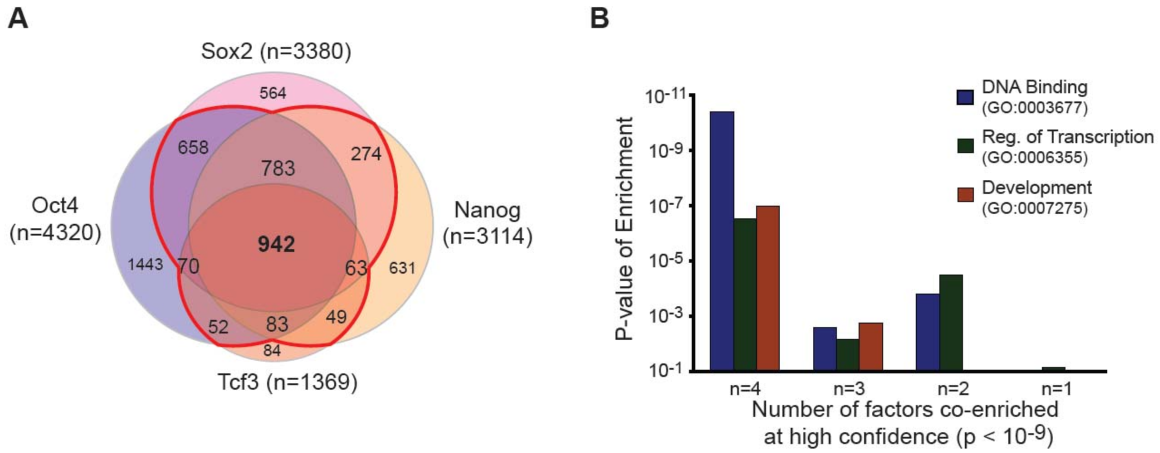


Figure S2. **Promoters for known genes occupied by Oct4/Sox2/Nanog/Tcf3 in mES cells**. **a.** Overlap of genes whose promoters are within 8kb of sites enriched for Oct4, Sox2, Nanog, or Tcf3. Not shown are the Nanog:Oct4 overlap (289) and Sox2:Tcf3 overlap (26). Red line deliniates genes considered occupied by Oct4/Sox2/Nanog/Tcf3. **b**. Enrichment for selected GO-terms previously reported to be associated with Oct4/Sox2/Nanog binding (Boyer et al., 2005) was tested on the sets of genes occupied at high-confidence for 1 to 4 of the tested DNA binding factors. Hypergeometric p-value is shown for genes annotated for DNA binding (blue), Regulation of Transcription (green) and Development (red).

