Stem Cell Reports

Supplemental Information

Transcriptome Signature and Regulation in Human Somatic Cell Reprogramming

Yoshiaki Tanaka, Eriona Hysolli, Juan Su, Yangfei Xiang, Kun-Yong Kim, Mei Zhong, Yumei Li, Kartoosh Heydari, Ghia Euskirchen, Michael P. Snyder, Xinghua Pan, Sherman Morton Weissman, and In-Hyun Park

SUPPLEMENTAL FIGURE LEGENDS

Figure S1. Transcriptome profiling of intermediate states during hiPSC reprogramming.
(A) Schematic diagram illustrating the processing of human reprogramming intermediates.
(B) Cell counts of intermediate cell populations collected by FACS sorting. The percentage represents the number of cells expressing fibroblast or pluripotent markers calculated by Flowjo vX 0.7 software. The number represents the cell count recovered from FACS.
(C) GSEA of GO biological processes between day 0 and 3 after introduction of episomal vectors (pEP4 E02S ET2K, pEP4 E02S EN2L and pCEP4-M2L) and empty transfected or infected retroviral vector. –log10(FDR) and log10(FDR) of up- and down-regulated gene sets are shown, respectively. * FDR < 0.05.
(D) Fibroblast marker expression in each intermediate stage. Y-axis represents relative gene expression normalized to fibroblasts. Each class is composed of seven (Fib), six (I), three (II), four (III) and four populations (ESC/iPSC).
(E) Heatmap represents differentially-expressed genes ($p<0.05$ by T test and 1.5 fold change) between type III and ESC/iPSC. Relative expression values to the median expression values across eight libraries with log2 scale are represented by green (low expression) and red (high expression) colors.
(F) Overrepresentation of GO terms is shown by bar plot. Dashed line represents 0.05 FDR.
(G) GSEA of differentially-expressed genes in distinct cell populations in (Tanabe et al., 2013) was applied to transition pairs of distinct reprogramming stages. If gene sets are upregulated, -log10(FDR) is shown in red. If gene sets are downregulated, log10(FDR) is shown in blue.
(H) Principle component analysis of single-cell qPCR data.
(I) Percentage of intermediate stages in each cell population.

Figure S2. Identification of ESC-specific alternative splicing (AS) by our transcriptome dataset.
(A) Overview of pipeline to identify alternative splicing.
(B-C) Identification of known ESC-specific transcript variants, (B) *MBD2* and (C) *FOXP1* in human and mouse ESCs. These variants are specifically expressed in human type III-stage cells, iPSCs, and human and mouse ESCs.

Figure S3. Characterization of pCCNE1 isoform and ASE.
(A) Exon 9 of *CCNE1* is highly conserved among vertebrates. CCNE1 protein sequences were aligned by ClustalW2 in EMBL-EBI. Protein sequences coded by exon 9 are shown by black arrow. α-helix structure and centrosomal localization signal sequence were represented by red and blue line, respectively.
(B) Exon 9 skipping of *CCNE1* in parental human dermal fibroblast (HDF, gray), nuclear transfer stem cell (NT, purple), hESC (red), retrovirus-derived iPSCs (iPSC-R, blue) and Sendai virus-derived iPSCs (iPSC-S, green) (Ma et al., 2014).
(C) Expression of *uCCNE1* and *pCCNE1* in transgene-free iPSCs (Lister et al., 2011).
(D) Exon 9 skipping of *CCNE1* in polycistronic vector-derived iPSCs (Friedli et al., 2014).
(E) qPCR confirmation of *uCCNE1* and *pCCNE1* expression in four distinct clones derived from the lab's own retroviral pMIG-OSKM polycistronic construct (three technical replicates) (error bar, s.d.).
(F-G) qPCR of (F) *uCCNE1* and (G) *pCCNE1* in D551 fibroblasts 11 days after infection with

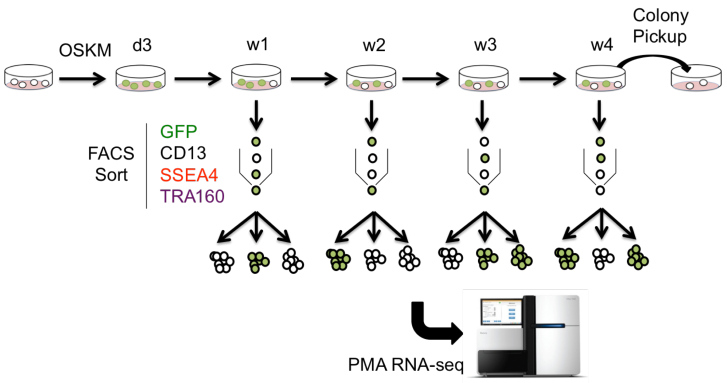
OSKM, uCCNE1, pCCNE1, or empty vector retrovirus. N.I. denotes non-infected fibroblasts. (error bar, s.d.).
(H) A model of regulation of *pCCNE1*.
(I) Validation of (Figure 3H) by double SSEA4/ TRA160 staining of reprogrammed cells. Right panel represents immunofluorescence with SSEA4 and TRA160 in hiPSC colonies generated after pCCNE1 overexpression (two biological replicates).
(J) ASE in polycistronic vector-based iPSC reprogramming (Friedli et al., 2014).

Figure S4. Effect of NOTCH signaling on iPSC reprogramming.
(A) Schematic representation of the reprogramming experiments to determine the effect of NOTCH inhibitor DAPT or NOTCH activation ligand DLL4 during different stages of reprogramming.
(B-C) The count difference of AP stained colonies in (B) DAPT- and (C) DLL4- treated cells from non-treated cells. Black, red, and blue represent treatment at whole, early, and late time points, respectively.

Figure S5. Relationship between type III/iPSC signatures and endogenous OCT4 and SOX2.
(A) GO analysis in three main principle components (PC1, 2, and 3). In each PC, top and bottom 500 genes ranked by factor loading were used for GO analysis. Dashed line represents 0.05 FDR.
(B) Venn diagram showing target genes of NANOG, PRDM14, and LIN28A.
(C) Percentage of NANOG target genes in fibroblast-type I, type II, and type III-ESC/iPSC groups (* $p < 0.05$ by hypergeometric test).
(D) Endogenous OSKM expression patterns during mouse iPSC reprogramming.
(E) Ratios of target genes by OSKM in fibroblast and type I, type II, and type III and ESC/iPSC. Gene sets are shown by pie chart.

