Supplemental Data

Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells

Laurie A. Boyer, Tong Ihn Lee, Megan F. Cole, Sarah E. Johnstone, Stuart S. Levine, Jacob P. Zucker, Matthew G. Guenther, Roshan M. Kumar, Heather L. Murray, Richard G. Jenner, David K. Gifford, Douglas A. Melton, Rudolf Jaenisch, and Richard A. Young

Table of Contents:

Supplemental Text

Supplemental Figures

Human ES Cell Controls

Immunohistochemical Analysis of Pluripotency Markers

For analysis of pluripotency markers, cells were fixed in 4% paraformaldehyde for 30 minutes at room temperature and incubated overnight at 4°C in blocking solution (5 ml Normal Donkey Solution:195 ml PBS + 0.1% Triton-X)(Figure S3). After a brief wash in PBS, cells were then incubated with primary antibodies to Oct-3/4 (Santa Cruz sc-9081), SSEA-3 (MC-631; Solter and Knowles, 1979), SSEA-4 (MC-813-70; Solter and Knowles, 1979), Tra-1-60 (MAB4360; Chemicon International), and Tra-1-81 (MAB4381; Chemicon International) in blocking solution overnight at 4°C. Following incubation with primary antibody, cells were incubated with either rhodamine red or FITC-conjugated secondary antibody (Jackson Labs) for 2-5hrs at 4°C. Nuclei were stained with 4',6-diamidine-2-phenylidole dihydrochloride (DAPI). Epifluorescent images were obtained using a fluorescent microscope (Nikon TE300). Our analysis indicated that >80% of the H9 cells were strongly positive for all pluripotency markers. Alkaline phosphatase activity of human ES cells was analyzed using the Vector Red Alkaline Phosphatase Subtrate Kit (Cat. No. SK-5100; Vector Laboratories) according to manufacturer's specifications and the reaction product was visualized using fluorescent microscopy.

Teratoma Formation

Teratomas were induced by injecting 2-5 x 10^6 cells into the subcutaneous tissue above the rear haunch of 6 week old Nude Swiss (athymic, immunocompromised) mice. Eight to twelve weeks post-injection, teratomas were harvested and fixed overnight in 4% paraformaldehyde at 4°C. Samples were then immersed in 30% sucrose overnight before embedding the tissue in O.C.T freezing compound (Tissue-Tek). Cryosections were obtained and 10µm sections were incubated with the appropriate antibodies as above and analyzed for the presence of the following differentiation markers by confocal microscopy (LSM 210): neuronal class II β-tubulin, Tuj1 (ectoderm; MMS-435P Covance); striated muscle-specific myosin, MF20 (mesoderm; kind gift from D. Fischman), and alphafetoprotein (endoderm; DAKO) (Figure S4). 4',6-diamidine-2-phenylidole dihydrochloride (DAPI) staining was used to identify nuclei. Antibody reactivity was detected for markers of all three germ layers confirming that the human embryonic stem cells used in our analysis had maintained differentiation potential.

Embryoid Bodies (EB)

ES cells were harvested by enzymatic digestion and EBs were allowed to form by plating ~1 X 10^6 cells/well in suspension in 6-well non-adherent, low cluster dishes for 30 days. EBs were grown in the absence of leukemia inhibitory factor (LIF) and basic fibroblast growth factor (bFGF) in culture medium containing 2x serum replacement. EBs were then harvested, fixed for 30 minutes in 4% paraformaldehyde at room temperature, and placed in 30% sucrose overnight prior to embedding the tissue in O.C.T. freezing compound (Tissue- Tek). Cryosections were obtained as described for teratoma formation. Confocal images were obtained for all three germ layer markers again confirming that the H9 cells used in our analysis had maintained differentiation potential (data not shown; results similar to those shown in Figure S4).

Chromatin Immunoprecipitation Assays

Antibodies

The Nanog (AF1997 R&D Systems) and Sox2 (AF2018 R&D Systems) antibodies used in this study were immunoaffinity purified against the human protein and shown to recognize their target protein in Western blots and by immunocytochemistry (R&D Systems Minneapolis, MN). Multiple Oct4 antibodies directed against different portions of the protein were used for location analysis (AF1759 R&D Systems, sc-8628 Santa Cruz, sc-9081 Santa Cruz), some of which were immunoaffinity purified and have been shown to recognize their target protein in Western blots and by immunocytochemistry. Prior to conducting the experiments with Agilent arrays, we compared these three Oct4 antibodies by performing location analysis with self-printed promoter arrays and found that they performed similarly. ChIP experiments carried out with AF1759 and sc-8628 were hybridized to the Agilent 10-array sets. In addition, our immunofluorescence results indicated that a nuclear protein was detected only in undifferentiated ES cells with the Oct4 antibody (sc-9081) (Figure S3; compare ES cell with MEF). E2F4 antibodies (sc-1082) were obtained from Santa Cruz Biotech and have been shown to specifically recognize previously reported E2F4 target genes (Table S2) (Ren et al., 2002; Weinmann et al., 2002).