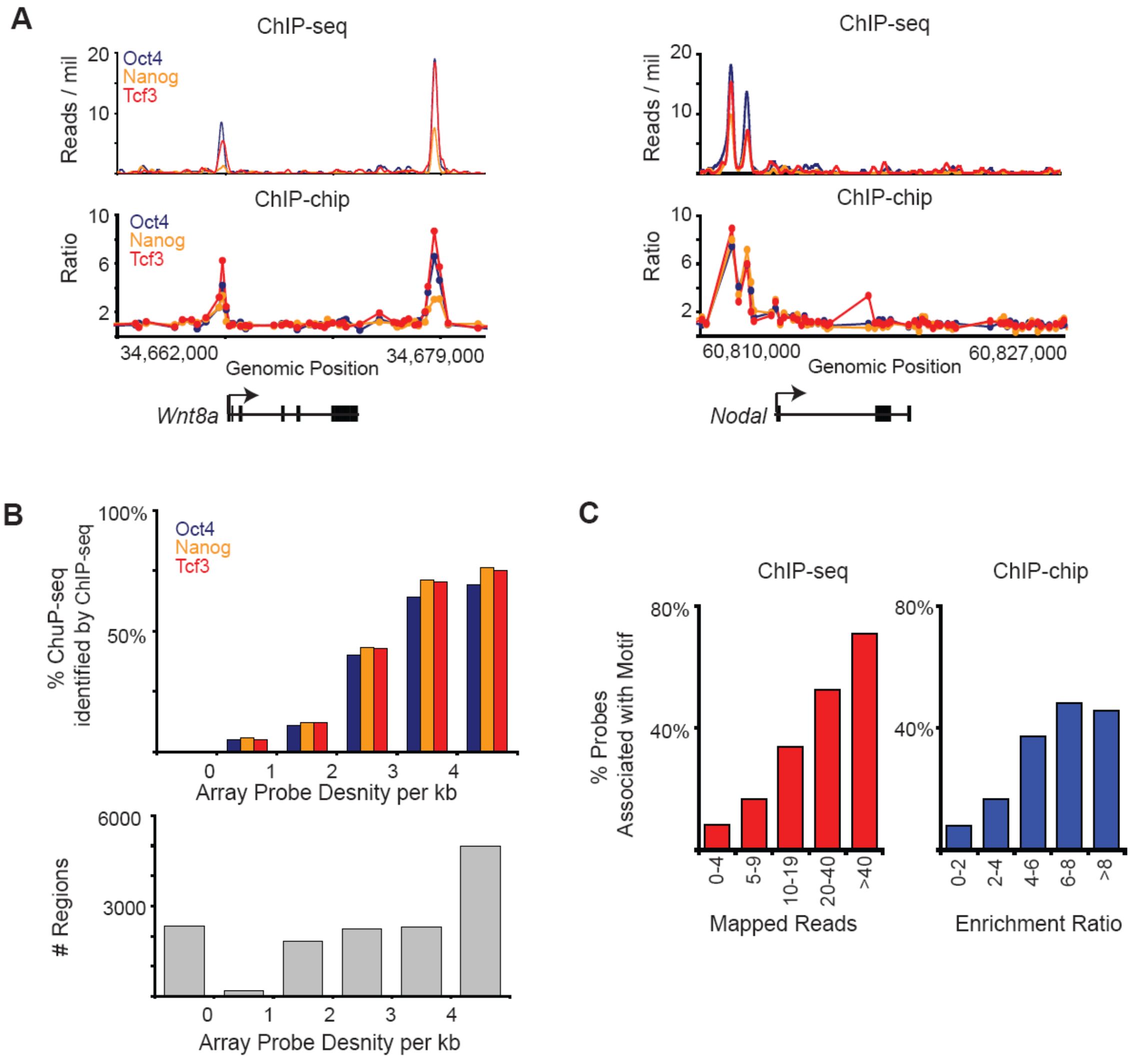


Figure S3. Comparison of ChIP-seq and ChIP-chip genome wide data for Oct4, Nanog and Tcf3. a. Binding of Oct4 (blue), Nanog (orange) and Tcf3 (red) across 17kb surrounding the Wnt8a and Nodal genes (black below the graph, arrow indicates transcription start site) as in **figure 1b**. (upper) Binding derived from ChIP-seq data plotted as reads per million. (lower) Binding derived from ChIP-chip enrichment ratios (Cole et. al., 2008) **b**. Poor probe density prevents detection of ~1/3 of ChIP-seq binding events on Agilent genome-wide tiling arrays. Top panel shows the fraction of regions that are occupied by Oct4/Sox2/Nanog/Tcf3 at high-confidence in mES cells as identified by ChIP-seq that are enriched for Oct4 (blue), Nanog (orange) and Tcf3 (red) on Agilent genome-wide microarrays (Cole et al., 2008). Numbers on the x-axis define the boundaries used to classify probe densities for the histogram. Bottom panel illustrates a histogram of the microarray probe densities of the enriched regions identified. **c.** Comparison of motif association. At the set of genome-wide ChIP-chip probe positions, we examined

the assocation between an Oct4 DNA motif and ChIP-chip and ChIP-seq enrichment. Probes / Bins were considered positive if they were associated with a high scoring motif within a 200 bp window (+/-100 bp). The background motif occurance for all probe positions is 8.2% (left most group). 1297 ChIP-seq bins and 421 ChIP-chip probes are included in the top categories respectively.