Supplemental Figure 1

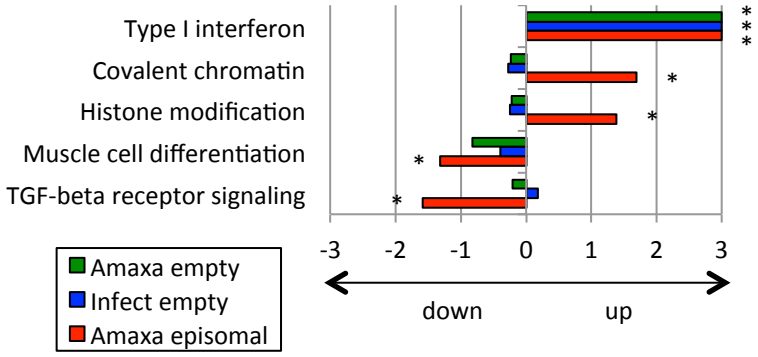
A



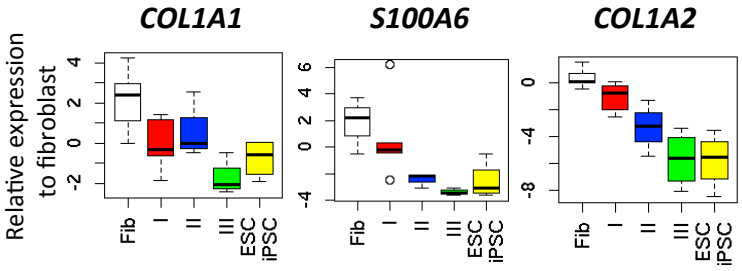
B

Date	Markers				Percentage of cell population	# of collected cells (x10⁴)
	GFP	CD13	SSEA4	TRA160		
Week 1	+	-	NA	NA	4.00%	4.1
Week 1	+	+	NA	NA	72.10%	27.6
Week 2	+	-	-	NA	36.48%	10
Week 2	-	-	+	NA	8.14%	1.4
Week 2	+	-	+	NA	1.28%	3.2
Week 2	+	+	-	NA	38.16%	20
Week 2	+	+	+	NA	4.08%	10.5
Week 3	+	NA	-	-	75.41%	10
Week 3	-	NA	+	-	1.91%	2.7
Week 3	+	NA	+	-	2.80%	4.8
Week 3	-	NA	+	+	5.11%	23.2
Week 4	+	NA	-	-	43.75%	10
Week 4	-	NA	+	-	12.61%	10
Week 4	+	NA	+	-	0.48%	1.5
Week 4	-	NA	+	+	4.30%	24