Chromatin Immunoprecipitation

Protocols describing all materials and methods can be downloaded from http://jura.wi.mit.edu/young/hESRegulation/.

Human embryonic stem cells were grown to a final count of $5x10^7$ – $1x10^8$ cells for each location analysis reaction. 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 and 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 power 7 for 10 x 30 second pulses (90 second pause between pulses) at 4°C while samples were immersed in an ice bath. The resulting whole cell extract was incubated overnight at 4°C with 100 µl of Dynal Protein G magnetic beads that had been preincubated with 10 µg of the appropriate antibody. Beads were washed 5 times with RIPA buffer and 1 time with TE containing 50 mM 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. Immunoprecipitated DNA and whole cell extract DNA were then purified by treatment with RNAseA, 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 Invitrogen Bioprime random primer labeling kits (immunoenriched DNA was labeled with Cy5 fluorophore, whole cell extract DNA was labeled with Cy3 fluorophore). Labeled DNA was combined (5 – 6 µg each of immunoenriched and whole cell extract DNA) and hybridized to arrays in Agilent hybridization chambers for 40 hours at 40°C. Arrays were then washed and scanned.

Control ChIPs

Location analysis experiments were performed with both rabbit and goat IgG in human ES cells as a negative control. We did not find any enrichment for sequences occupied by Oct4, Sox2 and Nanog (Figure S5). Control ChIP experiments were also performed with E2F4. We did not observe any substantial overlap among the Oct4, Sox2, and Nanog targets and E2F4 target genes. Additional control experiments were performed to address the potential cross-reactivity of the antibodies to other family members (Figure S5). We carried out ChIP with Oct4 (sc-8628 Santa Cruz), Sox2 (AF2018 R&D Systems), and Nanog (AF1997 R&D Systems) antibodies in HepG2 cells that do not express these factors, but which express other POU and HMG domain proteins. This experiment did not yield any significantly enriched targets.

Array Design

The following describes the design of the 10-slide promoter arrays that contain approximately 400,000 features used in this study. Arrays were produced by Agilent Technologies (www.agilent.com)

Selection of Regions and Design of Subsequences

To select well-characterized transcription start sites, we first collected the coordinates of all transcription start sites described in five different databases: RefSeq, Ensembl, MGC, VEGA (www.vega.sanger.ac.uk) and Broad (www.broad.mit.edu). The first three are commonly used databases for gene annotation, the last two are manually annotated databases covering subsets of the human genome from the Sanger Institute and Broad Institute, respectively. We then filtered for all transcription start sites that appeared in any two of these five databases (start sites separated by less than 500 bp in any of the databases were considered identical for this step). In cases where there were multiple start sites with different genomic coordinates, we selected the start site that would result in the longest transcript. A total of 18,002 start sites were selected. All sequences and coordinates are from the May 2004 build of the human genome (NCBI build 35), using the repeatmasked (-s) option which separates the genome into masked and unmasked subsequences. We used the program ArrayOligoSelector (AOS, Bozdech et al., 2004) to score 60-mers for every unmasked subsequence greater than 62 bp across all promoter regions. The scores for each oligo were retained but not put through the built-in AOS selection process.

The collection of scored 60-mers was divided by promoter and sorted by genomic position. Each set of 60-mers was then filtered based on the oligo scoring criteria. AOS uses a scoring system for four criteria: GC content, self-binding, complexity and uniqueness. For our most stringent filter, we selected the following ranges for each parameter: GC content between 30 percent and 100 percent, self-binding score less than 100, complexity score less than or equal to 24, uniqueness greater than or equal to –40.

From this subset of 60-mers, we selected oligos designed to cover the promoter region with an estimated density of one probe every 280 basepairs. To achieve more uniform tiling, we instituted a simple method to find probes within a particular distance from each other. Starting at the upstream end of the region, we selected the first qualified probe, then selected the next qualified probe that was between 150 bp and 280 bp away. If there were multiple, eligible probes, we chose the most distal probe within the 280 bp limit. If there were no probes within this limit, we continued scanning until we found the next acceptable probe. The process was then repeated with the most recently selected probe until we reached the end of the promoter region.

For regions that were not covered by high quality probes, we returned to the full set of scored 60-mers and filtered using less stringent criteria. This gave us an additional set of 60-mers that we then used to fill gaps in our coverage. After this second pass, we identified gaps in our coverage and added oligos that were properly spaced and best fit our criteria regardless of whether they passed the filter cutoffs. This iterative process gave us a compromise between optimal probe quality and optimal probe spacing. For each start site, we selected the region 8 kb upstream and 2 kb downstream of the site for tiling.

Compiled Probes and Controls

The design process described was used to generate a set of 10 Agilent microarrays containing a total of 399,309 features designed for 18,002 transcription start sites representing 17,917 unique genes. Each array contains between 39,904 and 39,961 features. The probes are arranged such that array 1 begins with the first qualified transcription start site on the left arm of chromosome 1, array 2 picks up where array 1 ends, array 3 picks up where array 2 ends, and so on. There are some gaps in coverage that reflect our inability to identify high quality unique 60-mers: these tend to be unsequenced regions, highly repetitive regions that are not repeat masked (such as telomeres or gene families) and certain regions that are probably genome duplications. As an estimate of probe density, approximately 96% of all 60-mers are within 400 bp of another 60-mer; approximately 90% of all 60-mers are within 280 bp of another 60-mer.

