Cell Stem Cell, Volume *7*

Supplemental Information

Chromatin Structure and Gene Expression Programs of Human Embryonic and Induced Pluripotent Stem Cells

Matthew G. Guenther, Garrett M. Frampton, Frank Soldner, Dirk Hockemeyer, Maya Mitalipova, Rudolf Jaenisch, and Richard A. Young

TABLE OF CONTENTS

Supplemental References

SUPPLEMENTAL FIGURES

SUPPLEMENTAL TABLES

A) Regions with different H3K4me3 between male and female pluripotent cells

B) Regions with different H3K27me3 between male and female pluripotent cells

C) Regions with different H3K4me3 between pluripotent and fibroblast cells

D) Regions with different H3K27me3 between pluripotent and fibroblast cells

E) Regions with different H3K4me3 between ES and iPS cells

F) Regions with different H3K27me3 between ES and iPS cells

Table S4: Chromatin differences between ES and iPS cells do not reflect cell of origin

A) H3K4me3 differences do not reflect cell of origin

B) H3K27me3 differences do not reflect cell of origin

Table S5: The probesets differentially expressed between ES and iPS cells and between pluripotent and fibroblast cells in this study, Chin et al., Maherali et al., and Yu et al. and the numbers of differentially expressed genes and probesets overlapping between these datasets

A) Differentially expressed probesets between ES and iPS cells in this study, Guenther et al.

B) Differentially expressed probesets between ES and iPS cells in Chin et al.

C) Differentially expressed probesets between ES and iPS cells in Maherali et al.

D) Differentially expressed probesets between ES and iPS cells in Yu et al.

E) Differentially expressed probesets between pluripotent and fibroblast cells in this study, Guenther et al.

F) Differentially expressed probesets between pluripotent and fibroblast cells in Chin et al.

G) Differentially expressed probesets between pluripotent and fibroblast cells in Maherali et al.

H) Differentially expressed probesets between pluripotent and fibroblast cells in Yu et al.

I) The numbers of differentially expressed genes and probesets overlapping between Guenther et al., Chin et al., Maherali et al., and Yu et al..

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

ES, iPS, and Fibroblast Cells and Cell Culture

All primary human fibroblasts cells described in this paper (PDB-AG20442 and GM-M01660) were purchased from the Coriell Cell Repository (Camden, NJ). Fibroblasts were cultured in fibroblast medium (Dulbecco's modified Eagle's medium [DMEM] supplemented with 15% fetal bovine serum [FBS; Hyclone], 1 mM glutamine [Invitrogen], 1% nonessential amino acids [Invitrogen], and penicillin/streptomycin [Invitrogen]).

hiPS cell lines iPS A1, iPS C1, iPS4, iPS A6 (Hockemeyer et al. 2009); hiPS cell lines iPS PDB^{2lox}-17, iPS PDB^{2lox}-21, iPS PDB^{2lox}-5, iPS PDB^{2lox}-22, iPS PDB^{1lox}-17puro-5, iPS PDB^{1lox}-17puro-10, iPS PDB^{1lox}-17puro-33, iPS PDB^{1lox}-21puro-20, iPS PDB^{1lox}-21puro-26, and iPS PDB^{1lox}-21puro-28 (Soldner et al. 2009); hES cell lines BG01 and BG03 (National Institutes of Health code: BG01 and BG03; BresaGen, Inc., Athens, GA); hES cell lines WIBR1, WIBR2, WIBR3, and WIBR7 (Lengner et al., 2010; Whitehead Institute Center for Human Stem Cell Research) and hES cell line H9 (NIH Code:WA09, Wisconsin Alumni Research Foundation, Madison, WI) were maintained on mitomycin C-inactivated mouse embryonic fibroblast (MEF) feeder layers in hESC medium (DMEM/F12 [Invitrogen] supplemented with 15% FBS [Hyclone], 5% KnockOut Serum Replacement [Invitrogen], 1 mM glutamine [Invitrogen], 1% nonessential amino acids [Invitrogen], 0.1 mM β-mercaptoethanol [Sigma], and 4 ng/ml FGF2 [R&D Systems]). Cultures were passaged every 5 to 7 days either manually or enzymatically with collagenase type IV (Invitrogen; 1.5 mg/ml). hiPS cell lines were passaged 15-25 times prior to ChIP-Seq and gene expression analysis.