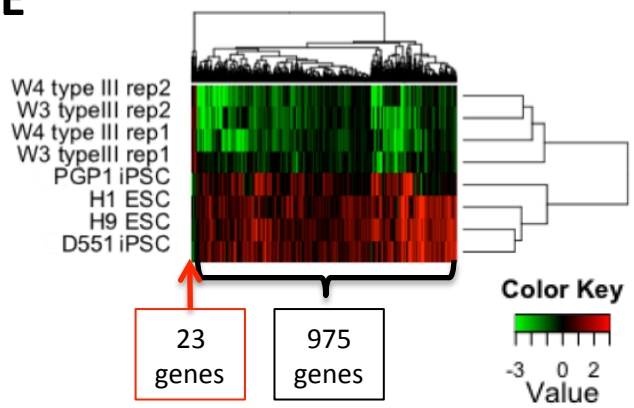
C



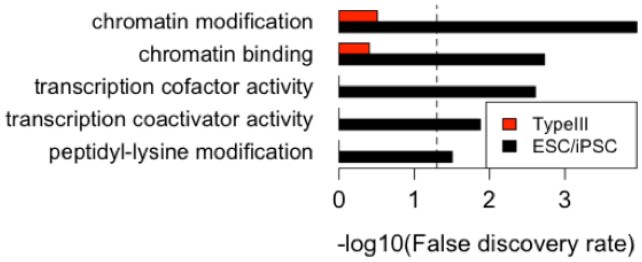
D



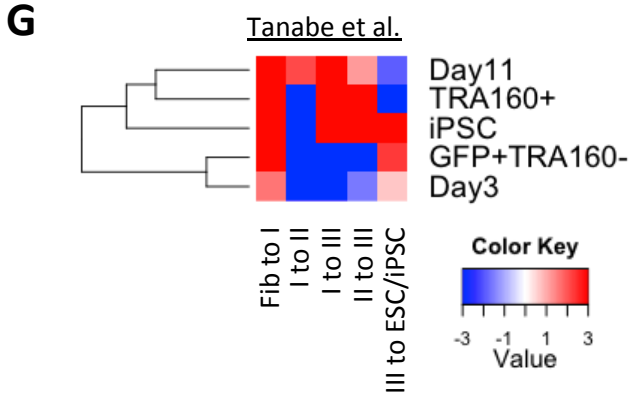
E



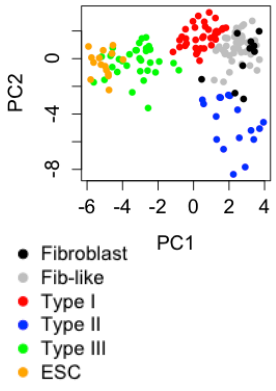
F



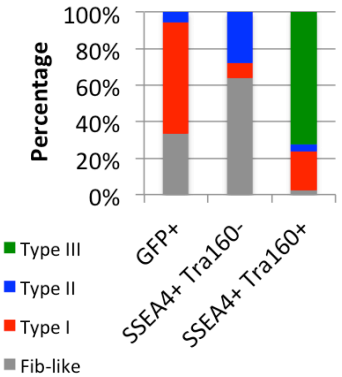
G



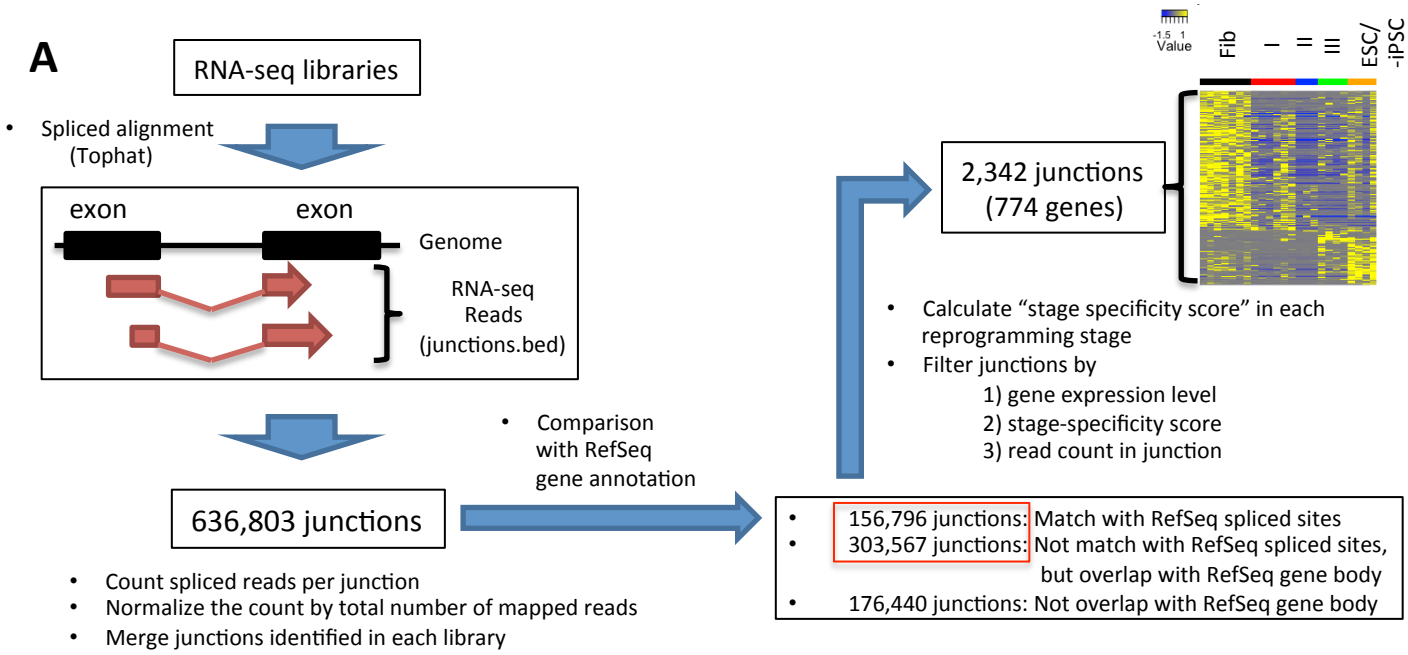
H

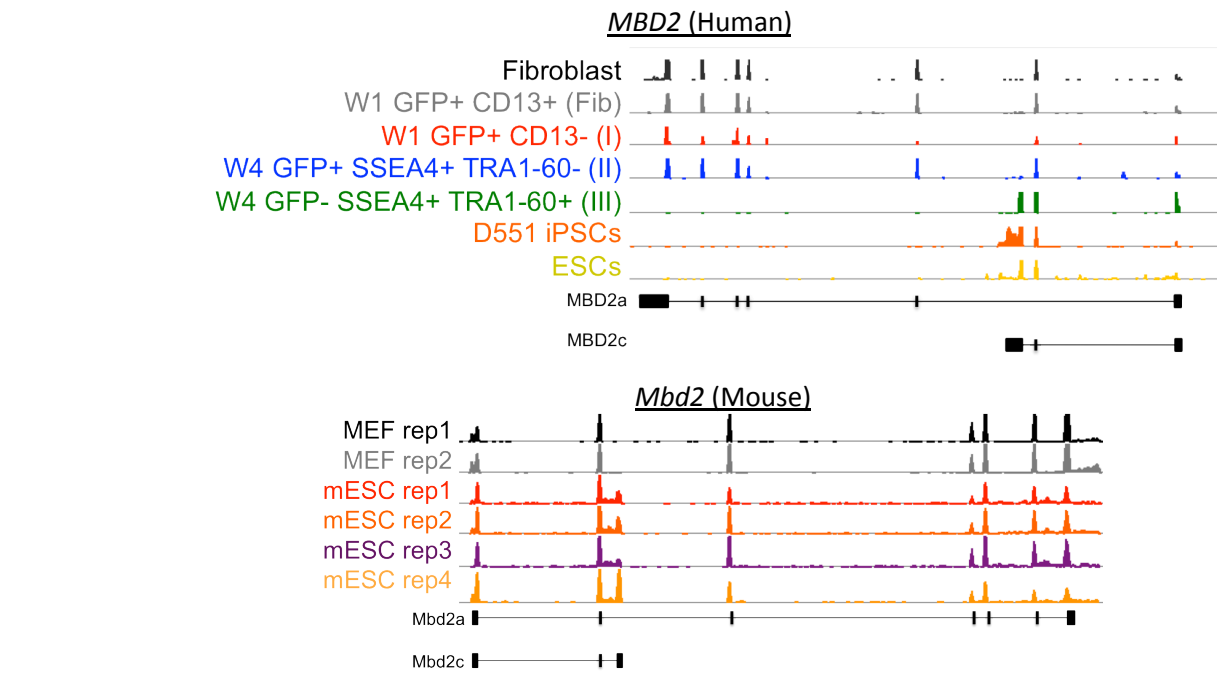


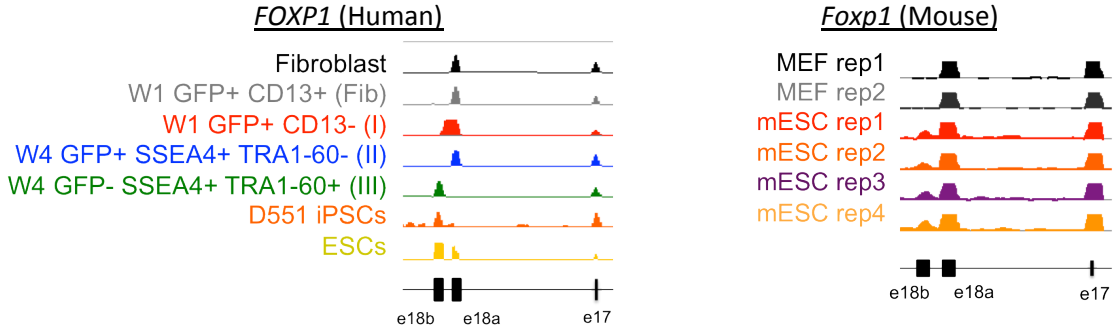
I



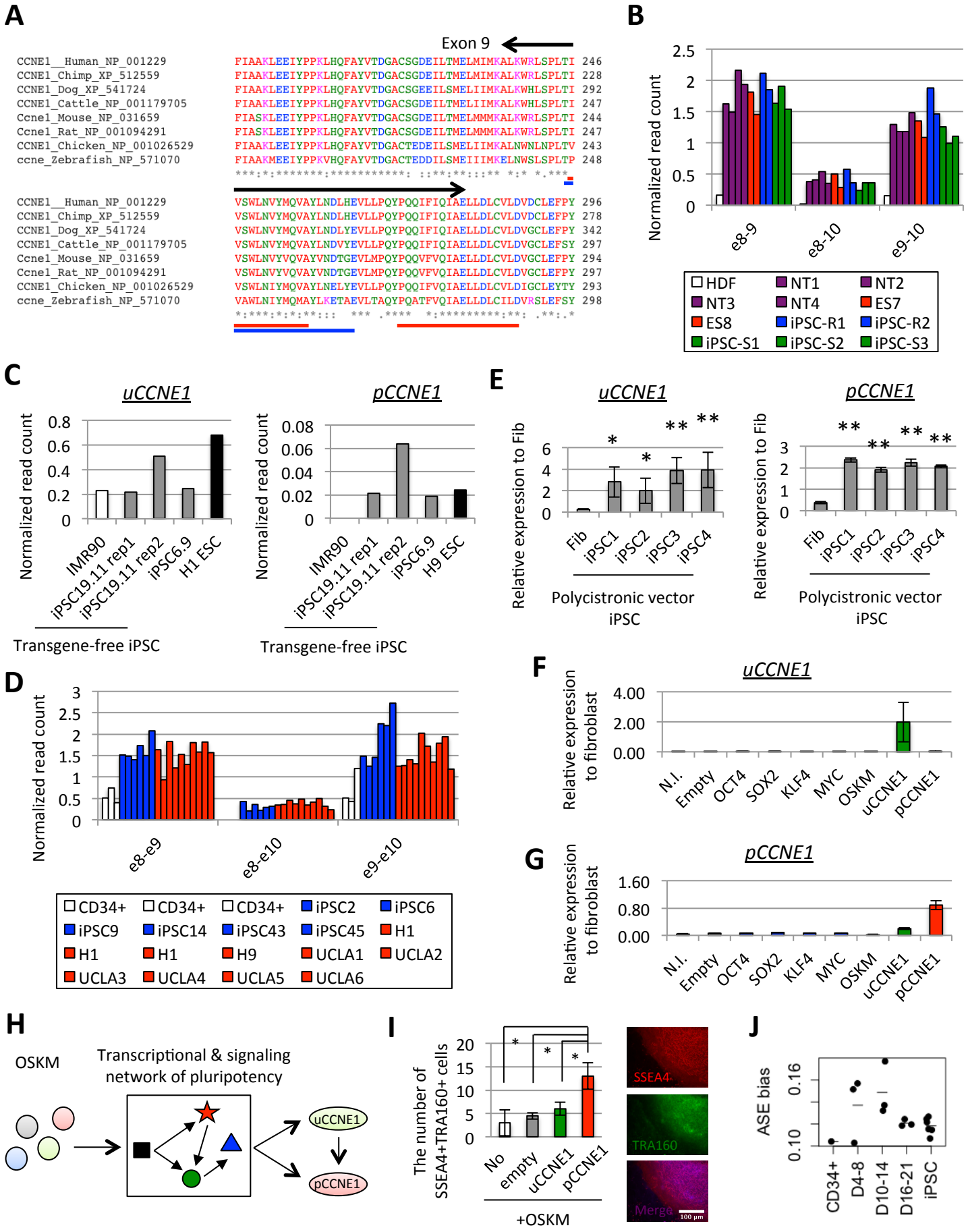
Supplemental Figure 2





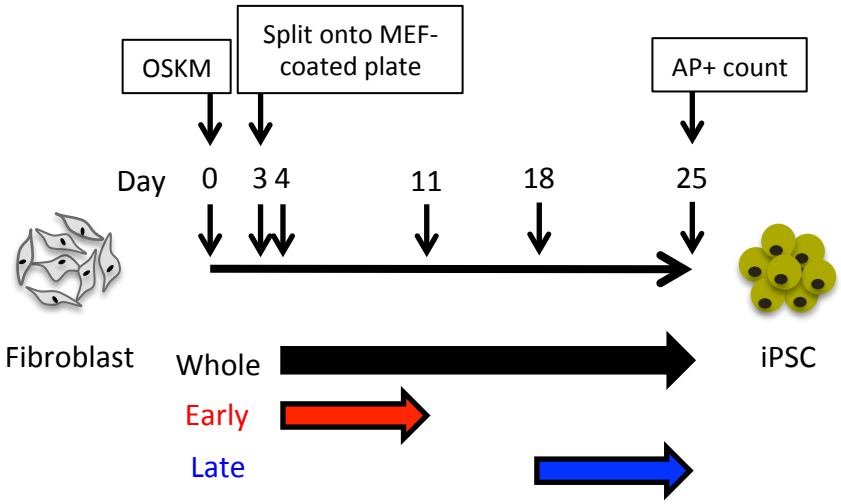


Supplemental Figure 3

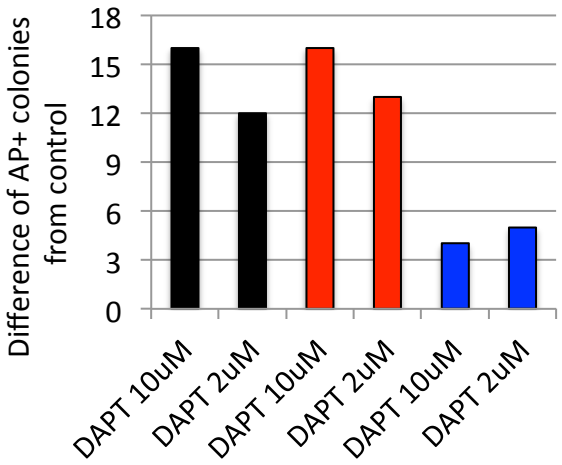


Supplemental Figure 4

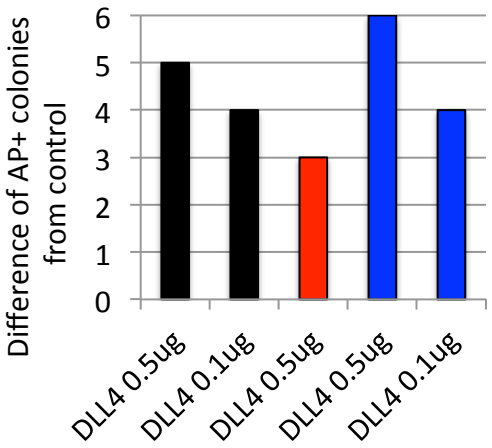
A



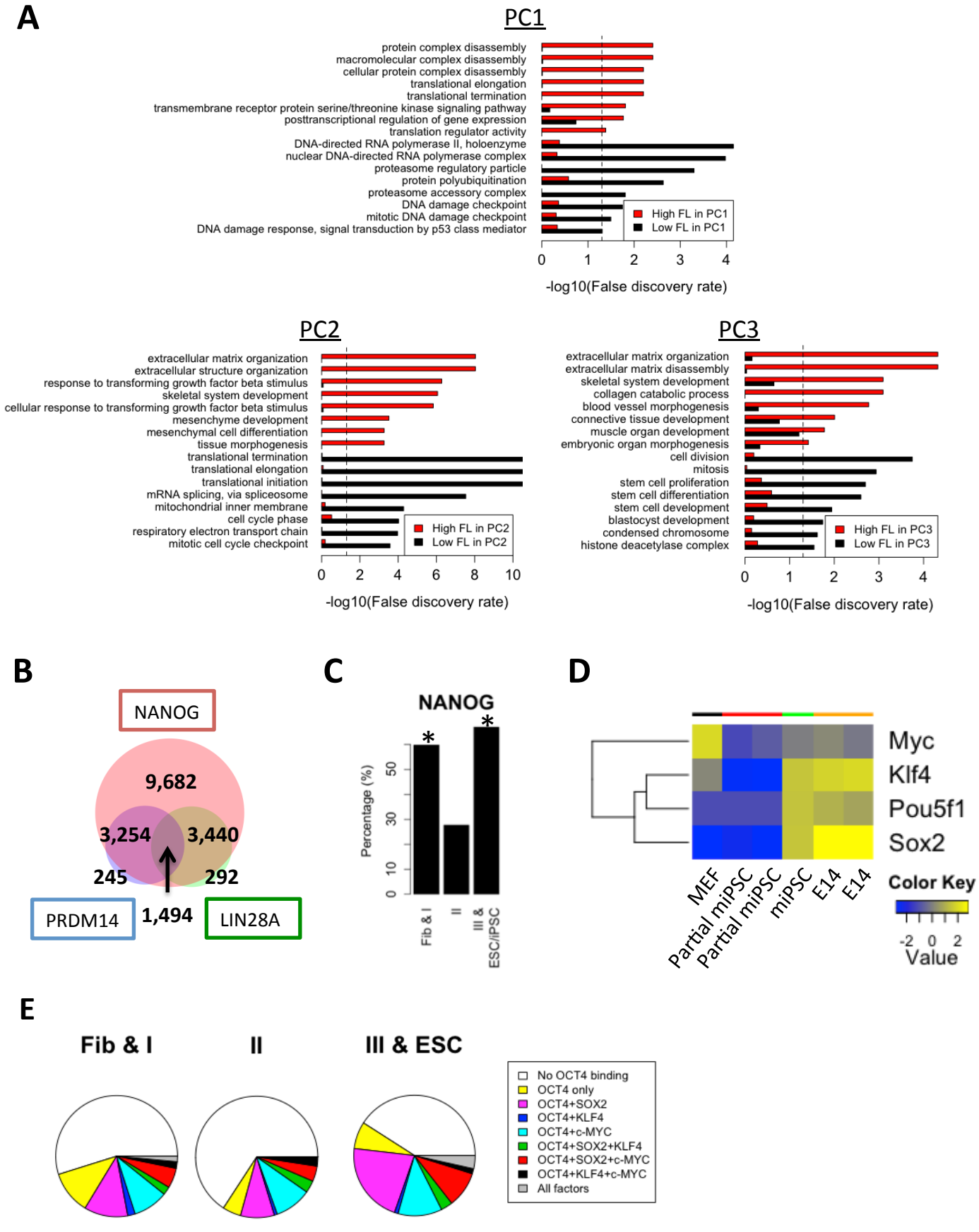
B



C



Supplemental Figure 5



SUPPLEMENTAL TABLES

Table S1. List of differentially-expressed genes between type III and ESC/iPSC stages
Table S2. Gene sets used for GSEA in (A) Tanabe et al. datasets, (B) stem cell functions, (C) signaling pathways, and (D) cancer-related genes
Table S3 List of genes in fibroblast-type I, type II, and type III-ESC/iPSC groups
Table S4. List of endogenous OSKM-specific regions