We added several sets of control probes (2,043 total) to the array designs. On each array, there are 40 oligos designed against six *Arabidopsis thaliana* genes and printed in triplicate. These *Arabidopsis* oligos have been carried over from previous array designs and were intended for eventual use with spike-in controls. These oligos were BLASTed against the human genome and did not register any significant hits. An additional 543 *Arabidopsis* oligos were selected as negative controls based on their failure to show any significant BLAST hits against the human genome. Since E2F4 chromatin immunoprecipitations can be accomplished with a wide range of cell types and have provided a convenient positive control for ChIP-Chip experiments (for putative regulators where no prior knowledge of targets exist, for example), we added a total of 80 oligos representing four proximal promoter regions of genes that are known targets of the transcriptional regulator E2F4 (NM_001211, NM_002907, NM_031423, NM_001237). Each of the four promoters is represented by 20 different oligos that are evenly positioned across the region from 3 kb upstream to 2 kb downstream of the transcription start site. We also included a control probe set that provides a means to normalize intensities across multiple slides throughout the entire signal range. There are 384 oligos printed as intensity controls; based on test hybridizations, this set of oligos gives signal intensities that cover the entire dynamic range of the array. Twenty additional intensity controls, representing the entire range of intensities, were selected and printed fifteen times each for an additional 300 control features. We also incorporated 616 “gene desert” controls. To design these probes, we identified intergenic regions of 1 Mb or greater and designed probes in the middle of these regions. These are intended to identify genomic regions that are least likely to be bound by promoter-binding transcriptional regulators (by virtue of their extreme distance from any known gene). We have used these as normalization controls in situations where a factor binds to a large number of promoter regions. In addition to these 2,043 controls, there are 2,256 controls added by Agilent (standard) and a variable number of blank spots bringing the total number of features on each slide to 44,290.

	Start		*End*		
Slide	*Chr*	*Pos*	*Chr*	*Pos*	*Probes*
1	1	5575	1	224646230	39961
2	1	224694779	3	108726269	39909
3	3	109290599	5	147564193	39937
4	5	147665548	7	106280884	39935
5	7	106395416	10	15044190	39925
6	10	15119596	11	129697251	39905
7	11	129802259	14	94119500	39930
8	14	94140702	17	41335175	39938
9	17	41603407	20	30042900	39940
10	20	30054185	Y	57685547	39930

Replicate Data Sets

Multiple batches of ES cells were cultured and each was tested for expression of pluripotency markers and the potential to differentiate into derivatives of the three embryonic lineages. Independent batches of ES cells were used to perform independant ChIP experiments as described above for each of the three transcription factors. Biological replicates were performed with the same Nanog (AF1997 R&D Systems) and Sox2 (AF2018 R&D Systems) antibody or two different antibodies against Oct4 (Sc-8628 Santa Cruz; AF1759 R&D Systems). ChIPs for each of the three different transcription factors were hybridized to independent Agilent array sets.

Array Scan and Data Extraction

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.

Data Normalization and Analysis

GenePix software was used to obtain background-subtracted intensity values for each fluorophore for every feature on the array. To obtain set-normalized intensities, we first calculated, for each slide, the median intensities in each channel for a set of 1,420 control probes that are included on each array. We then calculated the average of these median intensities for the set of 10 slides. Intensities were then normalized such that the median intensity of each channel for an individual slide equaled the average of the median intensities of that channel across all slides.

Each slide contains 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 set-normalized intensities of all other features.

To correct for different amounts of genomic and immunoprecipitated DNA hybridized to the chip, the set-normalized, negative control-subtracted median intensity value of the IP-enriched DNA channel was then divided by the median of the genomic DNA channel. This yielded a normalization factor that was applied to each intensity in the genomic DNA channel.

Next, we calculated the log of the ratio of intensity in the IP-enriched channel to intensity in the genomic DNA 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. The X scores for an array are assumed to be normally distributed which allows for calculation of a p-value for the enrichment ratio seen at each feature.

Identification of Bound Regions

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 neighbors. 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 labeled DNA fragments hybridized to the array. 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 center 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 Transcription Factor Bound Regions to Known Genes

The coordinates for the complete list of bound regions can be found in Tables S1, S3, S4, and S6 (see Index of Tables).

Comparisons to Known Genes

The location of all bound regions were compared to a composite database of genes compiled from three databases: RefSeq (Pruitt et al., 2005), Mammalian Gene Collection (MGC) (Gerhard et al., 2004), and Ensembl (Hubbard et al., 2005). This database was generated by compiling genes with Entrez Gene IDs, and adding additional genes or transcription start sites from the above databases as necessary. Transcripts that overlapped multiple non-nested genes on the same strand were not used. By this method, 22,200 unique genes were identified. Genes lacking formal names are identified by their transcript ID number. All coordinate information was downloaded in January 2005 from the UCSC Genome Browser (NCBI build 35). The annotated gene lists are available for download from our website (web.wi.mit.edu/young/hESregulation/).

