SUPPLEMENTAL DATA

Supplemental Text

Index of Supplemental Tables

Index of Supplemental Figures

Supplemental References

Supplementary Text

Growth Conditions for Human Embryonic Stem Cells

Human embryonic stem (ES) cells were obtained from WiCell (Madison, WI; NIH Code WA09) and grown as described (Cowan et al., 2004). Briefly, passage 34 cells were grown in KO-DMEM medium supplemented with serum replacement, basic fibroblast growth factor (FGF), recombinant human leukemia inhibitory factor (LIF) and a human plasma protein fraction. Detailed protocol information on human ES cell growth conditions and culture reagents are available at http://www.mcb.harvard.edu/melton/hues.

In order to minimize any MEF contribution to our analysis, H9 cells were cultured on a low density of irradiated murine embryonic fibroblasts (ICR MEFs) resulting in a ratio of approximately >8:1 H9 cell to MEF (Figure S1). The culture of H9 on low-density MEFs had no adverse affects on cell morphology, growth rate, or undifferentiated status as determined by immunohistochemistry for pluripotency markers (e.g. Oct4, SSEA-3, Tra-1-60; see below). In addition, H9 cells grown on a minimal feeder layer maintained the ability to generate derivates of ectoderm, mesoderm, and endoderm upon differentiation (see below).

Quality Control for Human Embryonic Stem Cells

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 S2). After a brief wash in PBS, cells were 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). Data is shown for Oct4 and SSEA-3. Our analysis indicated that >90% 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 Substrate 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 10 μm 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 S3). Nuclei were stained blue with 4',6-diamidine-2-phenylidole dihydrochloride (DAPI). Antibody reactivity was detected for markers of all three germ layers confirming that the human embryonic 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 on 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 (FGF) 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 have maintained differentiation potential (data not shown; results similar to those shown in Figure S3).

Antibodies

RNA polymerase II-bound genomic DNA was isolated from whole cell lysate using 8WG16, a mouse monoclonal antibody (Thompson et al., 1989). This antibody preferentially binds a form of RNA polymerase II that lacks phosphorylation at the C-terminal domain of the largest subunit of polymerase (Patturajan et al., 1999; Cho et al., 2001; Jones et al., 2004) although this preference can be subject to experimental conditions.

Suz12-bound genomic DNA was isolated from whole cell lysate with a Suz12 rabbit polyclonal antibody purchased from Upstate (07-379).

Eed-bound genomic DNA was isolated from whole cell lysate using an Eed mouse monoclonal antibody previously described (Hamer et al., 2002).

H3-K27me3-bound genomic DNA was isolated from whole cell lysate using rabbit polyclonal antibody purchased from Abcam (AB6002). Chromatin immunoprecipitations against H3K27me3 were compared to reference DNA obtained by chromatin immunoprecipitation of total histone H3 (Abcam AB1791; epitope derived from C-terminal 100 amino acids of histone H3) to normalize for nucleosome density.

Chromatin Immunoprecipitation

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

We performed independent immunoprecipitations for each whole-genome analysis. Human WA09 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. 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 power 7 for 10 x 30 second pulses (90 second pause between pulses). 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 preincubated with approximately 10 µg of the appropriate antibody. For cases where suppliers did not provide information regarding antibody concentration, 20 µl of the supplied solution was used per reaction. The immunoprecipitation was allowed to proceed overnight.

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 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 labeled with Cy3 fluorophore).

Labeled DNA was mixed (5-6 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. Whole genome arrays were hybridized in batches of 35 to 60 arrays.

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.

Array Design

Whole genome arrays

We designed a set of 115 60-mer oligonucleotide arrays to cover the non-repeat masked region of the sequenced human genome. Arrays were produced by Agilent Technologies (www.agilent.com).

Selection of regions and design of subsequences
We tiled the genome with variable density: transcription units (defined below) were tiled with higher density and non-transcription regions were tiled with a slightly lower density.

To define transcription units, we first selected transcripts from 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 also added all microRNAs from the Rfam database (Griffiths-Jones et al., 2003) and a small set of collected non-coding RNAs (manual selection).

The entire collection of transcripts was sorted by chromosomal order. We then extended each transcript 10 kb upstream to capture proximal promoter regions. Each of these extended transcripts was considered a "transcription unit". In cases where one or more transcription units overlapped, we merged the transcription units into a single, larger unit. We extracted DNA sequence for all transcription units. Separately, we extracted intervening genomic DNA ("intergenic units") between transcription units. All sequences and coordinates are from the May 2004 build of the human genome (NCBI build 35), using the repeatmasked (-s) option.

We then separated sequences into subsequences in order to efficiently process sequences for oligo selection. We first removed all unmasked regions 100 bp or smaller. The small size of these regions makes it more difficult to identify high quality oligos for use on the array. These small regions represented a small fraction of the genome and were often covered by neighboring probes designed against larger subsequences. For unmasked regions that were 101 to 300 bp long, we treated each strand (Watson and Crick) as a separate subsequence. This ensured that we would have two oligos to represent these subsequences if the region could not be covered by neighboring 60-mers. For regions that were 301 to 640 bp long, we divided the region into two, evenly sized subsequences. Unmasked regions greater than 640 bp were divided into evenly sized subsequences such that no individual subsequence was greater than 320 bp.

We used the program ArrayOligoSelector (AOS)(Bozdech et al., 2003) to score 60-mers for use on the array, but modified the oligo selection process. We had two primary reasons for this. First, AOS uses a relative quality scale in selecting oligos. For any particular subsequence, it generates scores based on four parameters to evaluate each 60-mer in the subsequence and looks for the best oligos within that set, ignoring the absolute quality of the oligo. As a result, lower quality oligos can be selected. Second, AOS does not have a parameter to set distance between oligos. Consequently, resolution is largely set by defining subsequence size but is still subject to highly variable placement within each

subsequence. For instance, if the desired tiling density is 300 bp, we would select subsequences 300 bp long. For any two adjacent subsequences, probes could be separated by as little as 0 bp (both probes were placed near the shared subsequence border) or as much as 480 bp (both probes placed at opposite subsequence ends).

To avoid selecting lower quality oligos, we ran AOS to derive scores for every 60-mer in all subsquences and then eliminated oligos based on these scores. AOS uses a scoring system for four criteria: GC content, self-binding, complexity and uniqueness. 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.

To achieve more uniform tiling, we instituted a method to find probes within a particular distance from each other for the transcription unit subsequences. We sorted all qualified probes into chromosomal order and identified gaps in the genomic sequence that were not covered by one or more 60-mers. These gaps typically represented regions that were repeat masked or generated regions of consistently low quality oligos. For our purposes, gaps that were greater than 640 bp long represented potential dead zones or "borders". Based on empirical experience with genome-wide location analysis technology, we conservatively estimated that we would not identify binding events that occurred more than 320 bp away from the genomic location of any particular probe. As a result, gaps that were longer than 640 bp long likely contained one or more basepairs within the gap that would not be detected even if we used the closest qualified oligos as probes. Using these borders, we split the set of all probes into "packages" containing all qualified probes between two borders.

For packages up to 300 bp long, we designed two probes where possible, one from each strand (Watson and Crick). This resulted in two different probes in the region, compensating for those instances where a small region would be found isolated by two borders from the nearest, potentially informative, neighboring probe. For packages greater than 301 bp long, we selected the first qualified probe in the package (lowest chromosomal coordinate), 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. If the most recently selected probe was within 250 bp of the next border, we automatically selected the qualified probe closest to the next border. This ensured that we were selecting probes as close to the ends of packages as possible.

For intergenic unit tiling, we generated subsequences and identified borders and packages as described for genic tiling. We divided packages into evenly sized segments where the maximum segment size was 480 bp. We then selected the qualified probe closest to the midpoint of each segment.

All probes from both transcription unit and intergenic unit tiling were combined and grouped by chromosome and sorted by position.

Compiled Probes and Controls

The design process described above led to the production of a set of 115 Agilent microarrays containing a total of 4,652,484 features. Each array contains 40,457 features except for array #115, which contains 40,386 features. The probes are arranged such that array 1 begins with 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. We estimate that only 10% of the total, non-repeat masked region is not covered by probes. As an estimate of probe density, 95% of all 60-mers are within 450 bp of another 60-mer; 80% of all 60-mers are within 350 bp of another 60-mer.

We added several sets of control probes (1,500 total) to the whole genome array designs. On each array, there are 40 oligos designed against five *Arabidopsis thaliana* genes that are printed in triplicate, and thus available for use with spike-in controls. These *Arabidopsis* oligos were BLASTed against the human genome and do not register any significant hits. 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 1,500 controls, there are 2,256 controls added by Agilent (standard) and 77 blank spots.

Promoter Array

This set of 10 arrays was designed to cover regions between -8 kb and +2 kb relative to the transcription start sites of 16,710 genes. See Boyer et al. (2005) for details of the design of the arrays.