Table S5. Summary of public ChIP-seq and RNA-seq data used in this study
Table S6. List of primer sets used in this study

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Cell culture. Detroit 551 fibroblasts (ATCC CCL110) were maintained in DMEM high glucose (Gibco) supplemented with 10% FBS and penicillin/streptomycin. Human ESCs and iPSCs were cultured on irradiated murine embryonic feeder cells (Millipore), and stem cell medium composed of DMEM/F12, 20% knockout serum replacement, 2mM non-essential amino acids, 2mM L-glutamine, 4ng/ml bFGF, and 0.1mM 2-mercaptoethanol. Retrovirus production was carried out as previously described (Park et al., 2008). OCT4, SOX2, KLF4, and c-MYC, cloned into the pMIG retrovirus backbone, were transfected individually along with pCMV-Gag-Pol, pCMV-VSVG, and the transfection reagent X-tremeGENE 9 (Roche) in 293T cells. The supernatant was collected at 24, 48, and 72 hours post-transfection, and spun at 23 000 rpm for 1.5 hours. The virus pellet was dissolved in DMEM medium followed by titration in 293T cells.

iPSC reprogramming and cell sorting. The reprogramming procedure was conducted as previously described (Park et al., 2008). Detroit 551 cells were seeded at 100 000 cells/well of a 6-well plate one day prior to infection. A retrovirus cocktail containing OCT4, SOX2, KLF4, and c-MYC was added to each well at MOI 5. The next day cells were washed 3 times with 1X PBS. On day 5-post infection, the cells were trypsinized and transferred into 10-cm culture dishes containing MEFs. One day later the medium was switched to KSR-based ESC medium and subsequently changed every other day. Prior to sorting the cells were detached using accutase, washed, and incubated in 20% FBS in 1X PBS with the following antibodies, according to manufacturer's recommended dilutions: anti-human CD13 (BD cat.# 555394), anti-human/mouse SSEA4 (R&D cat.# FAB1435A), anti-human TRA160 (BD cat.# 560193). Sorting was conducted using a BD FACSAria cell sorter. Then the cells were pelleted and quickly frozen in liquid nitrogen, or sorted directly in RLT + 2-mercaptoethanol lysis buffer (Qiagen).