Analysis of Error Rates in Location Analysis Experiments

Estimating a false positive and false negative rate is challenging as the estimates depend on perfect knowledge of a ground truth or confirmation by other experimental techniques that will each have their own bias. For the array platform used here, our experience with yeast provides an estimate of the error inherent in the platform. In this case, we selected a set of positives and negatives for the binding of Gcn4, a well-studied yeast transcription factor. The 84 positive genes were selected using three criteria: previous high confidence binding data ($P \leq 0.001$) (Harbison et al., 2004), the presence of a perfect or near perfect Gcn4 consensus binding site (TGASTCA) in the promoter region (-400bp to +50bp), and a greater then 2-fold change in steady state mRNA levels dependent on Gcn4 when shifted to amino acid starvation medium (Natarajan et al., 2001). The negative list of 222 genes was selected by weak binding ($P \geq 0.1$), absence of a motif near the presumed start site, and less then a 20% change in steady state mRNA levels in response to shift to amino acid starvation.

Using these positive and negative sets, we used ROC curve analysis (Statistics-ROC package for Perl) to evaluate a range of different IP/WCE ratio thresholds for false positive and false negative rates. Essentially, we examined a range of thresholds to denote "bound" and asked how many false positives and false negatives were detected at each threshold. Each gene was scored based on the maximum median-normalized IP/WCE ratio found in the region -250 to +50bp from the UAS. With the optimal cutoff for minimizing false positives (a 3.5 fold ratio), the data suggest a false positive rate of less than 0.5% and a false negative rate of ~20%. Thus, the oligo array platform is capable of generating extremely accurate, high quality data.

Comparing Binding and Expression Data

Processing Gene Expression Data

MPSS data: Three MPSS datasets were collected, two from a pool of the ES cell lines H1, H7 and H9 and one for HES-2 (Brandenberger et al., 2004; Wei et al., 2005). For each study, only MPSS tags detected at or over 4 transcripts per million (tpm) were used for further analysis. In addition, the data provided by Wei and colleagues (2005) allowed us to select only those tags that could be mapped to a single unique location in the human genome. For tags without a corresponding EntrezGene ID, IDs were assigned using the gene name or RNA accession numbers provided by the authors.

Gene expression microarray data: Four Affymetrix HG-U133 gene expression datasets were collected for the cell lines H1, H9, HSF1 and HSF6 (Abeyta et al., 2004; Sato et al., 2003). EntrezGene IDs were assigned to the probe sets using Affymetrix annotation or using RNA accession numbers provided by the authors. For each probeset, we counted the number of "Present" calls in the three replicate array experiments performed for each cell line. Many genes are represented by more than one probeset and, to enable comparison to MPSS data, we then found the maximum number of P calls for each gene (defined by unique EntrezGene ID). In each study, the cell lines were analyzed is triplicate. A gene was defined as detected if it was called "Present" in at least 2 of the 3 replicate experiments.

Defining Expressed Genes Using Multiple Expression Datasets

In order for a gene to be defined as expressed, we required that the gene fit one of three criteria: detected in at least one MPSS experiment and at least one Affymetrix experiment, consistently detected across all three MPSS experiments or consistently detected across all four Affymetrix experiments. As described above, a gene was considered detected if present at 4 tpm or more by MPSS analysis or if two out of three Affymetrix replicates called the gene "Present". These criteria allow us to capture the set of genes that were most consistently detected, including those genes where one experimental approach or the other is unable to detect expression due to technological limitations (for instance, genes detected by MPSS that are not included on the Affymetrix array).

Comparing Expression Patterns between ES Cells and Differentiated Cells

We examined the relative expression levels of genes bound by Oct4, Sox2 and Nanog in ES cells compared to differentiated cell and tissue types. In order to compare ES cells with as many human cell and tissue types as possible, we combined the data from three studies, all performed using the Affymetrix HG-U133A platform: 3 replicates of H1 ES cells (Sato et al., 2003), 3 replicates each of H9, HSF1 and HSF6 ES cells (Abeyta et al., 2004) and 2 replicates of 79 other human cell and tissue types (Su et al., 2004). To generate a measurement of the expression changes between undifferentiated ES cells and differentiated cells, each dataset was scaled to 150 using GCOS (Affymetrix). Then, for each gene, ratios were generated from the median signal intensity of each gene across all experiments. EntrezGeneIDs were assigned to each probe-set and for genes with multiple probe-sets, the expression ratios averaged. This resulted in a final set of 12,968 unique genes. For each gene, the significance of relative overexpression in the 12 ES cell experiments versus the 158 non-ES cell experiments was identified using a Mann-Whitney U-test. This metric was used to order genes shown in Figure 3A.

