Stem Cell Reports, Volume *8*

Supplemental Information

PDGFRα^+ Cells in Embryonic Stem Cell Cultures Represent the In Vitro Equivalent of the Pre-implantation Primitive Endoderm Precursors

Antonio Lo Nigro, Anchel de Jaime-Soguero, Rita Khoueiry, Dong Seong Cho, Giorgia Maria Ferlazzo, Ilaria Perini, Vanesa Abon Escalona, Xabier Lopez Aranguren, Susana M. Chuva de Sousa Lopes, Kian Peng Koh, Pier Giulio Conaldi, Wei-Shou Hu, An Zwijsen, Frederic Lluis, and Catherine M. Verfaillie

SUPPLEMENTAL FIGURES

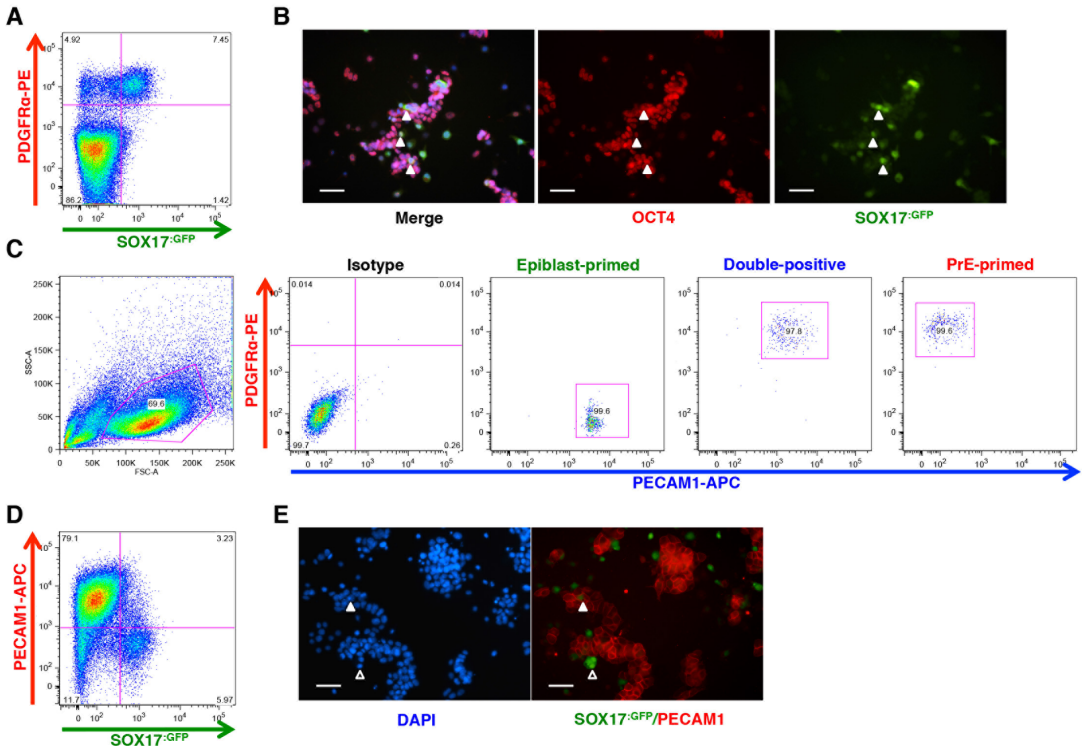

Figure S1. Co-expression of Sox17 with PDGFRα, Oct4 and PECAM1; related to Figure 1.

(A) Representative FACS analysis on *Sox17*$^{GFP/+}$ line for the expression of PDGFRα, n=3. Gating strategy was based on isotype controls.

(B) Immunostaining analysis for OCT4 on *Sox17*$^{GFP/+}$ line. Arrows indicate cells coexpressing OCT4 and SOX17. Scale bar=50μm, n=3.

(C) FSC/SSC plot, isotype controls, gating strategies and sorting purities (relative to all experiments involving sorting).

(D) Representative FACS analysis on *Sox17*$^{GFP/+}$ line for the expression of PECAM1, n=3. Gating strategy was based on isotype controls.

(E) Immunostaining analysis on *Sox17*$^{GFP/+}$ for the expression of PECAM1. Full arrow indicates cells co-expressing PECAM1 and SOX17. Empty arrow indicates cells expressing only SOX17. Scale bar=50μm, n=3.