ChIP-Seq Experiments and Analysis

Chromatin immunoprecipitation

Protocols describing chromatin immunoprecipitation (ChIP) materials and methods can be downloaded from http://web.wi.mit.edu/young/hES_PRC and have previously been described in detail (Lee et al. 2006).

Human ES, iPS or fibroblast cells were grown to a final count of ~5x10^7 cells to obtain starting material for six chromatin immunoprecipitations. Cells were chemically cross-linked by the addition of one-tenth volume of fresh 11%

formaldehyde solution for 15 minutes at room temperature. Cells were rinsed twice with 1X PBS, harvested by centrifugation, and flash frozen in liquid nitrogen. Cross-linked cells were stored at –80^{o}C prior to use.

Cells were re-suspended, lysed and sonicated to solubilize and shear cross-linked DNA. Sonication was performed using a Misonix Sonicator 3000 at a power of 27W for ten 30 second pulses with a 90 second pause between each pulse. Samples were kept on ice at all times.

The resulting whole cell extract was incubated overnight at 4 degrees C with 10µl of Dynal Protein G magnetic beads that had been pre-incubated with approximately 3 µg of the appropriate antibody. Each individual immunoprecipitation used 1/6 of the 3ml total, or ~8 x10^6 cells per IP. The immunoprecipitation was allowed to proceed overnight. Beads were washed three times (3 x 1.5ml) with RIPA buffer and one time (1x 1.5ml) with TE containing 50 mM NaCl. Bound complexes were eluted from the beads by heating at 65 degrees C with occasional vortexing and cross-linking was reversed by overnight incubation at 65 degrees C. Whole cell extract DNA (reserved from the sonication step) was also treated for cross-link reversal. Immunoprecipitated DNA and whole cell extract DNA were then purified by treatment with RNAse A, proteinase K and two phenol:chloroform:isoamyl alcohol extractions.

The ChIP antibodies used were ab8580 (Abcam) for H3K4me3 and ab6002 (Abcam) for H3K27me3.

ChIP-Seq sample preparation

All protocols for Solexa sample preparation and sequencing are provided by Illumina (http://www.illumina.com/). A brief summary of the technique, minor protocol modifications, and data analysis methods are described below.

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

Solexa sequencing

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

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

Images acquired from the Genome Analyzer were processed through the bundled image extraction pipeline (Illumina), which identified polony positions, performed base-calling and generated QC statistics.

Sequencing of the H3K27me3 ChIP from the hES BG03 cell line failed several quality control metrics and this sample not used for analysis except for gene track comparisons and the profile shown in Figure 1F.

Genomic mapping of ChIP-Seq data

ChIP-Seq reads were aligned using the software Bowtie (Langmead et al., 2009) to NCBI build 36.1 (hg18) of the human genome with default settings. Sequences uniquely mapping to the genome with zero or one mismatch were used in further analysis.

Public availability of ChIP-Seq data

Complete ChIP-Seq data are available from the Gene Expression Omnibus database (http://www.ncbi.nih.gov/geo/) under the accession number GSE22499.

ChIP-Seq density calculation and normalization of ChIP-Seq samples

The analysis methods used were derived from previously published methods

(Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007). The genome was divided into bins 100 base pairs in width, beginning at the first base of each chromosome. For identification of genomic regions with statistically significant differential ChIP-Seq occupancy 250 bp bins were used due to computer memory constraints. Each ChIP-Seq read was shifted 100 bp from its mapped genomic position and strand to the approximate middle of the sequenced DNA fragment. The ChIP-Seq density within each genomic bin was then calculated as the number of ChIP-Seq reads mapping within a 1kb window (+/- 500bp) surrounding the middle of that genomic bin.

In order to facilitate comparison of ChIP-Seq samples a quantile normalization method was used. In each ChIP-Seq sample the genomic bin with the greatest ChIP-Seq density was identified. The mean of these values was calculated and the bin with the greatest signal in each sample was assigned this mean value. This was repeated for all genomic bins from the greatest signal to the least, assigning each the average ChIP-Seq signal for all bins of that rank across all samples. H3K4me3 and H3K27me3 samples were subjected to quantile normalization as separate groups.