PMA RNA-seq library construction and Illumina sequencing. RNA was isolated from each intermediate population as well as D551 parental fibroblasts, iPSCs derived from PGP1 and D551 fibroblasts, and H1 and H9 ESCs. PMA RNA-seq library was prepared as previously described (Pan et al., 2013). Briefly, the cells were collected, washed and stored at -80 °C as pellet before processing. RNA was isolated using RNeasy Plus Micro Kit (Qiagen cat.# 74034). Then single stranded cDNA was transcribed using Superscript III in the presence of carrier RNA (Life Technologies cat.# 18080-051). Double-stranded cDNA was generated by using the above single-stranded reaction (unpurified) in the presence of E. Coli DNA Polymerase I, E.Coli DNA Ligase, and RNaseH. The reaction was purified in the presence of carrier DNA (Zymogen cat.# D4013) prior to the ligation reaction involving end-repair enzymes and T4 DNA ligase (End-It, Epicentre cat.# ER0720). Finally, the circularized double-stranded DNA product was amplified using Phi29 DNA polymerase (Epicentre cat.# RH031110), followed by gel purification. The product was then sonicated, and library preparation conducted using standard Tru-Seq Illumina kits, followed by sequencing in HiSeq 2000.

Data processing of RNA-seq. Human genomic sequence and RefSeq gene coordinate (version hg19) were downloaded from the UCSC genome browser. All RNA-seq reads were aligned to human reference genome (hg19) by Tophat (v2.0.10) using SAMtools (v0.1.18) and Bowtie (v2.1.0) with default parameters (Trapnell et al., 2009). Unmapped reads were trimmed from 3'end and the first 50bp retained to remove error-prone 3'end. These trimmed reads were

mapped to the human genome by Tophat again, and results from the first and second round mapping were merged. Then, Cufflinks (v1.2.1) was run to calculate Fragments Per Kilobase of exon model per Million mapped fragments (FPKM) by using RefSeq genes as reference annotation with "-G" option (Trapnell et al., 2010). PCA was performed to log2-transformed FPKM of 12,573 genes, which average FPKM values more than 1. Factor loading values were used to classify genes into three classes: fib/type I (PC1<0.2 and PC3>0.4), type II (PC1>0.2, PC2>-0.5 and PC3>0) and type III/ESC/iPSC-enriched genes (PC1<0.2 and (PC2+PC3)/2<-0.4). GO analysis was performed by hyperGTest function in GOstats in the Bioconductor package. Multiple-test correction was adjusted by Benjamini & Hochberg (BH) method using p.adjust function in R. The enrichment of signaling pathways and developmental genes was analyzed by GSEA (v2.0.14) software (Subramanian et al., 2005). Parameters for GSEA were set as 100 permutations of gene sets, classic enrichment statistic and signal-to-noise separation metric. 0.05 FDR was used as a cutoff for statistical significance. Gene sets used in this study were collected from public microarray data, databases and literature (Table S2).

Public microarray data analysis. Five microarray experiment data (GSE59435, GSE15603, GSE42379, GSE47489, and GSE18691) were used in this study (Chang et al., 2010; Hanna et al., 2009; Polo et al., 2012; Tanabe et al., 2013; Theunissen et al., 2014). Probe sets not overlapped with Refseq genes were removed, and those in same Refseq genes were collapsed by average. Differentially-expressed genes were identified with more than 3-fold changes and less than 0.05 FDR by T test and BH method. The datasets GSE59435 and GSE15603 were used to generate "naïve high" and "primed high" gene sets by comparison between naïve and primed ESC/iPSC in human and mouse, respectively (Table S2B). In Tanabe et al. dataset, "day3" was up-regulated genes at day 3 from fibroblasts. The datasets of "day11" was identified by comparison to the day 3 dataset, and "iPSC" was compared to day 11. All other gene sets were obtained from comparison to fibroblasts (Table S2A). Polo et al. dataset was used to compare the induction of stem cell function and signaling pathways during iPSC reprogramming in mouse (Fig. 2C and S3B). Chang et al. dataset was used to get ECC and ESC-specific genes in mouse.

Single-cell transcriptional analysis. Single cell gene expression data for fibroblasts, ESCs, and intermediate cells sorted by GFP^+, $SSEA4^+TRA1\text{-}60^-$ and $SSEA4^+TRA1\text{-}60^+$ were obtained from (Chung et al., 2014). Expression profiles were transformed to z-score in each cell, and then visualized by PCA. K-means clustering was performed to all intermediate cells with "centers=4" by kmeans function in R. Clusters, which are the nearest to fibroblasts and ESCs, were defined as fibroblast-like and type III group, respectively. A cluster between fibroblast-like and type III was categorized into type I. The farthest cluster from ESCs was classified into type II group.

Histone modification data analysis. Raw sequence data of ChIP-seq for H3K4me3, H3K27ac, H3K27me3 and H3K9me3 in fibroblast cells were downloaded from NCBI Short Read Archive (SRA) (Bernstein et al., 2010). ChIP-seq reads were mapped to hg19 genome by Bowtie2 with options "--local -D 15 -R 3 -N 1 -L 20 -i S,1,0.50 -k 1". The number of ChIP-seq reads in TSS±500bp was counted and then normalized by the total number of uniquely-mapped ChIP-seq reads to the genome. SRA IDs of ChIP-seq data used in this study were summarized in Table S5A.

Transcription factors and LIN28 target analysis. ChIP-seq raw data for initial binding of

OSKM in fibroblasts and OSKM, NANOG and PRDM14 in ESCs were obtained from NCBI SRA (Chia et al., 2010; Kunarso et al., 2010; Lister et al., 2009; Soufi et al., 2012). After mapping to hg19 genome by Bowtie2, their binding sites were identified by MACS2 peak caller with options "-g hs -q 0.05" (Feng et al., 2012). Refseq genes including transcription factor binding sites within 15k bp upstream and gene body regions were selected as targets. LIN28A binding sites in ESCs (GSM980593) were obtained from NCBI Gene Expression Omnibus (GEO) and reassigned from hg18 to hg19 using liftOver program (Wilbert et al., 2012). RefSeq genes including at least one LIN28A binding site in exons were selected as LIN28A targets. Overrepresentation of target genes in fibroblast-type I, type II and type III-ESC/iPSC gene group was evaluated by hypergeometric test with phyper function in R.