We further explored the hypothesis that bound genes are regulated by these transcription factors by taking advantage of the fact that Oct4 and Nanog are expressed in ES cells but their expression is rapidly downregulated upon differentiation. We compared the expression of Oct4, Sox2 and Nanog co-occupied genes in human ES cells with expression patterns in 79 differentiated cell types and focused the analysis on transcription factor genes because these were the dominant functional class targeted by the ES cell regulators (Figure 3B). We expected that for any set of genes, there would be a characteristic change in expression levels between ES cells and differentiated cells. The distribution of fold change ratios (log base 2) was calculated for transcription factors bound by Oct4, Sox2 and Nanog and transcription factors not bound by any one of the three factors. If Oct4, Sox2 and Nanog do not regulate the genes they occupy, then these genes should have the same general expression profile as the control population. We found, however, a significant shift in the distribution of expression changes for genes occupied by Oct4, Sox2 and Nanog (p-value < 0.001 using a two-sampled Kolmogorov-Smirnov test). The results for the H9 cell line are shown in Figure 3B. Similar results were obtained when using any other ES cell line or when using the average of all four ES expression datasets (data not shown).

Any factor-dependent effects on the profile could impact a combination of different characteristics, including the proportion of genes showing expression changes, the magnitude of changes or even whether the expression change is generally positive or negative. In general, these binding dependent effects on the profiles of sets of expression changes should be subtle. Many other factors are potentially contributing to the overall regulation of target genes and biologically relevant levels of gene expression changes may not be robust.

Gene Ontology Classification

Gene Ontology datasets were downloaded from the NCBI and gene ontology websites in February 2005. P-values were calculated using RefSeq genes that are both represented on the promoter array set and that have an associated ontology. Enriched terms (p-value < 10^{-6}) from all possible combinations of datasets are shown in Supplementary Table S7.

Oct4, Sox2, and Nanog Binding to the Oct4 Promoter Proximal Region

The oligo selection algorithm used for probe design has stringent criteria to assure the selection of unique and appropriately spaced probes covering each promoter of interest. However, this can result in an inability to find probes for some regions. In one case, the promoter region for Oct4 is poorly tiled on this set of arrays. As this promoter is one of the key targets in this study, we hybridized Oct4, Sox2 and Nanog immunoenriched material to a slide from a separate whole genome design that has more complete coverage of the Oct4 promoter region. Where possible, we used the exact same labeled, purified material (both IP and whole cell extract control) that was used on the original 10-slide set. The results indicate that all three factors co-occupy the same area of the Oct4 upstream (Figure S2).

Oct4 and Sox2 Binding to UTF1 and FGF4

UTF1 and FGF4 have been identified as key targets of Oct4 and Sox2 in mouse ES cells (Nishimoto et al., 1999; Yuan et al., 1995). It is not known if UTF1 or FGF4 play similar roles or whether these can be considered model target genes in human ES cells. The binding of these factors to their respective genes occurs at the 3' UTR and these sequences are not included in our current array design. The inclusion of the appropriate human sequences for FGF4 and UTF1 could serve as useful positive controls for our analysis, so we sought to determine whether these genes are also bound by Oct4 and Sox2 in human ES cells. We designed an array to contain the appropriate sequences. When possible, we used the exact same labeled, purified material (both IP and whole cell extract control) that was used on the original 10-slide set and hybridized labeled material from an Oct4 and Sox2 ChIP to this array. We found that Oct4 and Sox2 do occupy the 3'UTR of the co-activator UTF1, but that these factors are absent from FGF4 in human ES cells (Figure S6). This is consistent with the variable expression data with regard to FGF4 in human ES cells.

Distribution of Oct4, Sox2, and Nanog Binding Relative to Transcription Start Sites

We designed the arrays against the –8kb to +2kb region relative to each transcription start site because binding events in these regions are most likely to be connected with regulation of the associated gene. It is possible that a binding event controls a neighboring or even distal gene. Indeed, the further the binding event from the transcription start site of a gene, the more likely that event is associated with control of another gene.

It was determined how often we find binding events in various portions of the –8 kb to +2kb regions (Figure S7). We found that 35-50% of the binding sites occured within 1kb of a transcriptions start site. We found that only a small portion (6%) of all the binding events we identified occur in the -8kb to –7kb region. We then measured the distance from the binding events that occur in the -8kb to –7kb region to the closest transcription start site. The transcription start site for an adjacent gene occurred within 8kb in less than half of the cases (12, 31 and 37 binding events for Oct4, Sox2 and Nanog, respectively). For sites that are within 8kb of multiple genes both genes were assigned as candidate targets. It would be difficult and perhaps inappropriate to assign one but not both proximal promoters as likely targets since it is known that transcription factor binding events can affect multiple adjacent promoters.

Supplemental References

Abeyta, M.J., Clark, A.T., Rodriguez, R.T., Bodnar, M.S., Pera, R.A., and Firpo, M.T. (2004) Unique gene expression signatures of independently derived human embryonic stem cell lines. Hum. Mol.Genet. *13,* 601–608.

Bozdech, Z., Zhu, J., Joachimiak, M.P., Cohen, F.E., Pulliam, B., and DeRisi, J.L. (2003). Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol. *4*, R9.