Identification of ChIP enriched genomic regions and genes

Genomic bins with a normalized ChIP-Seq density greater than a defined threshold were considered enriched. Adjacent enriched bins were combined into enriched regions. For H3K4me3 a threshold of 30 normalized reads per kb and for H3K27me3 a threshold of 25 normalized reads per kb was used. A summary of the H3K4me3 and H3K27me3 occupied regions is provided in Table S1.

The genomic coordinates of the full set of transcripts from the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/) from the March 2006 version of the human genome sequence (NCBI Build 36.1, hg18) was downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables) on March 1, 2009. Genes were associated with H3K4me3 and H3K27me3 occupied genomic regions if the gene transcription start site (TSS) occurred within the region or if the distance from the TSS to the boundary of the region was less than or equal to 2 kb. If multiple regions were associated with a single gene, all of these gene are reported the region with the greatest peak ChIP-Seq density used. A summary of the genes associated with H3K4me3 and H3K27me3 occupied regions is provided in Table S1.

Pairwise comparisons of H3K4me3 and H3K27me3 occupied genes and peak heights

For each RefSeq gene the peak normalized ChIP-Seq density in the region from -2 kb to +2 kb of the transcription start site was examined. A gene was considered to have different ChIP-Seq occupancy between two cell lines for H3K4me3 if the peak signal at the transcription start site (+/- 2 kb) was greater than or equal to 30 units in one cell line and less than 20 units in the other cell lines. A gene was considered to have different ChIP-Seq occupancy between two cell lines for H3K27me3 if the peak signal at the transcription start site (+/-2kb) was greater than or equal to 25 units in one cell line and less than 15 units in the other cell line. The percentage of all RefSeq genes with different ChIP-Seq occupancy between the two samples was reported in Table S2.

To compare H3K4me3 or H3K27me3 peak heights between two samples the peak ChIP-Seq density in the region from -2 kb to +2 kb of the transcription start site was examined. For each gene that was occupied by that histone mark in at least one of the two samples the coefficient of variation was recorded. The average coefficient of variation of peak heights between the two samples was reported in Table S2.

ChIP-Seq density heatmaps and composite ChIP-Seq density profiles

For ChIP-Seq density heatmaps, genes were aligned using the position and direction of their transcription start sites. Heatmaps showing the H3K4me3 and H3K27me3 ChIP-Seq density around gene start sites (-4,500 bp to +4,500 bp) within 500bp bins were generated using Java Treeview (http://jtreeview.sourceforge.net/).

For composite ChIP-Seq density profiles, genes were aligned using the position and direction of their transcription start sites. The average ChIP-Seq density around the transcription start sites of all genes in 500 bp bins was calculated for ES cells and iPS cells.

Statistical method for identifying genomic regions with differential ChIP-Seq occupancy

A test statistic, the 'differential ChIP-Seq score' was created to quantify the degree of differential ChIP-Seq density at a given position in the genome for two groups of ChIP-Seq samples. For each set of four adjacent 250 bp genomic bins the differential ChIP-Seq score was calculated as the absolute value of the mean signal of the samples in group A (A) minus the mean signal in of the samples in the group B (B) divided by an estimate of the noise of these measurements ($NOISE_{AB}$), which is described in more detail below.

$$\text{differential ChIP-Seq score} = | A – B | / NOISE_{AB}$$

The value of the $NOISE_{AB}$ was calculated using the following method. First, in each set of four adjacent 250 bp bins across the genome, the mean and standard deviation of the ChIP-Seq signal for all samples, rounded to the nearest integer, was tabulated. Second, for each mean ChIP-Seq signal, the median standard deviation (stdev) was recorded. Third, a power function, predicting the noise in these ChIP-Seq datasets, was fit to this set of mean/stdev pairs. This function was of the form;

$$\mathrm{NOISE(MEAN)} = x * \mathrm{MEAN}^{\wedge}y$$

This power function provided a good representation of the dependence between the signal intensity and noise for these ChIP-Seq experiments across the full range of signal intensities, from zero to hundreds of reads per kilobase.