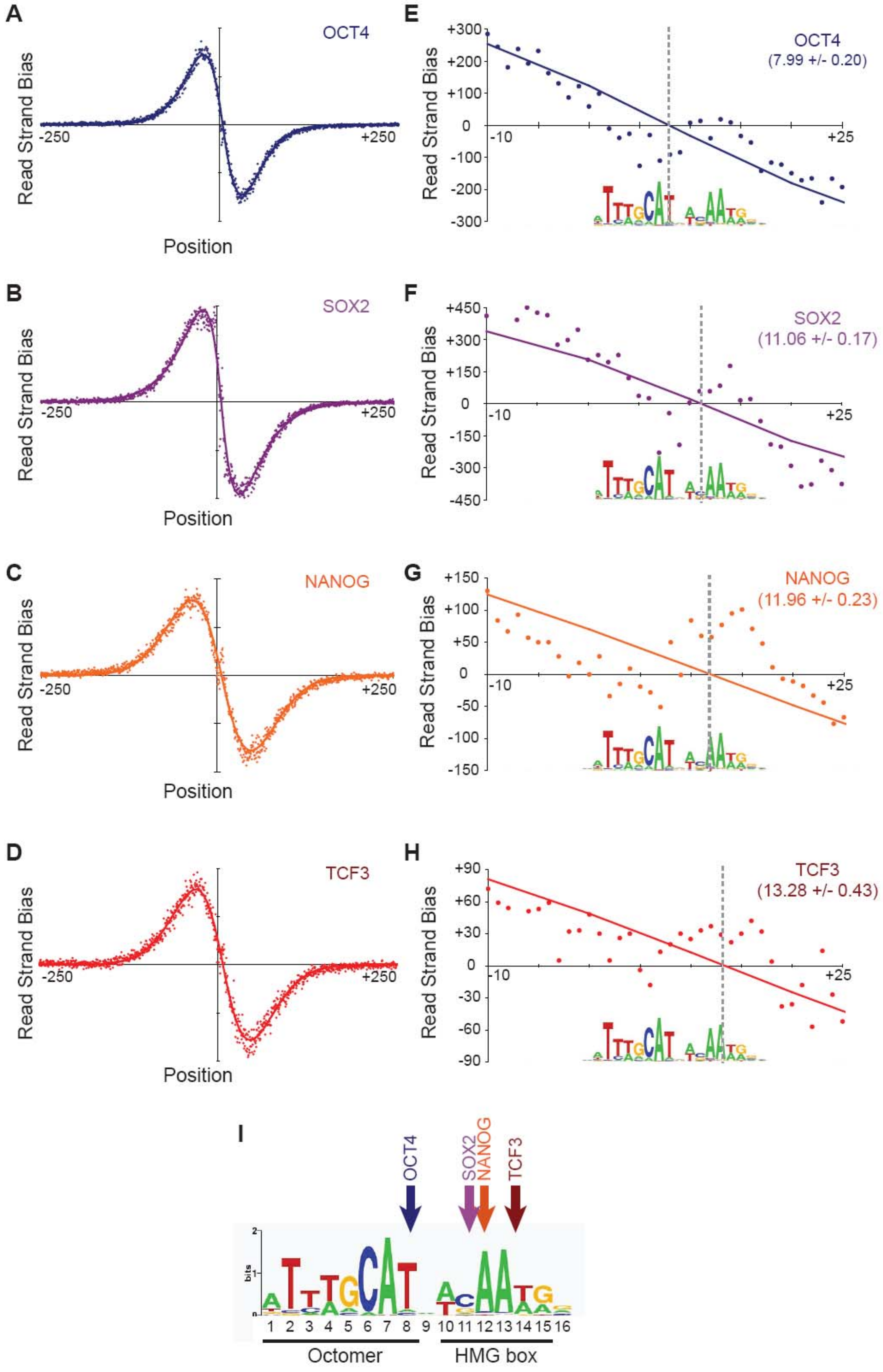

A
OCT4
Read Strand Bias
-250
+250
Position
B
SOX2
Read Strand Bias
-250
+250
Position
C
NANOG
Read Strand Bias
-250
+250
Position
D
TCF3
Read Strand Bias
-250
+250
Position
E
OCT4
(7.99 +/- 0.20)
+300
+200
+100
0
-100
-200
-300
-10
+25
Read Strand Bias
F
SOX2
(11.06 +/- 0.17)
+450
+300
+150
0
-150
-300
-450
-10
+25
Read Strand Bias
G
NANOG
(11.96 +/- 0.23)
+150
+100
+50
0
-50
-100
-150
-10
+25
Read Strand Bias
H
TCF3
(13.28 +/- 0.43)
+90
+60
+30
0
-30
-60
-90
-10
+25
Read Strand Bias
I
OCT4
SOX2
NANOG
TCF3
bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Octomer
HMG box

Figure S4. High resolution analysis of Oct4/Sox2/Nanog/Tcf3 binding based on Meta-analysis. a-d. Short sequence reads for **a.** Oct4, **b**. Sox2, **c.** Nanog, **d.** Tcf3 mapping within 250bp of 2000 highly enriched regions where the peak of binding was found within 50bp of a high quality Oct4/Sox2 motif were collected. Composite profiles were created at base pair resolution for forward and reverse strand reads centered on the Oct4/Sox2 motif (aligned at +1). The difference between the number of positive and negative strand reads are shown for each base pair (circles). The best fit line is shown for each factor (see Supplemental Text). **e-h** Zoomed in region of a-d showing 20bp surrounding the Oct4/Sox2 motif. Dashed line indicates the position where the best fit line crosses the X-axis. For reference, the motif is shown below each graph. **i.** Summary of meta-analysis for Oct4, Sox2, Nanog and Tcf3. Arrows indicate the nucleotide where each transcription factor switches from a positive strand bias to a negative strand bias. The octomer and HMG box motifs are indicated.