Analysis of endogenous OSKM. Endogenous OSKM gene expression was calculated using count of RNA-seq reads mapped to regions, which are not included in ectopic OSKM mRNA (Table S4). RNA-seq reads covering at least three base pairs in these regions were defined as endogenous OSKM-derived reads. The number of endogenous OSKM-derived reads was then normalized by total count of mapped reads. For endogenous OSKM analysis in mouse, we used RNA-seq data in MEFs, E14 mESCs, two partially-reprogrammed cells and a fully reprogrammed iPSC (Klattenhoff et al., 2013) (Table S5B).

Alternative splicing analysis. First, all splice junctions detected by Tophat were merged from all RNA-seq libraries (Fig. S2A). In each library, the number of spliced reads was counted at each splice junction and normalized by total number of mapped reads. In this study, splice junctions outside of RefSeq gene bodies were removed from subsequent analysis as part of novel transcripts or noises. To evaluate stage specificity of alternative splicing, we measured Jensen-Shannon (JS) divergence between the splice junction expression pattern and an extreme case of stage-specific expression, relying on (Cabili et al., 2011). Briefly, at each splice junction, the normalized read count r of library i were transformed to a density r' as:

$$r'_i = \frac{\log_2(r_i + 1)}{\sum_{j=1}^{n} \log_2(r_j + 1)}$$

$$r = (r_1, r_2, ..., r_n)$$

$$r' = (r'_1, r'_2, ..., r'_n)$$

where n is the number of RNA-seq libraries. The extreme case of stage-specific expression e was represented as:

$$e^S = (e_1^S, e_2^S, ..., e_n^S)$$

$$e_i^S = \begin{cases} \frac{1}{k} & if\ i \in S \\ 0 & if\ i \notin S \end{cases}$$

where k is the number of RNA-seq libraries belonging to stage S. Then, the JS divergence was calculated from Shannon entropy for each intermediate stage S as:

$$JS(S) = H\left(\frac{r'+s}{2}\right) - \frac{H(r')+H(s)}{2}$$

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i)$$

Finally, the stage specificity score was defined as:

$$Score(S) = 1 - \sqrt{JS(S)}$$

The stage specificity score, gene expression level, and average of the normalized read count were compared with all-pairwise comparison of five reprogramming stages. Finally, as AS candidates, we selected splice junctions, satisfying the following: 1) the difference of the stage specificity score is more than 0.35, 2) average gene expression level of both stages are more than 5 FPKM, and 3) the normalized read count of at least one compared intermediate population pair is more than 1 for one pair member and less than 0.05 for the other.

Expression of *CCNE1* splicing variants was measured by read counts mapped to exon8-exon9, exon9-exon10 (*uCCNE1*), and exon8-exon10 (*pCCNE1*) junctions. The count was normalized to the total number of mapped reads. The splicing pattern of *CCNE1* was also tested in mESC, mEpiSC, nuclear transfer human stem cell, Sendai virus-derived hiPSCs and polycistronic vector-derived hiPSCs using RNA-seq data from independent groups (Factor et al., 2014; Friedli et al., 2014; Ma et al., 2014; Yu et al., 2014). We also performed qPCR confirmation of *uCCNE1* and *pCCNE1* expression in parental D551 fibroblast and four hiPSC clones generated by in-house polycistronic pMIG vector (pMIG-4F).

We also tested CCNE1 splicing in transgene-free hiPSCs by public RNA-seq data (Lister et al., 2011). Since their read size is short (<50bp), we measured the expression level of *CCNE1* variants by a different approach. First, we built an index file from a fasta file including cDNA sequences of *uCCNE1* and *pCCNE1* by bowtie-build (v0.19.7). Then, we mapped RNA-seq read to the cDNA sequence with exact matching by bowtie ("-v 0 --sam -m 1 -a --best --strata" option). The normalized count of uniquely mapped reads to total number of reads was measured as expression level of each variant.

cDNA cloning and lentivirus construction. *CCNE1* isoforms (*pCCNE1* and *uCCNE1*) were PCR amplified from H9 ESC cDNA with primers containing restriction sites of EcoRI and XhoI (Table S6) using Quick-Load® Taq 2X Master Mix (NEB, cat.# M0271S). Each isoform was purified by 2% agarose gel and Zymoclean™ Gel DNA Recovery Kit (Zymo Research, cat.# D4002). Purified cDNA was cloned into the pMIG vector, and confirmed by Sanger sequencing. For retrovirus production, each clone was transfected along with pCMV-Gag-Pol, pCMV-VSVG into HEK293T cells with 70-80% confluence at a ratio of 2:1:1.5, together with X-tremeGENE 9. The medium was changed one day after transfection, and then collected at 48 and 72 hours post-transfection. After filtration and concentration, the retrovirus was titrated and drug selected in 293T cells prior to use.

Reprogramming with CCNE1 variant. D551 fibroblast cells were seeded at 25 000 cells/well in 12-well plate before the experiment. Fibroblasts were infected with OSKM retrovirus cocktail and either empty vector, pMIG-uCCNE1, or pMIG-pCCNE1 retrovirus with MOI 5, and cultured as described above. After four weeks, the cells were fixed and stained the using Alkaline Phosphatase (AP) Assay Kit (Sigma-Aldrich cat.# 86R-1KT). Immunostaining was also performed by adding anti-SSEA4 (BD Pharmingen, cat.# 560218) and anti-TRA160 (BD

Pharmingen, cat.# 560173) antibodies for 1 hour at 4 °C. Then, the cells were washed with PBS three times. Colonies stained with both markers were counted under a fluorescence microscope.