Transcription Factor Array

This array was designed to cover regions between –5 kb and +5 kb relative to the transcription start sites of 2,288 human genes encoding transcription factors as determined by GO classifications and manual annotation. Probes were designed essentially as

described above for the whole genome array although tiling density was slightly improved (1 probe approximately every 250 bp). There are a total of 2,079 control spots on the transcription factor array. The 40 Arabidopsis oligos and 80 E2F4 oligos described above for the whole genome design are each printed once. A total of 404 intensity controls are printed twice. A total of 1,085 "gene desert" controls (described above in the whole genome design) are each printed once. The intensity controls and "gene desert" controls are expanded sets of the controls described above for the whole genome design.

Data Normalization and Analysis

We used GenePix software (Axon) 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 the set of 1,500 control probes described above and included on each array. For multiple slide sets (whole genome and promoter array), we then calculated the average of these median intensities for all 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.

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 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. 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.

Identification of Bound Regions

Whole Genome Arrays

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 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 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. For RNA polymerase II, this algorithm identified 22,912 bound probe sets of RNA polymerase II ChIP-enriched DNA across the genome.

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. This final step reduced the 22,912 peaks to 10,244 bound regions. The bound regions had a median size of 950 bp.

The ES cell line we used (H9) has a female karyotype (XX). Nineteen (0.18%) of the RNA polymerase II bound regions mapped to the Y chromosome and 6 of these correspond to the promoters of known genes. Each of these 6 genes (ASMTL, CXYorf2, HIT000024005, PLCXD1, PPP2R3B and SYBL1) are also present on the X chromosome suggesting that all of these bound regions are duplicate measurements of X chromosome binding events caused by hybridization of X chromosome DNA to Y chromosome probes. Subtracting out these duplicates leaves 10,225 unique genomic regions bound by RNA polymerase II in ES cells.

Peak finding for genome-wide Suz12 binding data was carried out as described above for RNA polymerase II with the following modifications. Probe sets were marked as potentially bound if the p-value of average X (probe set p-values) was less than 0.0001 and 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 < 0.0005 or the center probe in the probe set has a single probe p-value < 0.0001 and one of the flanking probes has a single point p-

value < 0.01. This algorithm identified 16,438 bound probe sets of Suz12 ChIP-enriched DNA across the genome. As before, 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 1,000 bp of each other. This final step reduced the 16,348 peaks to 3,446 bound regions. The bound regions had a median size of 1,248 bp.

Unlike RNA polymerase II, Suz12 was often associated with large regions of DNA stretching over multiple kilobases of contiguous sequence. 28% of Suz12-bound regions were over 2 kb in size, compared with only 7% of RNA polymerase II-bound regions. In some instances, multiple large regions were clustered in close proximity as shown for the Hox clusters.

Promoter Array
Probe sets were marked as potentially bound if the p-value of average X (probe set p-values) was less than 0.001 and 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. This algorithm identified 7,074 bound probe sets of Suz12 ChIP-enriched DNA, 6,302 bound probe sets of Eed ChIP-enriched DNA and 8,205 bound probe sets of H3K27me3 ChIP-enriched DNA. As before, 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 1,000 bp of each other. This final step reduced the peaks to 1,415 (Suz12), 1,549 (Eed) and 1,885 (H3K27me3).

Transcription Factor Array
Probe sets were marked as potentially bound if the p-value of average X (probe set p-values) was less than 0.001 and 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. As before, 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 1,000 bp of each other. This algorithm identified 465 bound probe sets (299 bound regions) of Suz12 ChIP-enriched DNA in muscle cells, 7,199 bound probe sets (645 bound regions) of Suz12 ChIP-enriched DNA in ES cells, 1,375 bound probe sets (455 bound regions) of H3K27me3 ChIP-enriched DNA in muscle cells and 5,455 bound probe sets (775 bound regions) of H3K27me3 ChIP-enriched DNA in ES cells.

Controlling for the Effect of Murine Embryonic Fibroblast Feeder Cells

We performed two sets of experiments to measure the contribution of the murine embryonic fibroblasts (MEFs) to the RNA polymerase II binding data. In the first experiment, we grew a population of MEFs isolated from E13.5 embryos, irradiated and replated the cells for 24 hours, treated the cells with formaldehyde to crosslink polymerase to DNA and performed a chromatin IP. This DNA was then purified and labeled exactly as described for samples of ES cells. Labeled DNA was hybridized to self-printed arrays and analyzed as described previously (Odom et al., 2004). The results indicate that mouse

feeder cells are unlikely to contribute more than 1% false positives to RNA polymerase II chromatin immunoprecipitation results. Using our standard analysis, there are only 47 features that show enrichment with the mouse feeder cells RNA polymerase II chromatin immunoprecipitation. In contrast, there are typically 4,000-5,000 enriched features with human RNA polymerase II chromatin immunoprecipitation on self-printed arrays. In the second set of experiments, we obtained ES cells that were MEF-subtracted by preplating the cells on ungelatinized culture dishes for 1-2 hours at 37°C. The supernatant enriched for ES cells was then cross-linked as above and harvested for immunoprecipitation. The results were essentially the same with and without feeder cells. There are some differences, presumably due to the extra manipulations needed to separate the cells and the decreased cell number resulting from these manipulations. While it is technically possible that the oligonucleotide arrays perform differently from our self-printed arrays, these experiments generally suggest that the contribution of 8-12% of feeder cells is unlikely to have an effect on the final results.

Comparing Bound Regions to Known and Predicted Genes

The coordinates for the complete lists of RNA polymerase II-bound and Suz12-bound regions on the whole-genome arrays can be found in Table S1 and Table S7, respectively. Mapping the location of RNA polymerase II using genome-tiling arrays directly identified the physical location of active promoters in living cells, thus improving our confidence in transcription start sites previously inferred from RNA evidence. Mapping the location of Suz12 identified the location of genomic regions targeted by the chromatin regulator PRC2. This knowledge should be valuable for improving annotation of the genome and identifying regulatory elements that may not be detected by alternative methods.

Comparisons to Known Genes

The locations of RNA polymerase II-bound and Suz12-bound regions were compared relative to transcript start and stop coordinates of known genes compiled from five different databases: RefSeq (Pruitt et al., 2005), Mammalian Gene Collection (MGC) (Gerhard et al., 2004), Ensembl (Hubbard et al., 2005), University of California Santa Cruz (UCSC) Known Genes (genome.ucsc.edu)(Kent et al., 2002) and Human Invitational (H-Inv) full-length cDNAs (Imanishi et al., 2004). All coordinate information was downloaded in January 2005 from the UCSC Genome Browser (NCBI build 35). Of the 10,225 RNA polymerase II-bound regions, 6,741 (66%) occurred within 1 kb of gene starts from one of these 5 databases (Table S1). Of the 3,446 Suz12-bound regions, 2,113 (61%) occurred within 1 kb of gene starts from one of these 5 databases (Table S7).

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 Genes and H-Inv transcripts. Transcripts for which no EntrezGene annotation could be found in this manner were annotated manually. This resulted in a total of 7,106 EntrezGene genes being bound by RNA polymerase II (Table S2) and 1,893 EntrezGene genes being bound by Suz12 (Table S8).

The Distribution of Distances to Known Genes
The distances between each bound region and the closest RefSeq, Ensembl or MGC transcription start site were calculated and plotted as a histogram (Text: Figure 1E). As might be expected for RNA polymerase II, there is a higher frequency of binding events over the start sites of known genes. This distribution gradually tails off in both directions as the distance to the start site increases. Suz12 shows a similar, but broader distribution as a sizable subsetset of Suz12 binding events cover large regions of the genome. For comparison, the same distance calculation was made for all probes on chromosome 1.

Fraction of transcription start sites bound by RNA polymerase II or Suz12 in ES cells
We used several human gene databases to identify the fraction of annotated transcription start sites bound by RNA polymerase II or Suz12 in ES cells (Figure S4). For each database, we calculated the percentage of annotated transcription start sites that lie within 1 kb of a bound region (RNA polymerase II: MGC 42%, RefSeq 34%, Ensembl 28%, UCSC Known Genes 26% and H-Inv 26%; Suz12: MGC 6%, RefSeq 8%, Ensembl 7%, UCSC Known Genes 6% and H-Inv 4%).