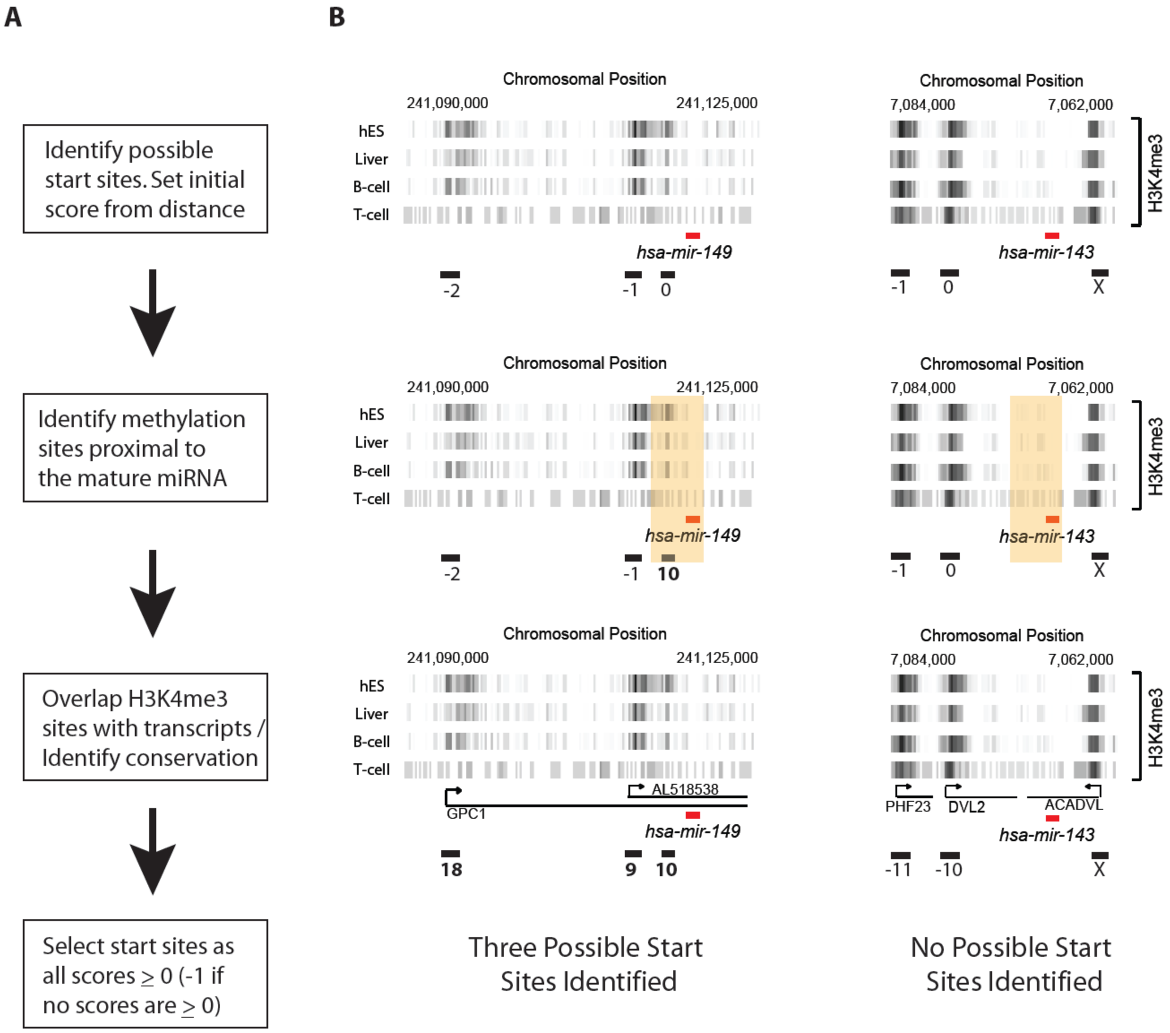

A
B
Identify possible start sites. Set initial score from distance
Identify methylation sites proximal to the mature miRNA
Overlap H3K4me3 sites with transcripts / Identify conservation
Select start sites as all scores ≥ 0 (-1 if no scores are ≥ 0)
Chromosomal Position
241,090,000
241,125,000
hES
Liver
B-cell
T-cell
hsa-mir-149
-2
-1
0
10
GPC1
AL518538
18
9
Chromosomal Position
7,084,000
7,062,000
H3K4me3
hsa-mir-143
-1
0
X
PHF23
DVL2
ACADVL
-11
-10
Three Possible Start Sites Identified
No Possible Start Sites Identified

Figure S5. Algorithm for Identification of miRNA promoters. a. Flowchart describing the method used to identify the promoters for primary miRNA transcripts in human and mouse. For a full description, see supplemental text. **b.** Two examples of identification of miRNA promoters. Top, Initial identification of possible start sites based on H3K4me3 enriched regions from four cell types. Enrichment of H3K4me3-modified nucleosomes is shown as shades of gray. Red bar represents the position of the mature miRNA. Black bars below the graph are regions enriched for H3K4me3. Initial scores are shown below the black bars. The region on the far right was excluded from the analysis (score = X) since it is downstream of the mature miRNA. Middle, Identification of candidate start sites <5kb upstream of the mature miRNA (yellow shaded area). Bottom, identification of candidate start sites that either initiate overlapping (left) or non-overlapping (right) transcripts. EST and transcript data is shown. Scores associated with identified genes are shown bold.

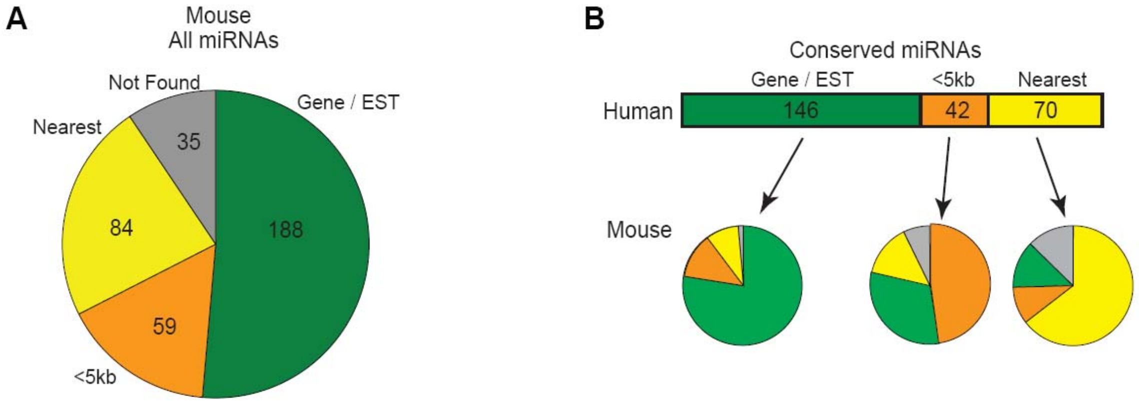


Figure S6. Summary of miRNA promoter classification. a. Promoters assigned to mature miRNAs were classified by the dominant feature of their scoring. Green: miRNAs that were found to have overlapping ESTs or genes confirming their promoters. Orange: miRNAs that were found to have a candidate start site within 5kb of the mature miRNA. Gray: miRNAs with either no candidates within 250kb of the mature miRNA or where all candidates had a score less then zero (**see Fig. S5b, right**). Yellow: miRNAs for which the closest candidate start site was selected solely on the basis of its proximity. **b**. The basis of miRNA promoter identification, including Gene or EST evidence (green), distance of <5 kilobases to mature miRNA (orange), nearest possible promoter to miRNA (yellow), tended to be conserved between human and mouse.