The value of NOISE_{AB} was then calculated as the mean of the maximum of the value of the power function at the mean signal for the samples in group A, NOISE(A), and the actual standard deviation of these measurements A and the maximum of NOISE(B) and the actual standard deviation of the measurements in group B.

$$\mathrm{NOISE}_{AB} = \mathrm{mean}(\ \max(\ \sigma_A, \mathrm{NOISE}(A)), \max(\ \sigma_B, \mathrm{NOISE}(B)\ )$$

$\sigma_{A\ =}$ STANDARD DEVIATION OF THE SIGNAL IS SAMPLES FROM GROUP A

$\sigma_{B\ =}$ STANDARD DEVIATION OF THE SIGNAL IS SAMPLES FROM GROUP B

To assess the statistical significance of a given differential ChIP-Seq score, a permutation method was used. The distribution of differential ChIP-Seq scores under the null hypothesis was modeled by shuffling the sample to group assignments and re-calculating the test statistic. Based on the permuted sample/group assignments the differential ChIP-Seq scores were re-calculated and tabulated. All possible combinations of sample/group assignments were used to determine the null distribution except for the actual assignments and the inverse of the actual assignments.

Using this null distribution of differential ChIP-Seq scores, a false discovery rate (FDR) associated with any differential ChIP-Seq score could be calculated. This was the fraction of the genomic bins in the null distribution with that score or greater, (P_{NULL}) divided by the fraction in the actual distribution with that score or greater (P_{ACTUAL}).

$$\mathrm{FDR}(\text{ differential ChIP-Seq score }) = P_{NULL} / P_{ACTUAL}$$

A false discovery rate threshold of 5% was used for the identification of genomic regions with statistically significant differential ChIP-Seq occupancy between male and female cells and between ES and iPS cells and an FDR threshold of 1% was used for the comparison of pluripotent to fibroblast cells. Adjacent sets of genomic bins that were identified as differentially occupied were combined into regions.

Gene Expression Experiments and Analysis

Sample preparation, hybridization, staining, scanning, and image analysis

5 μg total RNA was used to prepare biotinylated cRNA according to the manufacturer's protocol (Affymetrix One Cycle cDNA Synthesis Kit). Briefly, this method involves SuperScript II-directed reverse transcription using a T7-Oligo-dT promoter primer to create first strand cDNA. RNase H-mediated second strand cDNA synthesis is followed by T7 RNA Polymerase directed *in vitro* transcription, which incorporates a biotinylated nucleotide during cRNA amplification.

Samples were prepared for hybridization using 15 μg biotinylated cRNA in a 1X hybridization cocktail with additional hybridization cocktail components provided in the GeneChip Hybridization, Wash and Stain Kit (Affymetrix). GeneChip arrays (Human U133 Plus 2.0) were hybridized in a GeneChip Hybridization Oven at 45 degrees C for 16 hours at 60 RPM. Washing was performed using a GeneChip Fluidics Station 450 according to the manufacturer's instructions, using the buffers provided in the Affymetrix GeneChip Hybridization, Wash and Stain Kit. Arrays were scanned on a GeneChip Scanner 3000 and images were extracted and analyzed using the default settings of GeneChip Operating Software v1.4.

Public availability of gene expression data

Complete gene expression data are available from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE22499.

Previously published gene expression datasets

Three previously published datasets comparing gene expression profiles of human ES, iPS and fibroblast cells, using the Affymetrix U133 Plus 2.0 (GPL570) microarray platform were obtained from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) database. For Chin et al. (Chin et al., 2009) data were obtained from the GEO accession numbers GSE9865 and GSE16654 and data from GSE16654 were subjected to an inverse (backwards) logarithmic (base 2) transformation to return them to the same linear scale as the other datasets. For Maherali et al. (Maherali et al., 2008) data were obtained from the GEO database accession GSE12390 and used with no additional processing. For Yu et al. (Yu et al., 2009) data were obtained from the GEO database accession GSE15148 and subjected to an inverse (backwards) logarithmic (base 2) transformation to return them to the same linear scale as the other datasets. Each expression dataset was normalized and analyzed for statistically significant differential expression separately using the methods described below except in Figure 4, where all datasets were normalized and clustered as one group.