Comparisons to Predicted Genes
The locations of bound regions were also compared relative to transcript start and stop coordinates of predicted genes compiled from eight different databases; GenScan (Burge and Karlin, 1997), GeneID (Parra et al., 2000), FirstEF (Davuluri et al., 2001), ACEview (www.aceview.org), ECgene (Kim et al., 2005), UniGene (www.ncbi.nlm.nih.gov/UniGene), UCSC RetroFinder (Kent et al., 2003) and Non-human mRNAs (Kent et al., 2002). These gene models are generally derived through *ab initio* computational gene modeling (GenScan, GeneID and FirstEF) or EST clustering and alignment to the human genome (ACEview, ECgene, UniGene, UCSC RetroFinder and Non-human mRNAs). All predictions were derived from downloads of coordinates of predicted human genes mapped to NCBI build 35 of the public human genome sequence from UCSC in January 2005. Of the 3,484 RNA polymerase II-bound regions not mapping to a known gene, 2,110 mapped within 1 kb of the start site of a predicted gene (Table S1). Therefore, a total of 8,851 (87%) RNA polymerase II-bound regions corresponded to a known or predicted transcription start site. Of the 1,353 Suz12-bound regions not mapping to a known gene, 1,158 mapped within 1 kb of the start site of a predicted gene (Table S7). Therefore, a total of 3,271 (95%) Suz12-bound regions corresponded to a known or predicted transcription start site.

Candidate Novel Genes
We reasoned that RNA polymerase II bound regions located outside of known genes and relatively far from known transcription sites might represent novel genes. We identified 1,053 genomic regions bound by RNA polymerase II that lie over 10 kb away from and outside of any known gene (as defined as being present in any one of RefSeq, MGC, Ensembl, UCSC Known Genes or H-Inv databases) (Table S3). Of these, we calculated that 432 occur within 1 kb of the transcription start sites predicted by one or more of eight gene prediction algorithms (Table S4). These gene predictions are made based on *ab initio* computational gene modeling (GenScan, GeneID and FirstEF), EST clustering and alignment (ACEview (www.aceview.org), ECgene, UniGene (www.ncbi.nlm.nih.gov/UniGene), UCSC RetroFinder and Non-human mRNAs). All predictions were derived from downloads of coordinates of predicted human genes mapped to NCBI build 35 of the public human genome sequence from UCSC in January 2005.

While we favor the interpretation that RNA polymerase II-bound regions that are relatively distant from annotated transcript start sites represent promoters for novel genes, there are several other possibilities. These regions could also represent new, distal start sites for known genes. These distal regions might represent enhancers that are captured via long-range interactions with RNA polymerase II bound to proximal promoters. Finally, these regions could also represent regions that are spatially, but not linearly, co-localized, similar to the localization of separate regions of chromosomes in the nucleolus.

Promoters for miRNAs

RNA polymerase II and Suz12 were also found associated with microRNA genes in ES cells. MicroRNAs (miRNAs) are a non-coding class of small RNAs with significant regulatory potential (Bartel, 2004). In a few cases, miRNA primary transcripts have been characterized and shown to have the hallmarks of RNA polymerase II transcripts (Lee et al., 2004), but due to the rapid processing of primary miRNA transcripts, the location and classification of the majority of miRNA promoters remains unknown. We found RNA polymerase II associated with genes specifying 66 miRNAs in human ES cells, representing 29% of all annotated miRNAs (Table S5). RNA polymerase occupied the promoters of protein-coding genes harboring 35 intronic miRNAs, strengthening the proposal that miRNAs located within protein-coding genes are typically regulated by the promoters of the corresponding host genes. We also identified the promoters of 31 miRNAs that occur independently of protein-coding genes, providing global evidence that independently transcribed miRNAs are generally RNA polymerase II transcripts. This systematic identification of miRNA genes bound by RNA polymerase II overcomes many of the limitations to miRNA detection such as the small size of the mature species and the cross-hybridization of closely related miRNAs.

Similar analysis for Suz12-bound regions indicated that Suz12 binds the promoter regions of 34 miRNAs. These included mir-124, a miRNA preferentially expressed in brain tissue (Sempere et al., 2004) that can shift gene expression profiles towards that of brain (Lim et al., 2005). The observation that Suz12 occupies genes that specify both transcriptional and post-transcriptional regulators of development indicates that PRC2 functions to repress developmental transcriptional programs in ES cells at multiple levels.

Bound regions were assigned to miRNAs as follows. MiRNA clusters (data from Rfam, May 2005) were divided into two classes; intronic (inside known genes in the same orientation) and independent. Intronic miRNA genes were classified as bound if the promoter of their host gene was bound. For genes with alternative promoters, a promoter upstream of the miRNA had to be bound. Intronic miRNAs appeared to be transcribed from the promoters of their host genes; we did not observe any other RNA polymerase II binding close to intronic miRNAs. Intergenic miRNAs were classified as bound if RNA polymerase II or Suz12 binding was identified within 10 kb upstream of the miRNA, unless the bound region could be attributed to a neighboring gene. However, in most cases, the bound region was detected much closer to the DNA encoding the miRNA stem loops.

Comparing Binding and Human Expression Data

Transcription of genes bound by RNA polymerase II and Suz12
We collected 7 previously published ES cell expression datasets for comparison with our RNA polymerase II and Suz12 binding data. The expression data, gathered using massively parallel signature sequencing (MPSS) and Affymetrix gene expression arrays, were processed as follows:

MPSS data: Three MPSS datasets were collected, two from a pool of the ES cell lines H1, H7 and H9 (Brandenberger et al., 2004; Wei et al., 2005) and one for HES-2 (Wei et al., 2005). For each study, only MPSS tags detected at or over 4 transcripts per million (tpm) were used. In addition, the data provided by Wei and colleagues (Wei et al., 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 (Sato et al., 2003), H9, HSF1 and HSF6 (Abeyta et al., 2004). Each cell line was analyzed by the authors in triplicate. EntrezGene IDs were assigned to the probe-sets using Affymetrix annotation or using RNA accession numbers provides by the authors. For each probe-set, we counted the number of "Present" calls in the three replicate array experiments performed for each cell line. Most genes are represented by more than one probe-set and, to enable comparison to MPSS and RNA polymerase II binding data, we then found the maximum number of P calls for each gene (defined by unique EntrezGene ID). A gene was defined as detected if it was called "Present" in at least 2 of the 3 replicate arrays.

This provided 7 lists of genes expressed in ES cells, 3 from MPSS data and 4 from microarray data. We found that microarray analysis of H9 ES cells detected transcripts for 78% of the genes bound by RNA polymerase II in H9 cells that were present on the Affymetrix arrays. In total, the 7 expression experiments detected transcripts for 88% of genes bound by RNA polymerase II (Table S6). In contrast to genes bound by RNA polymerase II, the expression of genes bound by Suz12 was detected more rarely (Figure S12). We found that 20% (±6%) of the genes bound by Suz12 alone in H9 ES cells were expressed, depending on the expression dataset used. The expression of some of these genes may be due to the incomplete shut down of transcription by Suz12, variations in the genes bound by Suz12 in different cell culture conditions, or due to the detection of RNA transcripts that are present in a minority of differentiated cells. Transcription of genes bound by both Suz12 and RNA polymerase II is detected substantially more often than genes bound by Suz12 alone, consistent with the presence of RNA polymerase II.

ES cell expression relative to differentiated cell types
We examined the relative expression levels of genes associated with PRC2 and H3K27me3 in human ES cells (Text: Figure 2C). 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). We extracted data from the original CEL files from each array and scaled the data to a median signal of 150 in GCOS (Affymetrix). We then exported the data, created expression ratios using the median gene expression of each gene across all arrays, transformed the data into log base2 and median centered both gene and arrays (so that the median log2 expression ratio for each gene and each array is 0). EntrezGeneIDs were assigned to each probe-set and for genes with multiple probe-sets, the expression ratios averaged. This resulted in a set of 12,968 unique genes. Of these, 604 were bound by Suz12, Eed and H3K27me3 at high confidence.

In addition to examining the relative expression levels of Suz12-target genes in ES cells and differentiated cells, we also examined the Affymetrix absolute Present/Absent expression calls (Figure S12). Using this measurement, we found that RNA transcripts of Suz12-target genes were detected in ES cells much less frequently for RNA transcripts of RNA polymerase II target genes. However, in differentiated cells, RNA transcripts were detected for the two classes of genes more equally, indicating that many of the genes silenced by Suz12 in ES cells are transcriptionally active in differentiated cells.