Brandenberger, R., Khrebtukova, I. ,Thies. R.S., Miura, T., Jingli, C., Puri, R., Vasicek, T., Lebkowski, J., and Rao, M. (2004). MPSS profiling of human embryonic stem cells. BMC Dev. Biol. *4*, 10.

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.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E., Young, R.A. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431,* 99–104.

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 Database Issue*, D447–453.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. (2000). Functional discovery via a compendium of expression profiles. Cell *102*, 109–126.

Natarajan, K., Meyer, M.R., Jackson, B.M., Slade, D., Roberts, C., Hinnebusch, A.G., Marton, M.J. (2001). Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. Mol. Cell Biol. *21*, 4347–4368.

Nishimoto, M., Fukushima, A., Okuda, A., and Muramatsu, M. (1999). The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2. Mol. Cell Biol. *19*, 5453–5465.

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 Database Issue*, D501–504.

Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R.A., and Dynlacht, B.D. (2002). E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev. *16*, 245–256.

Sato, N., Sanjuan, I.M., Heke, M., Uchida, M., Naef, F., and Brivanlou, A.H. (2003). Molecular signature of human embryonic stem cells and its
comparison with the mouse. Dev Biol. *260*, 404–413.

Solter, D., and Knowles, B.B. (1979). Developmental stage-specific antigens during mouse embryogenesis. Curr. Top. Dev. Biol. *13 Pt 1*, 139–165.

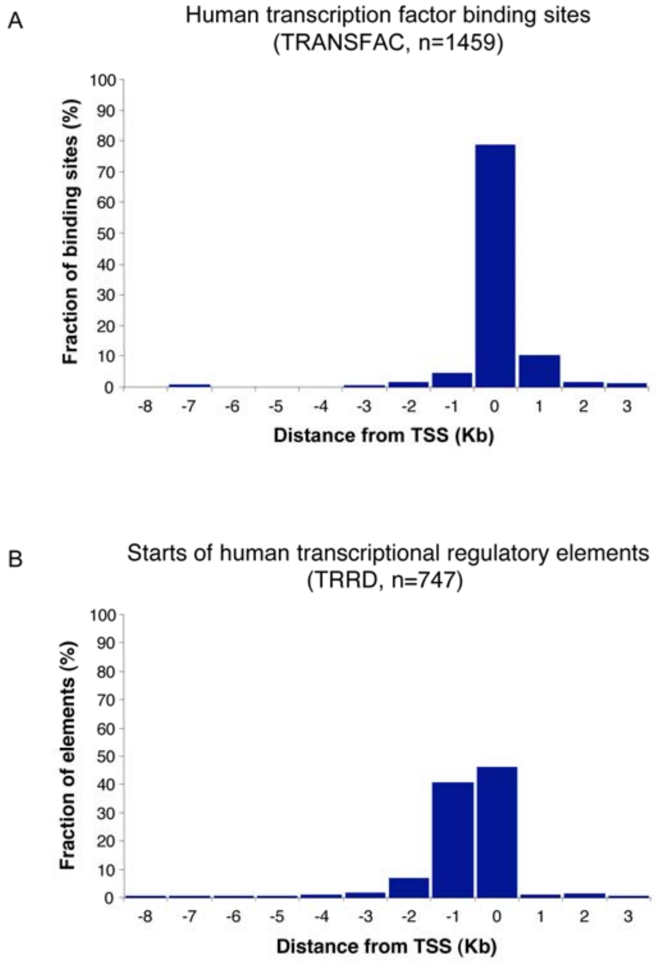
Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke. M.P., Walker, J.R., and Hogenesch, J.B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl .Acad. Sci. USA *101*, 6062–6067.

Wei, C.L., Miura, T., Robson, P., Lim, S.K., Xu, X.Q., Lee, M.Y., Gupta, S., Stanton, L., Luo, Y., Schmitt, J., Thies, S., Wang, W., Khrebtukova, I., Zhou, D., Liu, E.T., Ruan, Y.J., Rao, M., and Lim B. (2005). Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state. Stem Cells *23*, 166–185.

Weinmann, A.S., Yan, P.S. Oberley, M.J. Huang, T.H., Farnham, P.J. (2002). Isolating human transcription factor targets by coupling chromatin immunoprecipiation and CpG island microarray analysis. Genes Dev. *16*, 235–244.

Yuan, H., Corbi, N., Basilico, C., and Dailey, L. (1995). Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. Genes Dev. *9*, 2635–2645.