Expression data normalization

The data from each gene expression sample were floored at zero and linearly scaled to a mean expression signal of 500 units. Then, all expression signal values were increased by ten units to force all signals to be greater than one unit in logarithmic space. Subsequently, within each dataset, expression signal values were quantile normalized by assigning each probeset the average signal intensity for all probesets of the same rank across all samples. Each expression dataset was normalized separately except in Figure 4, where all datasets were normalized as one group.

Expression data annotation and identification of differentially expressed transcripts

Probeset annotations were downloaded from the NetAffx database (http://www.affymetrix.com/analysis/index.affx) on October 1, 2009.

Expression datasets were analyzed for statistically significant differential expression using the online NIA Array Analysis Tool (http://lgsun.grc.nia.nih.gov/ANOVA/). Expression data was transformed into log space by the webtool upon upload. All probesets were tested for differential expression using the following settings.
Threshold z-value to remove outliers: 10000
Error Model: Bayesian
Size of sliding window for averaging error variances: 500
Proportion of highest variance values to be removed: 0
Desirable degrees of freedom for Bayesian error model: 10
Number of permutations: 0

For identification of differentially expressed transcripts between ES and iPS cells an FDR threshold of 0.05 was used. For identification of differentially expressed transcripts between pluripotent cell lines and fibroblast cells an FDR threshold of 0.01 was used. We required that differentially expressed transcripts had at least a 1.5 fold change in signal intensity.

The probesets differentially expressed between human ES and iPS cells and between pluripotent and fibroblast cells in this study, Chin et al., Maherali et al., and Yu et al. are provided in Table S5. The numbers of overlapping probesets and genes between these four datasets are also provided in Table S5.

Heatmap display and hierarchical clustering of expression data

For heat map display expression data was normalized using a formula similar to a Z-score with the following modifications. Instead of using the mean and standard deviation of all samples, the mean and standard deviation was calculated within each cell type. Then, the mean of the within cell type means, and the mean of the within cell type standard deviations was used for Z-score normalization. This served to create a balanced color range, which was not biased towards groups of samples greater numbers of expression samples.

Centroid linkage, centered correlation distance, hierarchical clustering was performed using the software Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/). Heatmaps were generated using the software Java Treeview (http://jtreeview.sourceforge.net/). Cluster branches were flipped about tree nodes for optimal display.

RT-PCR validation of microarray based expression data

For RNA analysis, hES and hiPS colonies were mechanically isolated and pooled for RNA extraction. Total RNA was isolated from ES, iPS, and fibroblast cells using RNeasy MiniKit (Qiagen). One microgram of total RNA was reverse transcribed using the Invitrogen Superscript III First Strand Synthesis System with oligo-dT primers to produce cDNA. One microliter of cDNA (1/150 of cDNA synthesis reaction) was used for each individual quantitative PCR measurement. cDNA was amplified using TaqMan Pre-developed gene expression assays (20X mixture supplied by Applied Biosystems which included pre-optimized primers and probe;Applied Biosystems). Triplicate reactions were performed in a total of 20μl using Taqman universal PCR master mix in an Applied Biosciences 7500 Real Time PCR Thermocycler. The following probes were used to detect expression in each of four ES cell lines (BG03, WIBR2, WIBR1, WIBR3), four iPS cell lines (iPS 21, iPS C1, iPS 17, iPS A6), and two donor fibroblast cell lines (Fibroblast PDB, Fibroblast GM):

Positive control: POU5F1, Hs00999632_g1
Internal standard: GAPDH, Hs02786624_g1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Chin et al.) SOX9 - Hs00165814_m1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Chin et al.) CAT - Hs00156308_m1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Chin et al.) FN1 - Hs01549980_g1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Guenther et al.) PUS7L – Hs01094423_m1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Chin et al.) BMPR2 - Hs00176148_m1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Maherali et al.) IRX3 – Hs00735523_m1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Yu et al.) GREM1 – Hs00171951_m1

Detection of abundance was determined by measuring the point during cycling when amplification could first be detected, rather than the endpoint of the 40 cycle reaction. This cycle threshold (Ct) value corresponds to the fractional cycle number where the florescent Taqman probe increases above a fixed threshold (Auto Ct) determined by the ABI Prism 7000 Sequence Detection System software. The measured Ct value was used to calculate the estimated transcripts present in the test sample using relative quantization to the average internal standard GAPDH. Average Ct was calculated for each condition, a “delta Ct” value calculated by subtracting control GAPDH Ct. Expression was calculated as relative to Fibroblast PDB line, with fibroblast PDB expression normalized to 1.