Inverse correlation between the size of the Suz12 binding domain and gene expression
We found that, unlike RNA polymerase II, Suz12 was often associated with large regions of DNA stretching over multiple kilobases of contiguous sequence. For example, 28.3% of Suz12-bound regions were over 2 kb in size, compared with only 6.6% of RNA polymerase II-bound regions (Figure S8). To explore whether the size of the genomic region occupied by Suz12 had any functional implications, we measured how RNA polymerase II co-occupancy and gene expression varied according to Suz12 coverage (Figure S13). Suz12 bound regions were assigned to RefSeq genes if they occurred within 1 kb of a transcription start site. For genes associated with multiple bound regions, the regions were collapsed, unless the bound regions occupied alternative promoters, in which case the largest region was selected. Then for each gene, we determined whether the gene was co-occupied by RNA polymerase II and whether or not the gene was transcribed. Genes have to pass one of two criteria to be classified as transcribed: either RNA transcripts could be detected in all three MPSS datasets or RNA transcripts could be detected in all four Affymetrix gene expression microarray datasets. We discovered that the greater the extent of Suz12 binding, the less frequently the gene was transcribed and the less frequently the target gene was occupied by RNA polymerase II. Genes associated with Suz12 over 4 kb of sequence were 8-times less likely to be transcribed in ES cells (from 24% of RefSeq genes to 3%) and 4-times less likely to be bound by RNA polymerase II (from 36% to 9%). This suggests that transcriptional repression of genes is facilitated by the presence of Suz12 across large regions of DNA.

Expression changes upon ES cell differentiation
We also compared the expression level of genes between pluripotent ES cells and differentiated ES cells (expression data from Sato et al, 2003). The pluripotent ES cells (H1 cell line) were cultured on Matrigel in MEF-conditioned medium and then differentiated (non-lineage directed) on Matrigel in non-conditioned medium for 26 days and both samples were analyzed in triplicate on Affymetrix HG-U133A arrays. We extracted data from the original CEL files and scaled the data to a median signal of 150 in GCOS (Affymetrix). We then exported the data and, for each probe-set, calculated the ratio

of the average signal in differentiated cells to the average signal in pluripotent cells. EntrezGeneIDs were assigned to each probe-set and for genes with multiple probe-sets, the expression ratios averaged. We then selected only those genes that had transcripts detectable in either pluripotent or differentiated ES cells (gene called "P" in at least 2 of the 3 replicates), to avoid analyzing expression ratios consisting of only noise.

To test whether Suz12 bound genes were preferentially upregulated upon differentiation (Text: Figure 6, Table S13), we compared the distribution of expression ratios for genes bound by Suz12 but not RNA Pol II with the distribution of expression ratios for all genes. As a control, we also compared the distribution of expression ratios for genes bound by neither Suz12 nor RNA Pol II (i.e. genes repressed by other means) with the distribution of expression ratios for all genes. We chose to present data for genes not bound by RNA polymerase II because this was the stricter comparison (genes bound by RNA polymerase II are less likely to increase in expression as they are already being transcribed). However, the preferential induction of genes bound by Suz12 is also apparent without first filtering for RNA polymerase II occupancy (data not shown).

Estimating Error Rates

We used sequence-specific PCR to estimate false positive rates for the whole-genome array data (Figure S5). For RNA polymerase II, a subset of the bound probe sets were selected and primer pairs designed to amplify between 100 and 200 bp within each bound probe set. Primers were tested for specificity using BLAST and ePCR. A total of 192 primer pairs were selected, where each primer had 10 or fewer matches to the genome and the pair predicted a single amplicon. For RNA polymerase II IP samples, 10 ng of immunoenriched DNA was used as input to the PCR. For whole cell extract (WCE) samples, a range of unenriched DNA amounts (90, 30 and 10 ng of DNA) was used. The PCR was performed for 28 cycles and products were visualized on an agarose gel stained with SYBR Gold (Amersham) and quantified using ImageQuant (Amersham). Only PCR reactions giving single bands with intensities ordered according to the WCE concentration were used. Genomic regions were considered enriched if the 10 ng IP sample showed either 1.5-fold or greater enrichment compared to the 30 ng WCE sample or greater than 1-fold enrichment compared to the 90 ng WCE sample. Genomic regions were considered not enriched if the band intensity of the 10 ng IP was less than half that of the 30 ng WCE or less than the 10 ng WCE. A total of 119 primer pairs yielded a clear enriched/not enriched decision. 114 of these showed enrichment, indicating a false positive rate of 4.4%. Using this set of PCR results, we were also able produce receiver-operator curves showing how changes in peak identification criteria would affect the false positive and false negative rates. The results suggest that our selected criteria are useful for maximizing the identification of true positives.

Two lines of evidence suggest that the false negative rate is approximately 30%. Estimating a false negative rate is generally much more problematic than measuring a false positive rate because the measurement of a false negative rate assumes perfect knowledge of the true positives in the dataset. As every method will have its own error rate, determining a set of true positives is challenging, if not impossible. Despite this important caveat, we have used both sequence-specific PCR and a comparison with expression datasets to estimate a false negative rate.

To obtain an estimate of the false positive rate for our sequence specific PCR reactions, we designed 49 primer pairs against regions of the genome that had no indication of RNA polymerase II binding (p-value for average X and center probes X > 0.3) despite being in densely tiled regions. We reasoned that any substantial PCR amplification in this region was more likely to reflect a false positive in the PCR then to reflect binding of a very large fraction of the genome to the initiation form of RNA polymerase II. From these PCRs, we measured a false detection rate of ~9%. We then designed a series of PCR primers against probes 'expressing' a broad range of p-values between these absolute negatives and our positive list. 60 of these pairs produced positive PCR amplifications. Correcting for the expected false detection rate of the PCR, we calculate a probe based false negative rate of ~33%.

We also used sequence-specific PCR to estimate false positive and false negative rates for the whole-genome Suz12 array data. For estimating false positives, a total of 108 primer pairs yielded a clear enriched/not enriched decision. 105 of these showed enrichment, indicating a false positive rate of 2.8%. Correcting for the expected false detection rate of the PCR, we calculated a probe based false negative rate of 27%.

Binding of Suz12, Eed and H3K27me3

We used a microarray containing probes for the promoters of 16,710 genes to measure the correlation between Suz12 binding, Eed binding and H3K27 methylation. This array detected binding of Suz12 to 1,039 genes, Eed to 909 genes and H3K27me3 to 1,007 genes. (Text: Figure 2A, Table S9). Due to the strict significance threshold we use to call define a DNA binding event (see Identification of Bound Regions section), any set of genes we define as being bound is conservative, with a false negative rate of ~30% (see Estimating Error Rates section). We therefore compared the binding ratios between Suz12, Eed and H3K27me3 to determine whether the genes that were only called bound by one factor were also bound by the other factors, although at a significance that fell below our strict threshold (Figure S6). For genes bound by any one of Suz12, Eed or H3K27me3, we aligned the binding ratios from our Suz12 IP, our Eed IP and our H3K27me3 IP. We found that the binding patterns of Suz12, Eed and H3K27me3 followed one another, even at genes where the binding of only one factor was highly significant by our analysis. From this we conclude that Suz12, Eed and H3K27me3 are present at essentially the same set of genes in ES cells, although we cannot rule out that there is specific binding by these factors at a small number of genes.

The high degree of overlap between the Suz12, Eed, and H3K27me3 targets indicates that Suz12 defines an active PRC2 complex at these genes. As a critical subunit of the PRC2 complex, Suz12 has widely accepted roles in euchromatic gene silencing and dosage compensation, where Suz12 and H3K27me3 are transiently enriched on the Xi during X-inactivation (Plath et al., 2003; Silva et al., 2003; de la Cruz et al., 2005). However, alternative roles for Suz12 have been proposed that suggest Suz12 may function independently of PRC2 and H3K27me3. For example, Suz12 mutations are suppressors of position-effect variegation (PEV) and can interact with the heterochromatin protein 1α (HP1α), indicating a role in heterochromatin-linked gene silencing (Birve et al., 2001;

Yamamoto et al., 2004). Suz12 is also required for germ cell development independent of other PcG proteins and can exhibit different protein expression profiles compared to Eed and EZH2 (Birve et al., 2001; de la Cruz et al., 2005). While the vast majority of Suz12 co-localizes with Eed to regions of H3K27 methylation, non-overlapping targets may be representative of these alternative Suz12 roles that are independent of other PcG proteins.

GO Classification for RNA Polymerase II and Suz12 Bound Genes

We identified Gene Ontology classification terms (http://www.geneontology.org) enriched for RNA polymerase II-bound and Suz12-bound genes (defined as being within 1 kb of an annotated TSS in either the RefSeq, MGC or Ensembl databases). Hypergeometric distributions were calculated to determine enriched terms, using for reference the total number of genes annotated to that GO term. Categories with p-values $< 10^{-5}$ are indicated in Table S10.

Many of the classifications enriched for Suz12-bound genes were related to development, transcriptional regulation and signaling and are further described in the main text. Among the remaining enriched classifications, we noted an additional category of interest. Over 100 ion channel genes are bound by Suz12 (L-type calcium channels, voltage-gated and inward rectifying potassium channels). This is consistent with a role for PRC2 in blocking differentiation. L-type calcium channels are involved in the neural vs epidermal cell fate decision and direct activation of these channels results in neural induction (Moreau et al., 1994; Leclerc et al., 2001).