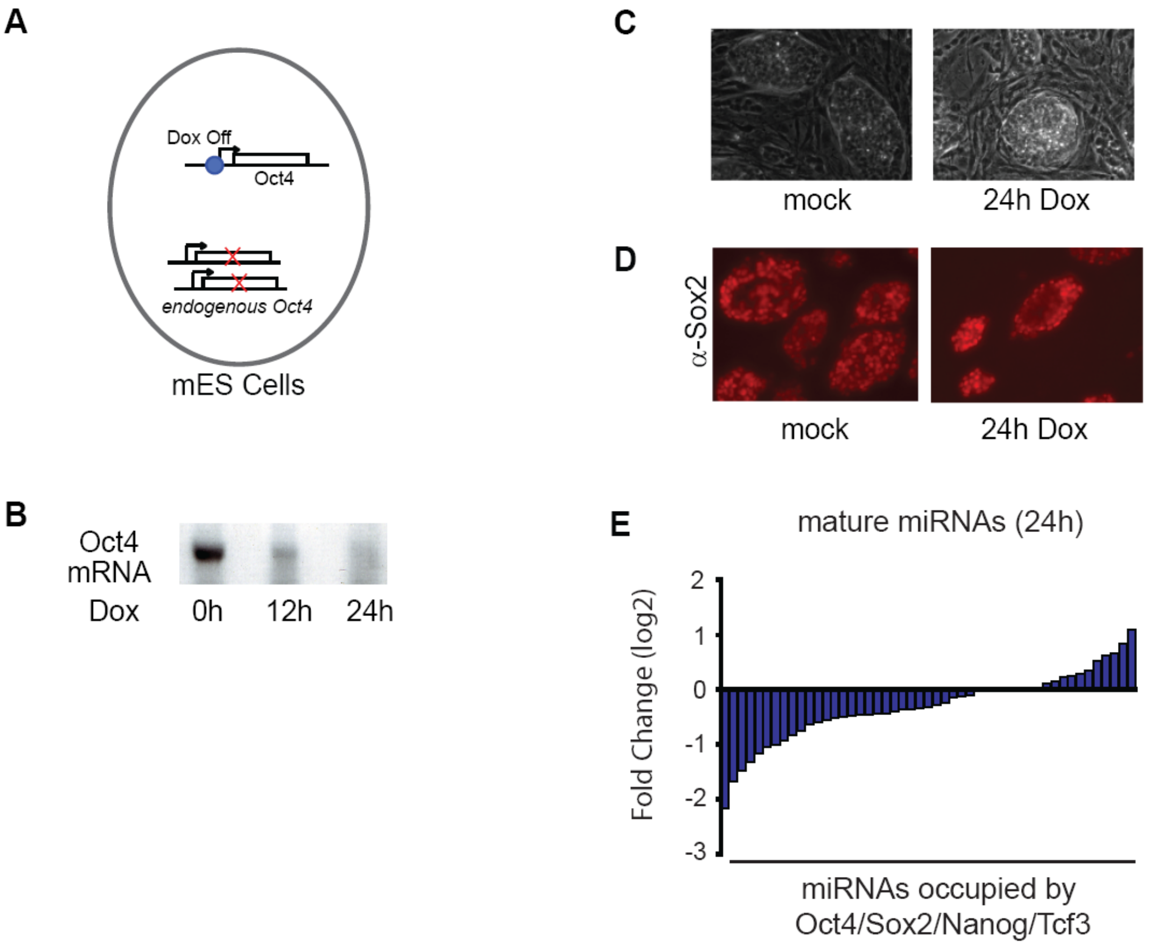


Figure S7. Regulation of miRNAs by Oct4. a. In an engineered murine cell line (Niwa et al., 2000), endogenous *Oct4* is deleted, and Oct4 expression is maintained by a Dox-repressible transgene. **b**. By 24 hours of Dox-treatment, Oct4 mRNA levels are reduced as shown by reverse transcription (RT)-PCR. **c**. 24 hours following Dox-treatment, cells remain ES-like by morphology. **d.** 24 hours following Dox-treatment Sox2 protein can still be detected by immunofluoresence.. **e.** Changes in levels of Oct4/Sox2/Nanog/Tcf3 occupied mature miRNAs based on Solexa sequencing of small RNAs. Fold change was calculated by comparing normalized read counts from untreated cells and cells 24 hours after Dox treatment. A full list of miRNA reads can be found in **Table S9**. Details about the normalization procedure are contained in the supplemental text.