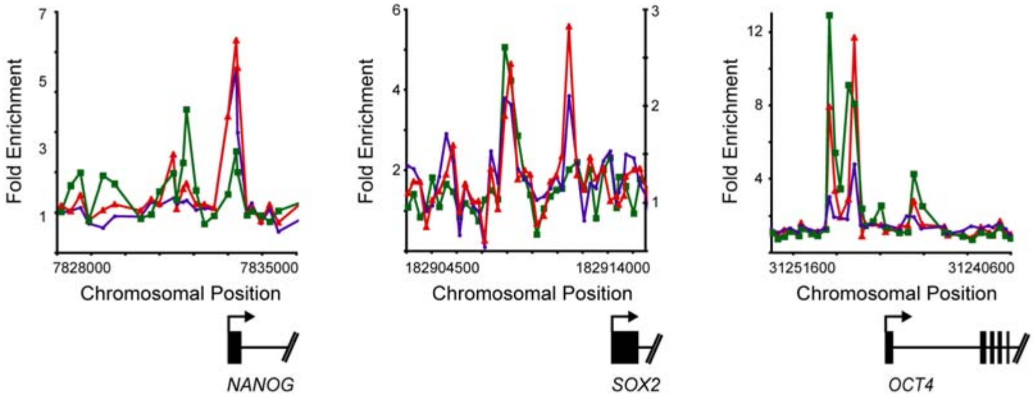
Figure S1. Distribution of Transcription Factor Binding Sites and Transcriptional Regulatory Elements Relative to Transcription Start Sites



(A) Distribution of transcription factor binding sites from TRANSFAC from –8kb to +3kb around the transcription start site.

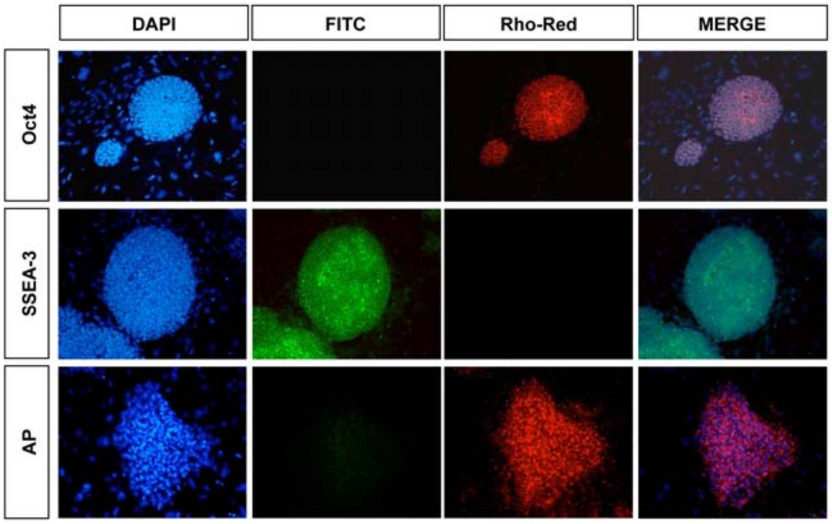
(B) Distribution of functional regulatory elements from the TRRD (database of transcriptional regulatory regions, http://www.bionet.nsc.ru/trrd/34/) from –8kb to +3kb around the transcription start site.

Figure S2. Oct4, Sox2, and Nanog Cooccupy Each of Their Promoters



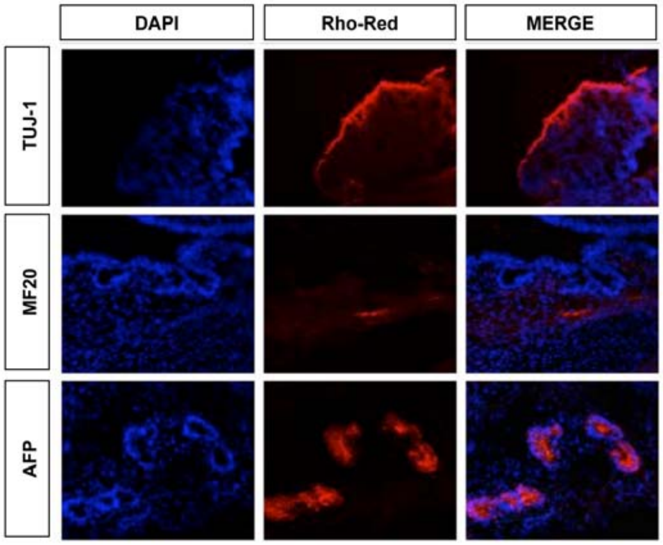
Plots display unprocessed ChIP enrichment ratios for all probes within a genomic region. Genes are shown to scale relative to their chromosomal position. Exons and introns are represented by thick vertical and horizontal lines, respectively. The start and direction of transcription are denoted by arrows. Green, red, and purple lines represent Nanog, Sox2, and Oct4 bound regions, respectively.