We identified 252 annotated human homeodomain transcription factors using PFam (Bateman et al., 2002) and EntrezGene. Of these, 150 (60%) were bound by Suz12. Most of these were associated with extended domains of Suz12 binding. Given the considerable number of homeodomain transcription factors bound by Suz12, we searched for other families of transcription factors enriched in the set of Suz12-bound targets. Genes annotated to the molecular function GO terms GO:003700 (transcription factor activity), GO:0030528 (transcription regulator activity), or GO:003705 (RNA polymerase II transcription factor activity/enhancer binding); or the biological process GO terms GO:006355 (regulation of transcription, DNA-dependent), or GO:0045449 (regulation of transcription) were defined as transcription factors. Suz12-bound genes in this set for which a SwissProt ID could be retrieved were input into the PANDORA software package (Kaplan et al., 2003) using domain annotation to search for enriched molecular domains at standard resolution. For reference, the same analysis was performed for transcription factor genes bound by RNA polymerase II. Results from the first level of classification are depicted in Figure S7, expressed as a percentage of the number of total transcription factors placed in that category by PANDORA.

Comparing Suz12-Bound Regions to domains of conservation

Regions of genomic conservation were obtained from the PhastCons database stored at UCSC (http://genome.ucsc.edu). PhastCons identifies genomic segments of conservation based on a two-state phylogenetic hidden Markov model with a state for conserved regions and a state for non-conserved regions. Each conserved element is assigned a log-odds

score equal to its log probability under the conserved model minus its log probability under the non-conserved model. The elements are then assigned a conservation score, which is a log transformation of the log-odds score and scales from 0 to 1000 (http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=phastConsElements). A LoD score of 100 corresponds to a conservation score of ~500. Conserved elements overlapping exons in the Refseq and UCSC Known Gene database were removed for most analyses. For cases in which Suz12-bound regions or TSS proximal regions (-8kb to +2kb around known start sites) contained multiple conserved elements, the top conservation score was used.

To calculate the significance of overlap between Suz12 binding and conserved domains, we tested randomly generated genomic regions that reflected the variation in size of Suz12 binding regions and the array coverage of the genome. The randomizations were performed by finding, for each bound region, a random region of equal size in the genome that was present on the array. Each of these random regions was then tested to see if it overlapped a conserved element and scored as above. Multiple runs of the randomization were performed and P-values were determined by assuming a binomial distribution with an expectation derived from the randomized regions. For comparison, the same analysis was performed for RNA polymerase II.

Generating Suz12-deficient Mouse Cells and Analysis of their Expression Pattern

Generation of Suz12 -/- mouse cells

To generate Suz12-deficient mouse cells, a targeting vector was constructed (Figure S10) from a BAC DNA clone containing the Suz12 gene isolated from a mouse genomic library derived from R1 embryonic stem cells (O. Ohara and H.Koseki, unpublished). The targeting construct had two homology sequences, a 6.1 kb EcoRI/XbaI fragment that lies 5' of the 11th exon of the locus and a 3.1 kb KpnI/XbaI fragment that lies 3' of the 14th exon and extends just past the stop codon, removing DNA encoding amino acids 482 – 741 containing the VEFS domain which is involved in interactions with EZH2 (Yamamoto et al., 2004). For the negative selection, the HSV tk cassette from pPNT vector was added. Successful integration replaced a 6.0 kb fragment containing four exons with a neomycin-resitance (Neor) gene cassette from pHR68 (gifted by Dr. T. Kondo) in a reverse orientation relative to Suz12 transcription. This vector was introduced into R1 embryonic stem cells as described previously (Akasaka et al., 1996) and four homologous recombinants were obtained. Suz12 heterozygous ES cells were introduced into recipient blastocysts and germline transmission of the null allele was obtained with no apparent phenotypic differences between wildtype and heterozygous animals. Suz12 +/- mice were backcrossed six times onto a C57BL/6. Genotyping was performed by Southern blotting against BamHI-digested genomic DNA to detect the appearance of a 3.5 kb fragment (generated by cutting at a BamHI site introduced with the neo cassette) and the loss of a 6.2 kb fragment that would occur with endogenous BamHI sites (Figure S10).

Suz12 -/- cell lines were derived from blastocysts from crosses between heterozygous Suz12 mutant animals based on conventional protocols (Hogan et al., 1994). Loss of wild-type Suz12 protein in Suz12 -/- cells was confirmed by Western blotting (Figure S10). The homozygous mutant cells display reduced ability to tri-methylate H3K27 (data not shown) indicating that PRC2 complex function is disrupted in these cells. The cells retain

some characteristics of ES cells, such as cellular morphology, relatively normal levels of Oct4 and Nanog expression and the ability to proliferate in culture, while gaining some characteristics of differentiated cells, such as upregulation of developmental transcription factors (as described below). Moreover, Suz12 -/- embryos in this study arrest development at 7.75 dpc, similar to that as previously described for Suz12, Eed, and Ezh2 null embryos (Schumacher et al., 1996; O'Carroll et al., 2001; Pasini et al., 2004).

Microarray expression analysis
Total RNA was purified from the two replicate wild-type mouse ES cell lines and the replicate Suz12 -/- cell lines using TRIzol. RNA from each Suz12 -/- cell line was labeled with Cy5 using the Low RNA Input Fluorescent Linear Amplification Kit (Agilent) and hybridized to Mouse Development Arrays (G4120A, Agilent) with Cy3 labeled total RNA from the corresponding wild-type cells. Each experiment was also repeated, swapping the dyes, giving a total of four expression datasets, with each of the two biological replicates being represented by two technical replicates. The arrays were scanned with an Agilent microarray scanner and the processed signals and expression ratios were obtained using Feature Extraction software (Agilent). We filtered the data to remove features with signal intensities not significantly above background in both channels. Average expression ratios were generated through inter-slide and intra-slide comparisons between the signals for Suz12 -/- cells and wild-type ES cells for each replicate. The average ratios between the self-self comparisons within each replicate set were also calculated and this population was then defined as the null-distribution. Expression ratios were then compared to this null-distribution and the number of standard deviations from the mean calculated. The expression of a gene was considered to be significantly altered in Suz12 -/- cells if the expression ratio between Suz12 -/- and wild-type ES cells was over 2 standard deviations from the mean of the null-distribution and the expression ratio for the same gene in the self-self comparisons was less than 1 standard deviation from the mean of the null distribution.

Comparing mouse expression data with human binding data and human expression data
We reasoned that genes bound by Suz12 in human ES cells have orthologs in mice that should be upregulated in Suz12 -/- mouse cells. A significant overlap in the genes bound by Suz12 in human ES cells and the genes upregulated in Suz12 -/- mouse cells would support a role for Suz12 in the repression of its target genes in ES cells. We expected that the overlap in genes bound by Suz12 in human ES cells and genes upregulated in Suz12 -/- mouse cells would be incomplete because of 1) potential differences in Suz12 occupancy in human and mouse ES cells, 2) possible repression of PRC2 target genes by additional mechanisms 3) the effects of the Suz12 -/- on genes downstream of Suz12-target genes, due to the fact that many of these are transcriptional regulators and 4) false positive and negative errors in both binding and expression analysis.

The mouse microarrays contained 5341 features with a mouse EntrezGene ID. Features representing duplicate genes were averaged, giving single expression values for 4266 unique genes. Using the Agilent GeneName field, we then used the Homologene database to identify orthologous human genes. Homologene listed a human ortholog for 3971 of the 4266 mouse genes. We determined that 557 of these 3971 genes with human orthologs were significantly upregulated in Suz12 -/- mouse cells.

We compared the set of 3971 human-mouse gene orthologs with the human Suz12 binding data and found that 346 of these 3971 genes were bound by Suz12 in human ES cells. By comparing the set of 346 bound genes with the set of 557 upregulated genes, we found that 70 (20%) of the 346 genes bound by Suz12 in human ES cells were upregulated in Suz12 -/- cells. This overlap is significant ($6x10^{-4}$) and given the complexities associated with human-mouse comparisons, strongly supports a role for Suz12 in the repression of its target genes in ES cells. Strikingly, 8 of the 10 most upregulated developmental transcription factors in Suz12 -/- cells were bound by Suz12 in human ES cells. The identities of the genes bound by Suz12 in human ES cells and upregulated in Suz12 -/- mouse cells are listed, together with their expression changes, in Table S14.

To determine the degree to which Suz12 bound genes were preferentially upregulated in Suz12 -/- cells (Text: Figure 6C), we performed the same analysis previously used to determine whether Suz12 bound genes were preferentially upregulated upon human ES cell differentiation (p.16 of Supplemental Data).

To compare the expression changes that occur in Suz12 -/- cells with those that occur upon human H1 ES cell differentiation (Text: Figure 6D), we identified a set of 182 genes that were present in both filtered datasets and were bound by Suz12 in human ES cells (the human ES cell differentiation dataset was filtered as described on page 16).