Allele-specific gene expression analysis. For estimation of allelic bias in the intermediate states, potential variant sites were first called from each RNA-seq mapping result using *mpileup* command in SAMtools with "-Bugf" options and *view* command in BCFtools (v0.1.17) with "-bvcg" options. Resultant variations were filtered by *varFilter* command in vcfutils.pl script with default parameters. Indel, deletion or more than two alternate non-reference alleles were removed from subsequent ASE analyses. Variant sites covered by all D551 samples were used to calculate ASE ratio as following formula:

$$ASE\,ratio = \frac{\left(Count\ of\ reads\ with\ nonreference\ allele\right)}{\left(Count\ of\ reads\ with\ reference\ allele\right) + \left(Count\ of\ reads\ with\ nonreference\ allele\right)}$$

Variant sites with more than 0.8 or less than 0.2 average ASE ratio were removed as sequence errors or mutant gene expression from a small cell population. The bias of ASE is also measured as averaged absolute value of the difference between ASE ratio and 0.5.

SNP expressions of *RPN* and *P4HB* were identified by PCR amplification of cDNA (Quick-Load® Taq 2X Master Mix). Primers were designed in exon-exon junctions to avoid contamination of genomic DNA (Table S6). Amplified cDNA was subjected to Sanger sequencing in the Keck DNA Sequencing Facility at Yale School of Medicine.

For validation of ASE in polycistronic vector-derived iPSC reprogramming, we analyzed RNA-seq data from (Friedli et al., 2014) in the same manner.

Electroporation Amaxa® Cell Line Optimization Nucleofector® Kit was used to electroporate plasmid into human D.551 cells with the nucleofector device program A-023. Either pMIG empty (5 μg), pMIG-OSKM (5 μg), or episomal plasmid DNA (11ug) was electroporated into 10^6 fibroblasts. Episomal vectors oriP/EBNA1used were from (Yu et al., 2009) as follows:
pCEP4-M2L containing MYC and LIN28 (2 μg)
pEP4EO2 SET2K containing OCT4, SOX2, and KLF4 (3 μg)
pEP4EO2 SEN2K containing OCT4, SOX2, NANOG, and KLF4 (3 μg)

SUPPLEMENTAL REFERENCES

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R.*, et al.* (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol *28*, 1045-1048.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev *25*, 1915-1927.

Chang, G., Miao, Y.L., Zhang, Y., Liu, S., Kou, Z., Ding, J., Chen, D.Y., Sun, Q.Y., and Gao, S. (2010). Linking incomplete reprogramming to the improved pluripotency of murine embryonal carcinoma cell-derived pluripotent stem cells. PLoS One *5*, e10320.

Chia, N.Y., Chan, Y.S., Feng, B., Lu, X., Orlov, Y.L., Moreau, D., Kumar, P., Yang, L., Jiang, J., Lau, M.S.*, et al.* (2010). A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature *468*, 316-320.

Chung, K.M., Kolling, F.W., Gajdosik, M.D., Burger, S., Russell, A.C., and Nelson, C.E. (2014). Single cell analysis reveals the stochastic phase of reprogramming to pluripotency is an ordered probabilistic process. PLoS One *9*, e95304.

Factor, D.C., Corradin, O., Zentner, G.E., Saiakhova, A., Song, L., Chenoweth, J.G., McKay, R.D., Crawford, G.E., Scacheri, P.C., and Tesar, P.J. (2014). Epigenomic comparison reveals activation of "seed" enhancers during transition from naive to primed pluripotency. Cell Stem Cell *14*, 854-863.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat Protoc *7*, 1728-1740.

Friedli, M., Turelli, P., Kapopoulou, A., Rauwel, B., Castro-Díaz, N., Rowe, H.M., Ecco, G., Unzu, C., Planet, E., Lombardo, A.*, et al.* (2014). Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. Genome Res *24*, 1251-1259.

Hanna, J., Markoulaki, S., Mitalipova, M., Cheng, A.W., Cassady, J.P., Staerk, J., Carey, B.W., Lengner, C.J., Foreman, R., Love, J.*, et al.* (2009). Metastable pluripotent states in NOD-mouse-derived ESCs. Cell Stem Cell *4*, 513-524.

Klattenhoff, C.A., Scheuermann, J.C., Surface, L.E., Bradley, R.K., Fields, P.A., Steinhauser, M.L., Ding, H., Butty, V.L., Torrey, L., Haas, S.*, et al.* (2013). Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. Cell *152*, 570-583.

Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet *42*, 631-634.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M.*, et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315-322.

Lister, R., Pelizzola, M., Kida, Y.S., Hawkins, R.D., Nery, J.R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S.*, et al.* (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature *471*, 68-73.

Ma, H., Morey, R., O'Neil, R.C., He, Y., Daughtry, B., Schultz, M.D., Hariharan, M., Nery, J.R., Castanon, R., Sabatini, K.*, et al.* (2014). Abnormalities in human pluripotent cells due to reprogramming mechanisms. Nature *511*, 177-183.

Pan, X., Durrett, R.E., Zhu, H., Tanaka, Y., Li, Y., Zi, X., Marjani, S.L., Euskirchen, G., Ma, C.,

Lamotte, R.H., *et al.* (2013). Two methods for full-length RNA sequencing for low quantities of cells and single cells. Proc Natl Acad Sci U S A *110*, 594-599.

Park, I.H., Lerou, P.H., Zhao, R., Huo, H., and Daley, G.Q. (2008). Generation of human-induced pluripotent stem cells. Nat Protoc *3*, 1180-1186.

Polo, J.M., Anderssen, E., Walsh, R.M., Schwarz, B.A., Nefzger, C.M., Lim, S.M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J., *et al.* (2012). A molecular roadmap of reprogramming somatic cells into iPS cells. Cell *151*, 1617-1632.

Soufi, A., Donahue, G., and Zaret, K.S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell *151*, 994-1004.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Tanabe, K., Nakamura, M., Narita, M., Takahashi, K., and Yamanaka, S. (2013). Maturation, not initiation, is the major roadblock during reprogramming toward pluripotency from human fibroblasts. Proc Natl Acad Sci U S A *110*, 12172-12179.

Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L., *et al.* (2014). Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. Cell Stem Cell.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Wilbert, M.L., Huelga, S.C., Kapeli, K., Stark, T.J., Liang, T.Y., Chen, S.X., Yan, B.Y., Nathanson, J.L., Hutt, K.R., Lovci, M.T., *et al.* (2012). LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. Mol Cell *48*, 195-206.

Yu, W., McIntosh, C., Lister, R., Zhu, I., Han, Y., Ren, J., Landsman, D., Lee, E., Briones, V., Terashima, M., *et al.* (2014). Genome-wide DNA methylation patterns in LSH mutant reveals de-repression of repeat elements and redundant epigenetic silencing pathways. Genome Res.