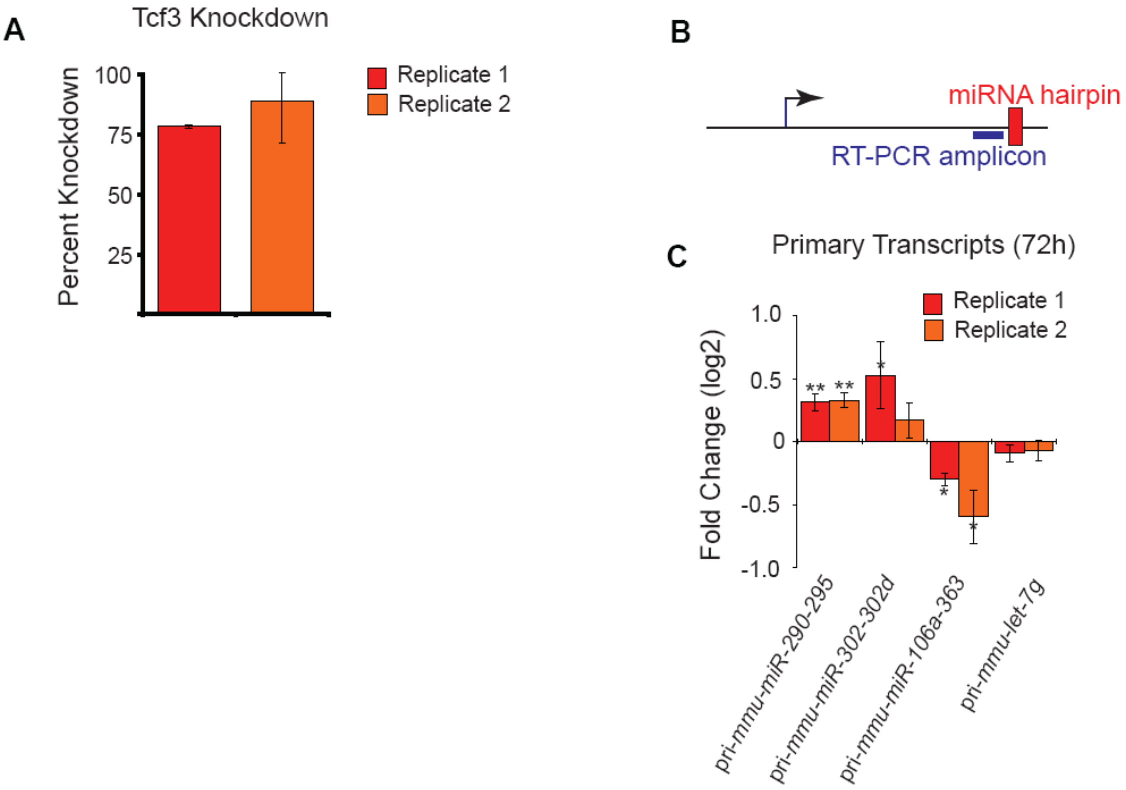


Figure S8. Regulation of miRNAs by Tcf3. Tcf3 was knocked down in V6.5 mES cells using lentiviral vectors containing shRNAs. **a.** RT-PCR confirmation of knockdown at 72 hours post-infection using Taqman probes against Tcf3 (relative to levels in cells infected with GFP control lentivirus). **b.** Schematic of the position of RT-PCR probes used to measure the levels of pri-miRNA transcripts in **Figure 3d** and part c.**c.** Results of quantitative reverse transcriptase(RT)-PCR analysis of probes designed to several pri-miRNAs occupied by Oct4/Sox2/Nanog/Tcf3. Change in the level of primary transcript compared to GFP control lentivirus are shown. * = $p < 0.05$, ** = $p < 0.001$ using a two-sampled t-test assuming equal variance. Standard deviation is indicated with error bars.