Sample Preparation and Analysis of Differentiated Muscle

Primary Human Skeletal Muscle Cells (HSkMCs) were obtained from Cell Applications, Inc. (San Diego, CA) and expanded according to supplier's protocols in growth medium. Upon reaching confluence, cells were shifted to differentiation medium in plates coated with collagen to promote attachment of differentiating cells, and medium was replaced every 2 days. Growth and differentiation media were supplied by Cell Applications, Inc. After 6 days of differentiation, cells had fused to form multinucleated myotubes. Cells were crosslinked and ChIP experiments were performed as described above. ChIP-chip data were analyzed as described above.

We observed Suz12 binding at a number of loci that were also Suz12-bound in ES cells, but loss of Suz12 and H3K27me3 was seen at several genes critical for the development of differentiated muscle tissue, including MyoD, Pax3, Pax7, and Six1. Previous work indicates Ezh2 is removed from the chromatin of the muscle-specific structural genes MCK and MHCII and replaced by transcriptional activators upon differentiation (Caretti et al., 2004). While Ezh2 levels appear to decline over the course of muscle differentiation, we note Suz12 binding in differentiated muscle tissue. We cannot rule out that Suz12 may be playing a different role in this terminally differentiated tissue than it plays in ES cells, and further work will be necessary to clarify this issue. However, the loss of Suz12 at factors necessary for development of muscle tissue and the retention of Suz12-binding at a number of genes important for other lineages suggests that removal of Suz12 accompanies the expression of key developmental regulators during cellular differentiation.

Comparing Suz12 Binding with Oct4, Nanog and Sox2 Binding

To explore how Suz12 might be targeted to genes, we compared our Suz12 binding data to our previous results from profiling the DNA binding of the ES cell transcription factors Oct4, Sox2 and Nanog (Boyer et al., 2005). We found that there was a significant overlap between the genes bound by Suz12 and the genes bound by Oct4, Sox2 or Nanog. Of the 1606 genes bound by Suz12 and present on the promoter arrays we used previously for Oct4, Sox2 and Nanog, 196 (12%) were also bound by Oct4, 148 (9%) by Sox2 and 271 (17%) by Nanog. 92 genes (6%) were bound by all three factors and a total of 342 (21%) by any one of the three factors. We found that genes bound by Suz12 and Oct4, Suz12 and Sox2 or Suz12 and Nanog were enriched for genes encoding transcriptional regulators of development when compared to genes bound by Suz12 alone ($p = 5.4x10^{-20}$, $3.5x10^{-10}$ and $2.4x10^{-18}$, respectively). There were 315 genes encoding developmental transcription factors that were present on the whole-genome and promoter arrays and bound by Suz12. Of these, 103 (33%) were also bound by Oct4, 81 (26%) were also bound by Sox2 and 107 (34%) were also bound by Nanog. 57 genes (18%) were bound by all three factors and a total of 143 (45%) were bound by any one of the three factors (Table S11).

Although Oct4, Sox2 and Nanog are all indispensable for ES cell propagation, mutations in each regulator display slightly different phenotypes (reviewed in Chambers, 2004 and Boiani and Scholer, 2005) suggesting each may have unique contributions to stem cell identity. This led us to ask if there were differences in the association of Oct4, Sox2 or Nanog with Suz12 when the factors were considered individually or in pairs. Direct analysis of the bound sites, compared to randomized binding data, showed that Oct4 was more associated with Suz12 bound regions then either Sox2 or Nanog. This association was consistent whether we compared the factors alone or in pairs. Surprisingly, sites bound by Sox2 or Nanog but not Oct4 were not particularly associated with Suz12 (Figure S14). These data point out subtle differences in the association of Oct4, Sox2 or Nanog with Suz12 that may eventually help identify regulatory mechanisms specific to each transcription factor.

Suz12, Oct4, Nanog and Sox2 binding and sequence conservation

The observation that a set of repressed genes bound by Oct4, Sox2 and Nanog were occupied by Suz12 and contain highly conserved non-coding DNA sequences led us to examine whether the DNA-binding regulators occupy conserved sequence motifs that might contribute to PRC2 targeting at these genes. When we examined the Oct4, Sox2 and Nanog binding sites at developmental TFs bound by Suz12, we found that ~50% of their bound regions overlapped conserved elements that had LoD conservation scores > 100. The association of these DNA-binding transcription factors with conserved elements at a substantial fraction of Suz12 occupied sites suggests that they may have some role in targeting PRC2 to conserved elements in many genes encoding developmental regulators. Oct4, Sox2 and Nanog were not found at all PRC2-occupied genes, however, so additional regulators must also be involved in PRC2-mediated silencing in ES cells.

Suz12, Oct4, Nanog and Sox2 binding and DNA motifs

If Oct4, Sox2 and Nanog are involved in targeting PRC2 to genes, we expect that other factors must influence Suz12 binding and thus explain why Suz12 is observed at only a subset of the genes bound by these transcription factors. We used MEME (Bailey and Elkan, 1995) to search for additional DNA sequence motifs that might discriminate

between genes bound by Oct4, Sox2, Nanog and Suz12 and genes bound by only Oct4, Sox2 and Nanog. There was one motif consisting of repeats of the dinucleotide GT that was specifically associated with the Oct4, Sox2, Nanog and Suz12-bound sites (Figure S15). This motif is similar to one previously associated with polycomb response elements in Drosophila (Ringrose et al., 2003). We also found DNA elements that were specifically associated with sites bound by Oct4, Sox2 and Nanog and not Suz12 (Figure S15). One of these elements contains DNA binding sites for multiple transcription factors, including HoxA3 and C/EBP. This suggests that Oct4, Sox2 and Nanog may act with other transcriptional regulators to positively regulate transcription at some genes, but in the absence of these other regulators, may recruit PcG proteins and thus negatively regulate transcription at other genes. Similar bimodal activities have been suggested for proteins involved in PcG targeting in Drosophila including GAGA and zeste (Kerrigan et al., 1991; Laney and Biggin, 1992; Strutt et al., 1997). Oct4, Sox2 and Nanog have previously been described as having both positive and negative roles in transcription (Yuan et al., 1995; Botquin et al., 1998; Nishimoto et al., 1999; Guo et al., 2002) and the association of Suz12 with only a subset of promoters bound by these regulators would be consistent with these observations.

Index of Tables

All tables can be found on the supporting website; the URLs below can be used to download the appropriate table.

Table S1. Regions bound by RNA polymerase II and their relationship to known and predicted genes.
http://web.wi.mit.edu/young/hES_PRC/TableS1.xls

Table S2. HUGO/EntrezGene identifiers for RNA Pol II bound, annotated genes.
http://web.wi.mit.edu/young/hES_PRC/TableS2.xls

Table S3. RNA polymerase II-bound regions that predict novel gene candidates.
http://web.wi.mit.edu/young/hES_PRC/TableS3.xls

Table S4. Gene models bound by RNA polymerase II.
http://web.wi.mit.edu/young/hES_PRC/TableS4.xls

Table S5. MicroRNA genes bound by RNA polymerase II and Suz12 in ES cells.
http://web.wi.mit.edu/young/hES_PRC/TableS5.xls

Table S6. Expression of genes bound by RNA polymerase II in ES cells.
http://web.wi.mit.edu/young/hES_PRC/TableS6.xls

Table S7. Regions bound by Suz12 and their relationship to known and predicted genes.
http://web.wi.mit.edu/young/hES_PRC/TableS7.xls

Table S8. HUGO/EntrezGene identifiers for Suz12-bound, annotated genes.
http://web.wi.mit.edu/young/hES_PRC/TableS8.xls

Table S9. Detection of Suz12, Eed and H3K27me3 occupancy using promoter arrays.
http://web.wi.mit.edu/young/hES_PRC/TableS9.xls

Table S10. Enriched gene ontologies among RNA Pol II-bound and Suz12-bound genes.
http://web.wi.mit.edu/young/hES_PRC/TableS10.xls

Table S11. Developmental transcription factors bound by Suz12.
http://web.wi.mit.edu/young/hES_PRC/TableS11.xls

Table S12. Developmental signaling proteins bound by Suz12.
http://web.wi.mit.edu/young/hES_PRC/TableS12.xls

Table S13. Expression of Suz12-bound genes during ES cell differentiation.
http://web.wi.mit.edu/young/hES_PRC/TableS13.xls

Table S14. Genes bound by Suz12 in ES cells and upregulated in Suz12 -/- mouse cells.
http://web.wi.mit.edu/young/hES_PRC/TableS14.xls

Table S15. Developmental regulators associated with PRC2 in ES cells and muscle.
http://web.wi.mit.edu/young/hES_PRC/TableS15.xls

Figure Legends

Figure S1. Human H9 ES cells cultured on a low density of irradiated murine embryonic fibroblasts.
Bright-field image of H9 cell culture.