Figure S3. Immunuohistochemical Analysis of Pluripotency Markers



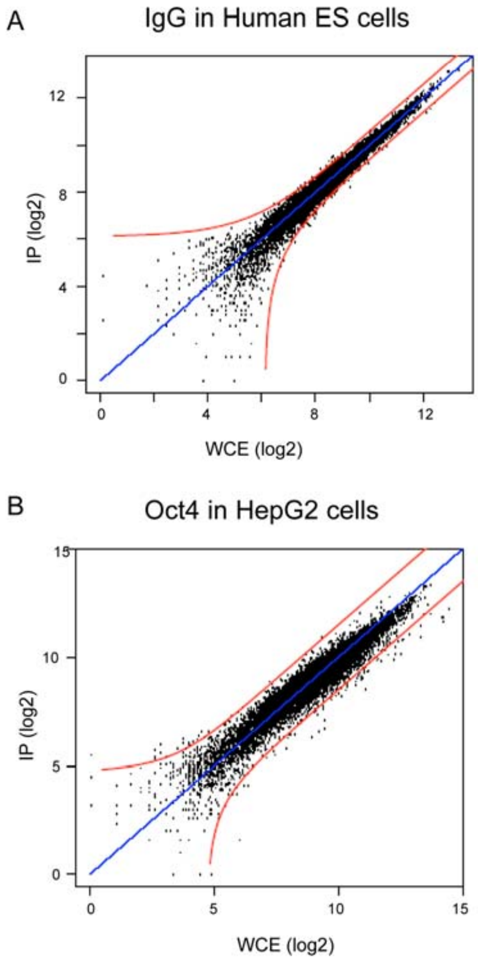
Human ES cells were analyzed by immunohistochemistry for the characteristic pluripotency markers Oct4 and SSEA-3. For reference, nuclei were stained with DAPI. Our analysis indicated that >>80% of the colonies were positive for Oct4 and SSEA-3. Alkaline phosphatase activity was also strongly detected in hES cells.

Figure S4. H9 Cells Maintain Differentiation Potential in Teratoma Assay



Teratomas were analyzed for the presence of markers for ectoderm (Tuj1), mesoderm (MF20) and endoderm (AFP). For reference, nuclei are stained with DAPI. Antibody reactivity was detected for derivatives of all three germ layers confirming that the human embryonic stem cells used in our analysis have maintained differentiation potential.

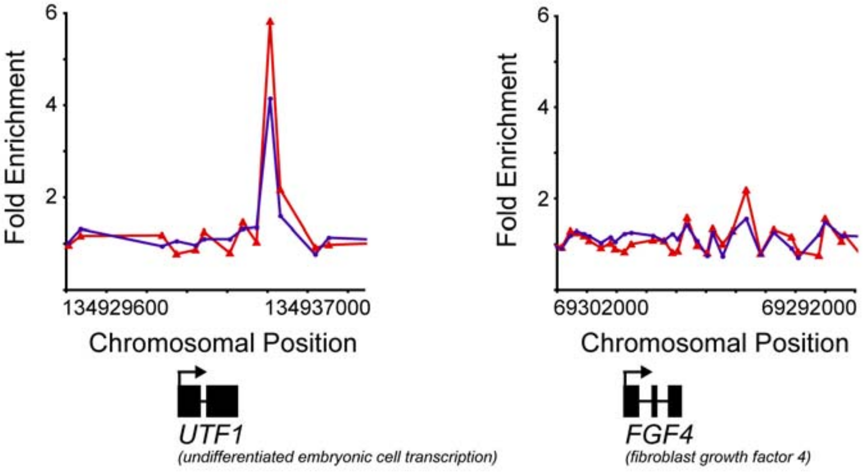
Figure S5. Control Chromatin Immunoprecipitations



(A) Oct4, Sox2, and Nanog targets were not enriched using preimmune sera in human ES cells. ChIP was carried out using rabbit or goat IgG to assess antibody specificity. Labeled IP material and control DNA were hybridized to self-printed promoter arrays. Background subtracted normalized log2 intensities are plotted. Red lines represent enrichment / exclusion p-values of $<10^{-3}$. Example shown is for the goat IgG control experiment.

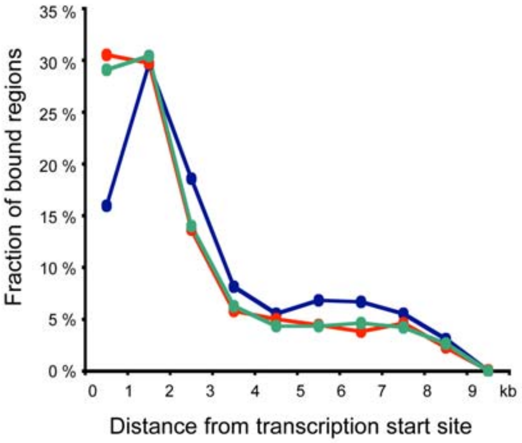
(B) Potential antibody cross-reactivity with other family members was assessed by performing ChIP experiments in HepG2 cells. Data were analyzed as above. Example shown for Oct4 (sc-9081) in HepG2.

Figure S6. Oct4 and Sox2 Binding to UTF1 and FGF4



Plots display unprocessed ChIP enrichment ratios for all probes within a genomic region. Genes are shown to scale relative to their chromosomal position. Exons and introns are represented by thick vertical and horizontal lines, respectively. The start and direction of transcription are denoted by arrows. Green, red, and purple lines represent Nanog, Sox2, and Oct4 bound regions, respectively.

Figure S7. Distribution of Oct4, Sox2, and Nanog Bound Regions Relative to Transciption Start Sites



Histogram of the distance between transcription factor bound regions and the nearest transcription start site. Green, red, and purple lines represent Nanog, Sox2, and Oct4 bound regions, respectively. A distance of 0 refers to bound regions that overlap a transcription start site.