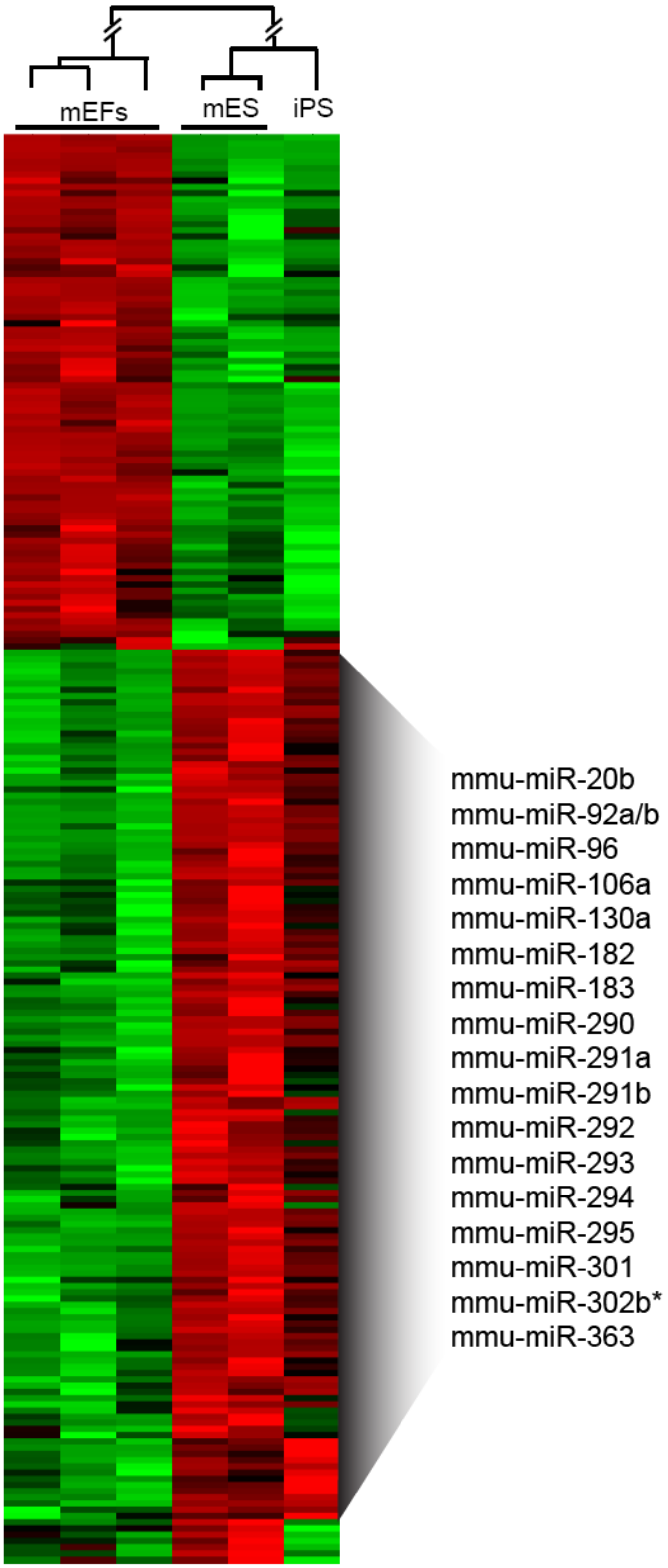


Figure S9. miRNA genes occupied by the core master regulators in ES cells are expressed in induced Pluripotent Stem cells (iPS).

RNA was extracted from MEFs (columns 1-3), mES cells (columns 4,5) and iPS cells (column 6) and hybridized to microarrays with LNA probes targeting all known miRNAs. Differentially expressed miRNAs enriched in either MEFs or mES cells are shown (FDR < 10%, see supplemental text, iPS cells were not used to determine differential expression). Data were Z-score normalized, and cell types were clustered hierarchically (top). Active miRNA promoters associated with Oct4/Sox2/Nanog/Tcf3 are listed to the right.

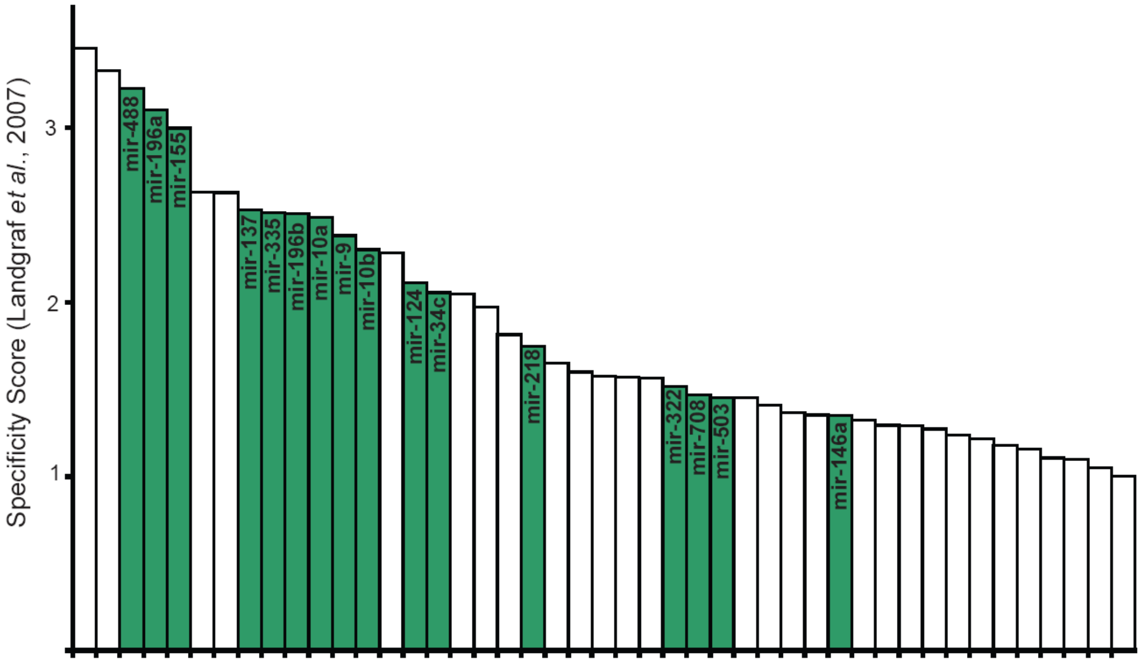


Figure S10. PcG occupied miRNAs are generally expressed in a tissue specific manner. Mature miRNAs derived from genes occupied by Suz12 and H3K27me3-modified nucleosomes were compared to the list of tissue specific miRNAs derived from the miRNA expression atlas (Landgraf et al., 2007). Vertical axis represents tissue-specificity and miRNAs with specificity score ≥1 are shown. miRNAs bound by Oct4/Sox2/Nanog/Tcf3 and expressed in mES cells are not shown (largely ES cell specific miRNAs). Among the tissue-specific miRNAs there is significant enrichment ($p < 0.005$ by hypergeometric distribution) for miRNAs occupied by Suz12 (green).