Figure S2. Analysis of human ES cells for markers of pluripotency.
Human embryonic stem 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 >90% of the ES cell colonies were positive for Oct4 and SSEA-3. Alkaline phosphatase activity was also strongly detected in human ES cells.

Figure S3. Analysis of human ES cells for differentiation potential.
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 S4. The fraction of annotated promoters bound by RNA polymerase II or Suz12.
The fraction of unique gene transcription start sites that lie within 1 kb of a genomic region bound by RNA polymerase II and Suz12. The total number of start sites in each database is as follows: MGC n=17,188; RefSeq n=19,349; Ensembl n=30,121; UCSC Known Genes n=42,160; H-Inv n=42,777.

Figure S5. Estimating error rates.
a. Example gel images showing PCR products amplified from 16 genomic regions judged to be bound by RNA polymerase II using the whole-genome arrays. Each primer-pair was used to amplify unenriched, whole cell extract (WCE) DNA (90, 30 and 10 ng) and immunoenriched (IP) DNA (10 ng). Enrichment in the IP DNA is indicated by a "+" and a lack of enrichment by a "-". PCR reactions judged to be inconclusive were labeled with an "N".
b. Example gel images showing PCR products amplified from 16 genomic regions judged not to be bound by RNA polymerase II using the whole-genome arrays. Each genomic region represents an annotated transcription start site.
c. Receiver-operator curve for RNA polymerase II binding in human ES cells. Curve compares percentage of true positives and false positives in binding events called from ChIP/chip compared to RT-PCR amplifications of anti-Pol II ChIP DNA. ROC curves were determined for all regions of the genome (blue) and for the subset of regions located within 1 kb of known transcription start sites (red).

Figure S6. Co-occupation of gene promoters by Suz12, Eed and H3K37me3.
Suz12 occupancy (top panel), Eed occupancy (middle panel) and H3K27me3 occupancy (bottom panel) at transcription start sites. Each row represents a gene considered occupied by either Suz12, Eed or H3K27me3 using our high-confidence gene calling algorithm (see sections on Data Normalization and Analysis and Identification of Bound Regions). The same genes are illustrated in each of the three panels. Each column represents the data from an oligonucleotide probe positioned relative to the start site as indicated by the gene

diagram below. The log binding ratios for each oligo are plotted for each protein; blue indicates enrichment of the immunoprecipitated factor (enrichment ratio >1). A scale for the binding ratios for each panel is shown. Each factor follows the same binding pattern. From this we conclude that Suz12, Eed and H3K27me3 are present at essentially the same set of genes and that our stringent gene calling algorithm sometimes calls a gene bound by one factor but not another factor because of the inherent false negative rate of ~30% (see Estimating Error Rates section).

Figure S7. Protein domain classification of Suz12- and Pol II-bound transcription factors.
Fraction of transcription factor categories bound by Suz12 (green) or RNA Pol II (blue). The percentage is expressed relative to all transcription factor genes assigned to that category by InterPro domain (PANDORA) annotation at the default resolution. Abbreviations are b-HLH (basic helix-loop-helix), NHR (nuclear hormone receptor), ETS (erythroblast transformation specific), b-Zip (basic leucine zipper), PHD finger (plant homeodomain finger), SMAD (Sma- and Mad-related) and FHA (forkhead-associated). n indicates the number of transcription factor genes assigned to a given category.

Figure S8. Suz12 occupies large regions of DNA.
Number of RNA polymerase II (blue bars, left hand axis) and Suz12 (green bars, right hand axis) bound regions of certain sizes (x axis). Unlike RNA polymerase II, Suz12 occupies over 2 kb of sequence at a significant number of genes.

Figure S9. H3K27me3 co-occupies large domains with Suz12.
a. Correlation between size of domains of Suz12 binding and H3K27me3 binding. The trend was calculated by computing the moving average of the size of H3K27me3 regions using a sliding window of 20 genes across the set of genes bound by Suz12 and H3K27 and ordered by size of Suz12 bound region. Sizes of bound regions were calculated from promoter arrays.
b. Binding profile of H3K27me3 (black) across ~500 kb regions encompassing Hox clusters A-D. Unprocessed enrichment ratios for all probes within a genomic region are shown (ChIP vs. whole genomic DNA). Approximate Hox cluster region sizes are indicated within black bars.

Figure S10. Generation of Suz12 -/- cells.
a. Targeted deletion of the Suz12 locus. Homologous recombination was used to replace the 5' portion of Suz12 with a neo cassette. Location of probe used for southern blot verification in (b) is shown. Restriction enzymes are denoted B, BamH1; E, EcoR1; X, Xba1.
b. Southern blot analysis of BamH1 digested genomic DNA from each genotype.
c. Western blot analysis of whole cell extracts derived from each genotype. Immunoblots were probed with anti-Suz12 (top) or anti-Lamin B (bottom).
d. Embryos generated from Suz12 heterozygous crosses were analyzed at different stages of development. At 7.75 dpc, normal as well as morphologically smaller embryos were evident. Genotyping analysis indicated that the abnormal embryos were homozygous for the Suz12 null allele confirming that Suz12 is required for early development.

Figure S11. Binding of Suz12 in differentiated muscle.
a. Suz12 binding profiles across the muscle regulator MYOD1 gene in H9 human ES cells

(green) and differentiated myotubes (grey). The plots show unprocessed enrichment ratios for all probes within a genomic region (ChIP vs. whole genomic DNA). Genes are shown to scale below plots (exons are represented by vertical bars). The start and direction of transcription are noted by arrows.
b. H3K27me3 profiles across the muscle regulator MYOD1 gene in H9 human ES cells (black) and differentiated myotubes (blue). The plots show unprocessed enrichment ratios for all probes within a genomic region (ChIP vs. whole genomic DNA). Genes are shown to scale below plots (exons are represented by vertical bars). The start and direction of transcription are noted by arrows.
c. Suz12 binding profiles across the muscle regulator PAX3 gene, as in **a.**
d. H3K27me3 profiles across the muscle regulator PAX3 gene, as in **b.**
e. Suz12 binding profiles across the muscle regulator PAX7 gene, as in **a.**
f. H3K27me3 profiles across the muscle regulator PAX7 gene, as in **b.**

Figure S12. Detection of genes bound by RNA polymerase II and Suz12 in human ES cell expression datasets.
Percentages of genes that are bound by RNA polymerase II only, RNA polymerase II and Suz12, and Suz12 only that are detected in 7 ES cell expression datasets and one differentiated cell expression dataset. The first four ES cell datasets and the differentiated cell dataset were generated using gene expression arrays (H1: U133A arrays (Sato et al., 2003); H9, HSF1 and HSF6: U133A+B arrays (Abeyta et al., 2004); differentiated tissues: U133A arrays (Su et al., 2004)). The percentages are relative to the fraction of bound genes that are represented on the arrays. The last three ES cell datasets were generated using MPSS (Brandenberger et al., 2004; Wei et al., 2005).

Figure S13. Relationship between size of Suz12 and RNA polymerase II co-occupancy and gene expression.
The percentage of genes with detectable RNA (grey bars) and associated with RNA polymerase II (blue bars) as a function of the extent of Suz12 binding. The frequencies for genes not bound by Suz12 are indicated on the left as controls.

Figure S14. Association of Oct4, Sox2 or Nanog with Suz12-bound regions.
a. The percentage of Oct4-bound regions (purple arrow), Sox2-bound regions (red arrow) or Nanog-bound regions (green arrow) that overlap with Suz12-bound regions are shown along the x-axis. Comparisons were made between promoter array data from Boyer et al., 2005 and whole genome Suz12 data presented here. The dashed line indicates the distribution of the expected overlap based on randomized data. For comparison, we also show the results for a fourth transcription factor, E2F4 (blue arrow).
b. The percentage of Sox2 and Oct4-cobound regions or Nanog and Oct4-bound regions (purple arrows) that overlap with Suz12-bound regions are shown along the x-axis. Comparisons were made between promoter array data from Boyer et al., 2005 and whole genome Suz12 data presented here. The dashed line indicates the distribution of the expected overlap based on randomized data. For comparison, we also show the results for Sox2-bound regions that are not bound by Oct4 (red arrow) and Nanog-bound regions that are not bound by Oct4 (green arrow).