SUPPLEMENTAL REFERENCES

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

Chin, M.H., Mason, M.J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G., Aimiuwu, O., Richter, L., Zhang, J.*, et al.* (2009). Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell *5*, 111-123.

Ghosh, Z., Wilson, K.D., Wu, Y., Hu, S., Quertermous, T., Wu, J.C. (2010) Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. PLoS One. 5(2):e8975.

Hockemeyer, D., Soldner, F., Cook, E.G., Gao, Q., Mitalipova, M., Jaenisch, R. (2008). A drug-inducible system for direct reprogramming of human somatic cells to pluripotency. Cell Stem Cell *3*, 346-353.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

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

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

Lengner, C.J., Gimelbrant, A.A., Cheung, W.A., Erwin, J.a., Guenther, M.G., Alagappan, R., Xu, P., Powers, D., Barrett, B.C., Young, R.A.*, et al.* (2010). Derivation of pre-X inactivation human embryonic stem cells under physiological oxygen concentrations. Cell *141*, 872-883.

Maherali, N., Ahfeldt, T., Rigamonti, A., Utikal, J., Cowan, C., and Hochedlinger, K. (2008). A high-efficiency system for the generation and study of human induced pluripotent stem cells. Cell Stem Cell *3*, 340-345.

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

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., *et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods *4*, 651-657.

Sharov, A.A., Dudekula, D.B., Ko, M.S. (2005) A web-based tool for principal component and significance analysis of microarray data. Bioinformatics. (10):2548-9.

Soldner, F., Hockemeyer, D., Beard, C., Gao, Q., Bell, G.W., Cook, E.G., Hargus, G., Blak, A., Cooper, O., Mitalipova, M., *et al.* (2009). Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. Cell *136*, 964-977.

Yu, J., Hu, K., Smuga-Otto, K., Tian, S., Stewart, R., Slukvin, II, and Thomson, J.A. (2009). Human induced pluripotent stem cells free of vector and transgene sequences. Science *324*, 797-801.

SUPPLEMENTAL FIGURE LEGENDS

Figure S1. Expression of genes associated with genomic regions having differential H3K4me3 and H3K27me3 occupancy between ES and iPS cells.
A. Expression data for genes differentially occupied by H3K27me3 between ES and iPS cells. Genes are ordered by the magnitude of differential H3K27me3 occupancy. Samples with higher than average expression are shown in red and samples with lower than average expression are shown in green (scale in standard deviations).
B. Expression data for genes differentially occupied by H3K4me3 between ES and iPS cells. Genes are ordered by the magnitude of differential H3K4me3 occupancy. Samples with higher than average expression are shown in red and samples with lower than average expression are shown in green (scale in standard deviations). See also main text Figure 3.

Figure S2. RT-PCR based validation of genes identified as differentially expressed by microarray data.
Expression of genes (relative to expression level in fibroblast cells) in ES and iPS cells. Triplicate reactions were performed using Taqman universal expression probes. All samples were normalized to Internal standard GAPDH gene. cDNA from the cell lines 1) Fibroblast PDB, 2) Fibroblast GM, 3) iPS PDB^{2lox}-21, 4) iPS C1, 5) iPS PDB^{2lox}-17, 6) iPS A6, 7) BG03, 8) ES WIBR2, 9) ES WIBR1, 10) ES

WIBR3 was used. The Taqman probes used were POU5F1 (Positive control); SOX9 (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Chin et al.); CAT (identified as differentially expressed in iPS vs ES cells in Chin et al.); FN1 (identified as differentially expressed in iPS vs ES cells in Chin et al.); PUS7L (identified as differentially expressed in iPS vs ES cells in Guenther et al.) BMPR2 (identified as differentially expressed in iPS vs ES cells in Chin et al.); IRX3 (identified as differentially expressed in iPS vs ES cells in Maherali et al.); GREM1 (identified as differentially expressed in iPS vs ES cells in Yu et al.). The results show that POU5F1 is highly expressed in ES and iPS cells relative to fibroblasts as expected and that the PUS7L gene is differentially expressed in iPS vs ES cells, consistent with microarray based expression analysis in the same cells. Other genes do not show consistent differences between ES and iPS cells in our cell lines, which is not consistent with previously published results. See also main text Figure 3.