Supplemental References

Bailey, T. L., and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol *3*, 21-29.

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res *34*, W369-373.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Beissbarth, T., and Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics *20*, 1464-1465.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., *et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947-956.

Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L. A., Lee, T. I., Levine, S. S., Wernig, M., Tajonar, A., Ray, M. K., *et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Calabrese, J. M., Seila, A. C., Yeo, G. W., and Sharp, P. A. (2007). RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. Proc Natl Acad Sci U S A *104*, 18097-18102.

Chang, T. C., Wentzel, E. A., Kent, O. A., Ramachandran, K., Mullendore, M., Lee, K. H., Feldmann, G., Yamakuchi, M., Ferlito, M., Lowenstein, C. J., *et al.* (2007). Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. Mol Cell *26*, 745-752.

Cole, M. F., Johnstone, S. E., Newman, J. J., Kagey, M. H., and Young, R. A. (2008). Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. Genes Dev *22*, 746-755.

Corney, D. C., Flesken-Nikitin, A., Godwin, A. K., Wang, W., and Nikitin, A. Y. (2007). MicroRNA-34b and MicroRNA-34c are targets of p53 and cooperate in control of cell proliferation and adhesion-independent growth. Cancer Res *67*, 8433-8438.

Fukao, T., Fukuda, Y., Kiga, K., Sharif, J., Hino, K., Enomoto, Y., Kawamura, A., Nakamura, K., Takeuchi, T., and Tanabe, M. (2007). An evolutionarily conserved mechanism for microRNA-223 expression revealed by microRNA gene profiling. Cell *129*, 617-631.

Gerhard, D. S., Wagner, L., Feingold, E. A., Shenmen, C. M., Grouse, L. H., Schuler, G., Klein, S. L., Old, S., Rasooly, R., Good, P., *et al.* (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res *14*, 2121-2127.

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell *27*, 91-105.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., and Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., *et al.* (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res *34*, D590-598.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., *et al.* (2005). Ensembl 2005. Nucleic Acids Res *33*, D447-453.

Johnson, D., Martazavai, A., Myers, R., Wold, B., (2007). Genome-wide mapping of inn vivo protein-DNA interactions. Science *316*, 1441-2.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.
Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853-858.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., *et al.* (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. Cell *129*, 1401-1414.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science *294*, 858-862.

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., Isono, K., *et al.* (2006a). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.

Lee, T. I., Johnstone, S. E., and Young, R. A. (2006b). Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc *1*, 729-748.

Loh, Y. H., Wu, Q., Chew, J. L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., *et al.* (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet *38,* 431-440.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Niwa, H., Miyazaki, J., and Smith, A. G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat Genet *24,* 372-376.

O'Donnell, K. A., Wentzel, E. A., Zeller, K. I., Dang, C. V., and Mendell, J. T. (2005). c-Myc-regulated microRNAs modulate E2F1 expression. Nature *435,* 839-843.

Okabe, S., Forsberg-Nilsson, K., Spiro, A. C., Segal, M., and McKay, R. D. (1996). Development of neuronal precursor cells and functional postmitotic neurons from embryonic stem cells in vitro. Mech Dev *59,* 89-102.

Pall, G. S., Codony-Servat, C., Byrne, J., Ritchie, L. & Hamilton, A. Carbodiimide mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. Nucleic Acids Res 35, e60 (2007).

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *33*, D501-504.

Rozen, S., Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. *132,* 365-86

Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre, N. C., Schreiber, S. L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419,* 407-411.

Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C. G., Zavolan, M., Svoboda, P., and Filipowicz, W. (2008). MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. Nat Struct Mol Biol *15,* 259-267.

Tam, W. L., Lim, C. Y., Han, J., Zhang, J., Ang, Y. S., Ng, H. H., Yang, H., and Lim, B. (2008). Tcf3 Regulates Embryonic Stem Cell Pluripotency and Self-Renewal by the Transcriptional Control of Multiple Lineage Pathways. Stem Cells.

Tucker, K. L., Wang, Y., Dausman, J., and Jaenisch, R. (1997). A transgenic mouse strain expressing four drug-selectable marker genes. Nucleic Acids Res *25,* 3745-3746.

Valoczi, A. et al. Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide probes. Nucleic Acids Res 32, e175 (2004).

Voorhoeve, P. M., le Sage, C., Schrier, M., Gillis, A. J., Stoop, H., Nagel, R., Liu, Y. P., van Duijse, J., Drost, J., Griekspoor, A., *et al.* (2006). A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. Cell *124*, 1169-1181.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B. E., and Jaenisch, R. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature *448*, 318-324.

Yi, F., Pereira, L., and Merrill, B. J. (2008). Tcf3 Functions as a Steady State Limiter of Transcriptional Programs of Mouse Embryonic Stem Cell Self Renewal. Stem Cells.