Figure S15. Motifs associated with DNA regions that are bound by Oct4, Sox2, Nanog and Suz12 or bound by Oct4, Sox2 and Nanog.
a. Consensus sequence of a motif associated with DNA regions bound by Oct4, Sox2, Nanog and Suz12. This motif was found in approximately 50% of the regions bound by Oct4, Sox2, Nanog and Suz12 and enriched 4.8-fold compared to regions bound by Oct4, Sox2 and Nanog but not Suz12.
b. Consensus sequence of a motif associated with DNA regions bound by Oct4, Sox2 and Nanog and not bound by Suz12. This motif was found in approximately 20% of the regions bound by Oct4, Sox2, Nanog and enriched 3.0-fold compared to regions bound by Oct4, Sox2, Nanog and Suz12. Putative transcription factor binding sites are labeled and indicated by black lines. Binding sites were identified with P-Match (http://www.gene-regulation.com) using the input sequence CCTGTAATCCCAGC and cut-off selection for matrix group to minimize the sum of false positives and negatives.
c. Consensus sequence of a motif associated with DNA regions bound by Oct4, Sox2 and Nanog and not bound by Suz12. This motif was found in approximately 15% of the regions bound by Oct4, Sox2, Nanog and enriched 2.4-fold compared to regions bound by Oct4, Sox2, Nanog and Suz12. No putative transcription factor binding sites were identified when examined as described in **b.** using the input sequence ATCTCGGCTCACTG. More lenient selections for the cut-off selection indicate potential binding sites for C/EBP, HoxA3, CdxA, Msx-1 and v-Myb.

Supplementary 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.

Akasaka, T., Kanno, M., Balling, R., Mieza, M. A., Taniguchi, M., and Koseki, H. (1996). A role for mel-18, a Polycomb group-related vertebrate gene, during theanteroposterior specification of the axial skeleton. Development *122*, 1513-1522.

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.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bateman, A., Birney, E., Cerutti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002). The Pfam protein families database. Nucleic Acids Res *30*, 276-280.

Birve, A., Sengupta, A. K., Beuchle, D., Larsson, J., Kennison, J. A., Rasmuson-Lestander, A., and Muller, J. (2001). Su(z)12, a novel Drosophila Polycomb group gene that is conserved in vertebrates and plants. Development *128*, 3371-3379.

Botquin, V., Hess, H., Fuhrmann, G., Anastassiadis, C., Gross, M. K., Vriend, G., and Scholer, H. R. (1998). New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. Genes Dev *12*, 2073-2090.

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.

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.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol *268*, 78-94.

Caretti, G., Di Padova, M., Micales, B., Lyons, G. E., and Sartorelli, V. (2004). The Polycomb Ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation. Genes Dev *18*, 2627-2638.

Cho, E. J., Kobor, M. S., Kim, M., Greenblatt, J., and Buratowski, S. (2001). Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. Genes Dev *15*, 3319-3329.

Cowan, C. A., Klimanskaya, I., McMahon, J., Atienza, J., Witmyer, J., Zucker, J. P., Wang, S., Morton, C. C., McMahon, A. P., Powers, D., and Melton, D. A. (2004). Derivation of embryonic stem-cell lines from human blastocysts. N Engl J Med *350*, 1353-1356.

Davuluri, R. V., Grosse, I., and Zhang, M. Q. (2001). Computational identification of promoters and first exons in the human genome. Nat Genet *29*, 412-417.

de la Cruz, C. C., Fang, J., Plath, K., Worringer, K. A., Nusinow, D. A., Zhang, Y., and Panning, B. (2005). Developmental regulation of Suz 12 localization. Chromosoma *114*, 183-192.

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.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. (2003). Rfam: an RNA family database. Nucleic Acids Res *31*, 439-441.

Guo, Y., Costa, R., Ramsey, H., Starnes, T., Vance, G., Robertson, K., Kelley, M., Reinbold, R., Scholer, H., and Hromas, R. (2002). The embryonic stem cell transcription factors Oct-4 and FoxD3 interact to regulate endodermal-specific promoter expression. Proc Natl Acad Sci U S A *99*, 3663-3667.

Hamer, K. M., Sewalt, R. G., den Blaauwen, J. L., Hendrix, T., Satijn, D. P., and Otte, A. P. (2002). A panel of monoclonal antibodies against human polycomb group proteins. Hybrid Hybridomics *21*, 245-252.

Hogan, B., Beddington, R., Constantini, F., and Lacy, E. (1994). Manipulating the Mouse Embryo: A Laboratory Manual, Cold Spring Harbor Press).

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.

Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M.*, et al.* (2004). Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol *2*, e162.

Jones, J. C., Phatnani, H. P., Haystead, T. A., MacDonald, J. A., Alam, S. M., and Greenleaf, A. L. (2004). C-terminal repeat domain kinase I phosphorylates Ser2 and Ser5 of RNA polymerase II C-terminal domain repeats. J Biol Chem *279*, 24957-24964.

Kaplan, N., Vaaknin, A., and Linial, M. (2003). PANDORA: keyword-based analysis of protein sets by integration of annotation sources. Nucleic Acids Res *31*, 5617-5626.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A *100*, 11484-11489.

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.

Kerrigan, L. A., Croston, G. E., Lira, L. M., and Kadonaga, J. T. (1991). Sequence-specific transcriptional antirepression of the Drosophila Kruppel gene by the GAGA factor. J Biol Chem *266*, 574-582.

Kim, P., Kim, N., Lee, Y., Kim, B., Shin, Y., and Lee, S. (2005). ECgene: genome annotation for alternative splicing. Nucleic Acids Res *33 Database Issue*, D75-79.

Laney, J. D., and Biggin, M. D. (1992). zeste, a nonessential gene, potently activates Ultrabithorax transcription in the Drosophila embryo. Genes Dev *6*, 1531-1541.

Leclerc, C., Rizzo, C., Daguzan, C., Neant, I., Batut, J., Auge, B., and Moreau, M. (2001). [Neural determination in Xenopus laevis embryos: control of early neural gene expression by calcium]. J Soc Biol *195*, 327-337.

Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. Embo J *23*, 4051-4060.

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S., and Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769-773.

Moreau, M., Leclerc, C., Gualandris-Parisot, L., and Duprat, A. M. (1994). Increased internal Ca2+ mediates neural induction in the amphibian embryo. Proc Natl Acad Sci U S A *91*, 12639-12643.

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.

O'Carroll, D., Erhardt, S., Pagani, M., Barton, S. C., Surani, M. A., and Jenuwein, T. (2001). The polycomb-group gene Ezh2 is required for early mouse development. Mol Cell Biol *21*, 4330-4336.

Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K.*, et al.* (2004). Control of pancreas and liver gene expression by HNF transcription factors. Science *303*, 1378-1381.
Parra, G., Blanco, E., and Guigo, R. (2000). GeneID in Drosophila. Genome Res *10*, 511-515.

Pasini, D., Bracken, A. P., Jensen, M. R., Denchi, E. L., and Helin, K. (2004). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. Embo J *23*, 4061-4071.

Patturajan, M., Conrad, N. K., Bregman, D. B., and Corden, J. L. (1999). Yeast carboxyl-terminal domain kinase I positively and negatively regulates RNA polymerase II carboxyl-terminal domain phosphorylation. J Biol Chem *274*, 27823-27828.

Plath, K., Fang, J., Mlynarczyk-Evans, S. K., Cao, R., Worringer, K. A., Wang, H., de la Cruz, C. C., Otte, A. P., Panning, B., and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. Science *300*, 131-135.

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.

Ringrose, L., Rehmsmeier, M., Dura, J. M., and Paro, R. (2003). Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster. Dev Cell *5*, 759-771.

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.

Schumacher, A., Faust, C., and Magnuson, T. (1996). Positional cloning of a global regulator of anterior-posterior patterning in mice. Nature *384*, 648.

Sempere, L. F., Freemantle, S., Pitha-Rowe, I., Moss, E., Dmitrovsky, E., and Ambros, V. (2004). Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. Genome Biol *5*, R13.

Silva, J., Mak, W., Zvetkova, I., Appanah, R., Nesterova, T. B., Webster, Z., Peters, A. H., Jenuwein, T., Otte, A. P., and Brockdorff, N. (2003). Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. Dev Cell *4*, 481-495.

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

Strutt, H., Cavalli, G., and Paro, R. (1997). Co-localization of Polycomb protein and GAGA factor on regulatory elements responsible for the maintenance of homeotic gene expression. Embo J *16*, 3621-3632.

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

Thompson, N. E., Steinberg, T. H., Aronson, D. B., and Burgess, R. R. (1989). Inhibition of in vivo and in vitro transcription by monoclonal antibodies prepared against wheat germ RNA polymerase II that react with the heptapeptide repeat of eukaryotic RNA polymerase II. J Biol Chem *264*, 11511-11520.

Wei, C. L., Miura, T., Robson, P., Lim, S. K., Xu, X. Q., Lee, M. Y., Gupta, S., Stanton, L., Luo, Y., Schmitt, J.*, et al.* (2005). Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state. Stem Cells *23*, 166-185.

Yamamoto, K., Sonoda, M., Inokuchi, J., Shirasawa, S., and Sasazuki, T. (2004). Polycomb group suppressor of zeste 12 links heterochromatin protein 1alpha and enhancer of zeste 2. J Biol Chem *279*, 401-406.

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.