Figure S1. Expression of genes associated with genomic regions having differential H3K4me3 and H3K27me3 occupancy between ES and iPS cells.

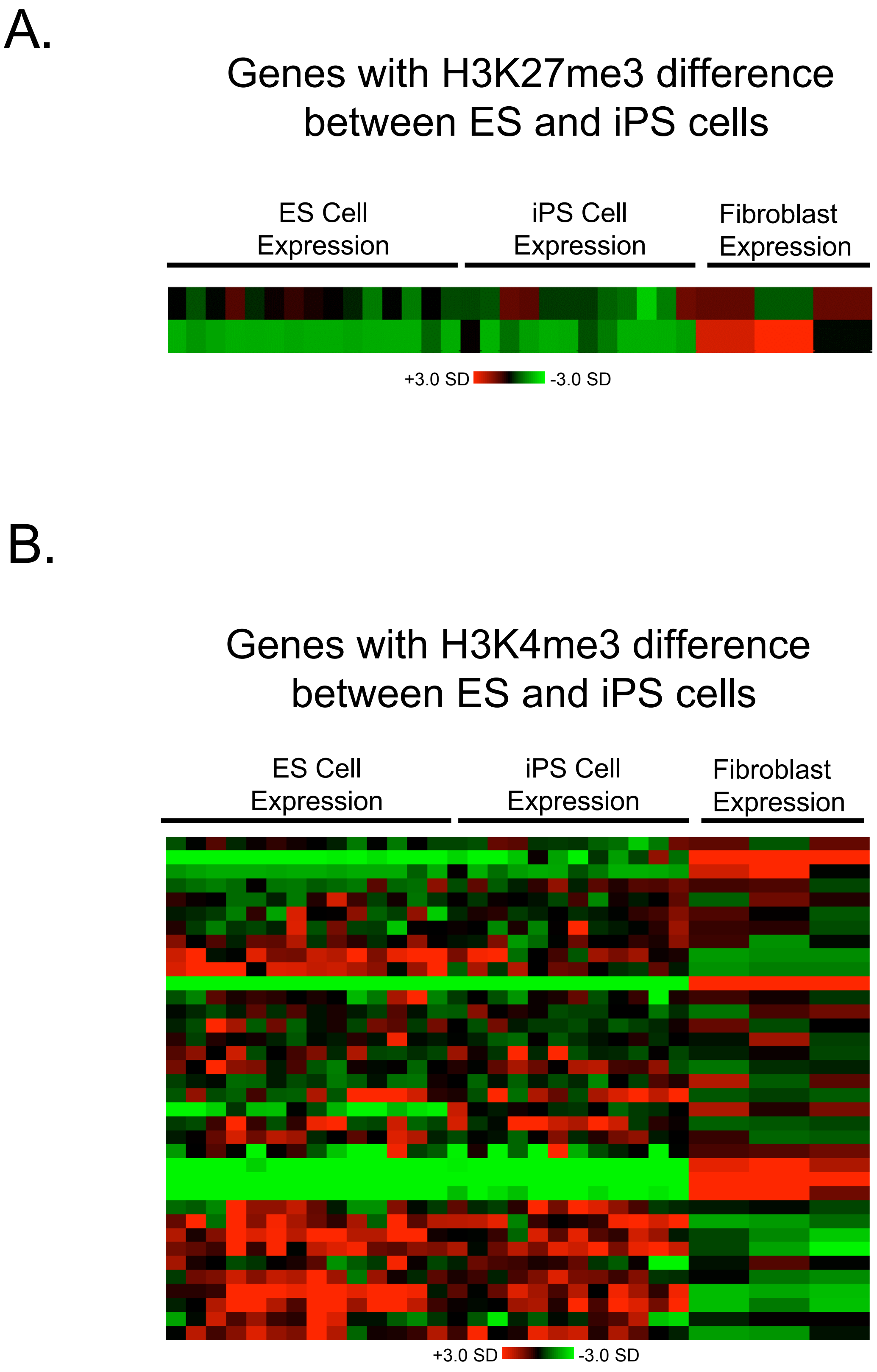
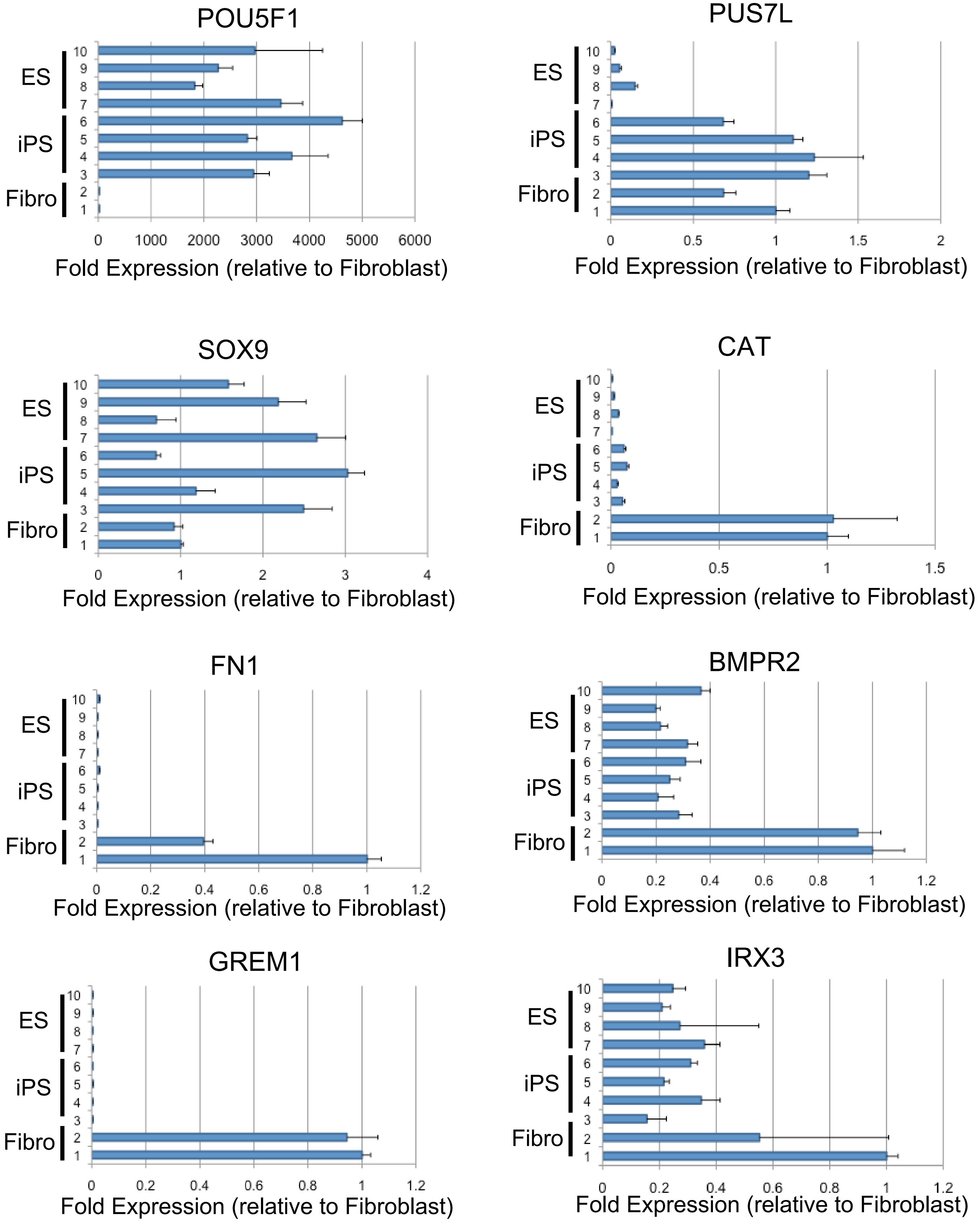


Figure S2. RT-PCR based validation of genes identified as differentially expressed by microarray data.


POU5F1
PUS7L
SOX9
CAT
FN1
BMPR2
GREM1
IRX3
ES
iPS
Fibro
Fold Expression (relative to Fibroblast)