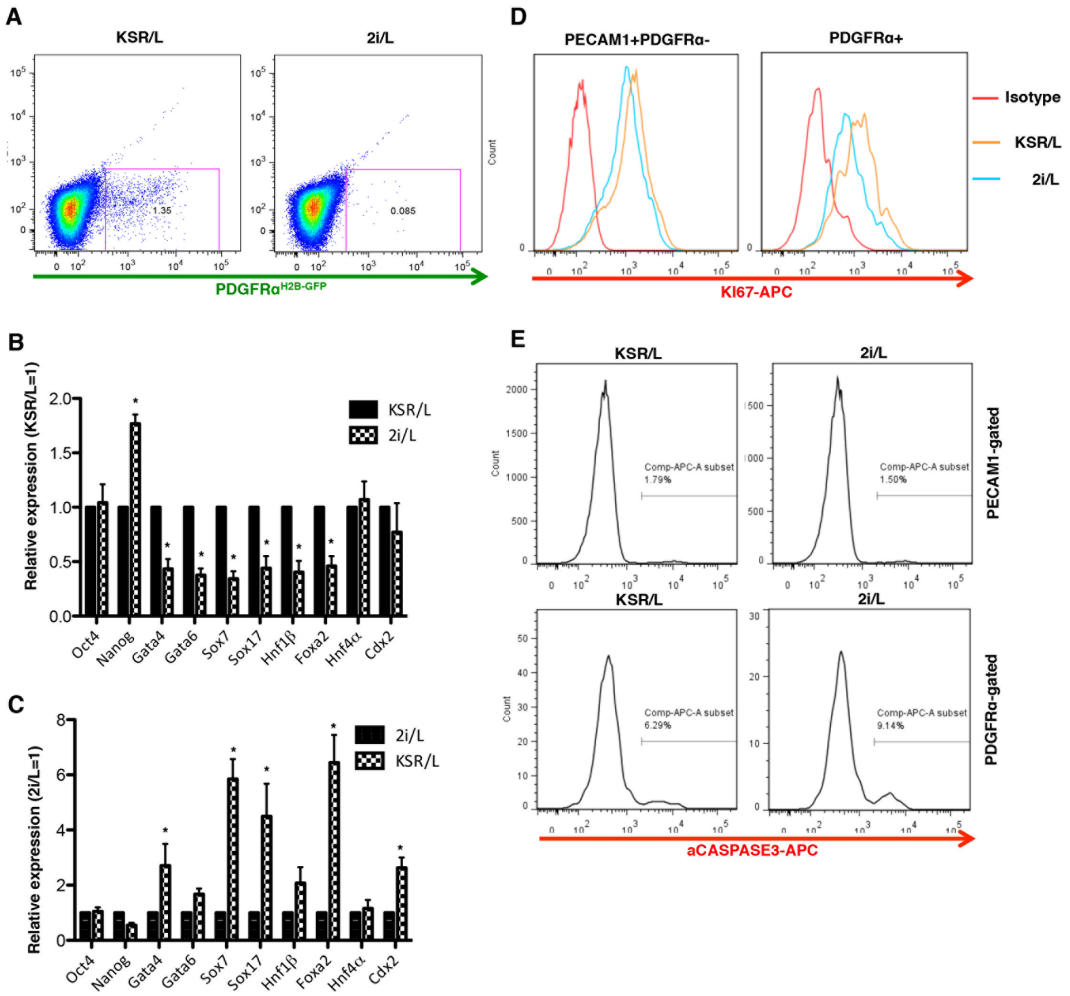

Figure S2. Effect of 2i on PDGFRα and PrE-related genes expression; related to Figure 1.

(A) Representative FACS analysis showing the percentage of *Pdgfrα*$^{H2B\text{-}GFP/+}$ cells cultured in KSR/L (left plot) or 2i/L (right plot), n=3.

(B) qRT-PCR analysis for embryonic and extraembryonic TFs on *Pdgfrα*$^{H2B\text{-}GFP/+}$ line (previously cultured in KSR/L) after 72 hours in 2i/L. Data are represented as Mean ± SEM of each transcript from three independent experiments (normalized to *β-Actin*), *p < 0.05, t test.

(C) qRT-PCR analysis for embryonic and extraembryonic TFs on *Pdgfrα*$^{H2B\text{-}GFP/+}$ line (previously cultured in 2i/L) after 72 hours in KSR/L. Data are represented as Mean ± SEM of each transcript from three independent experiments (normalized to *β-Actin*), *p < 0.05, t test.

(D) Representative FACS analysis for PDGFRα, PECAM1 and KI67 on R1 line, cultured in KSR/L or 2i/L. n=3.

(E) Representative FACS analysis for PDGFRα, PECAM1 and active Caspase3 on R1 line, cultured in KSR/L or 2i/L. n=3.

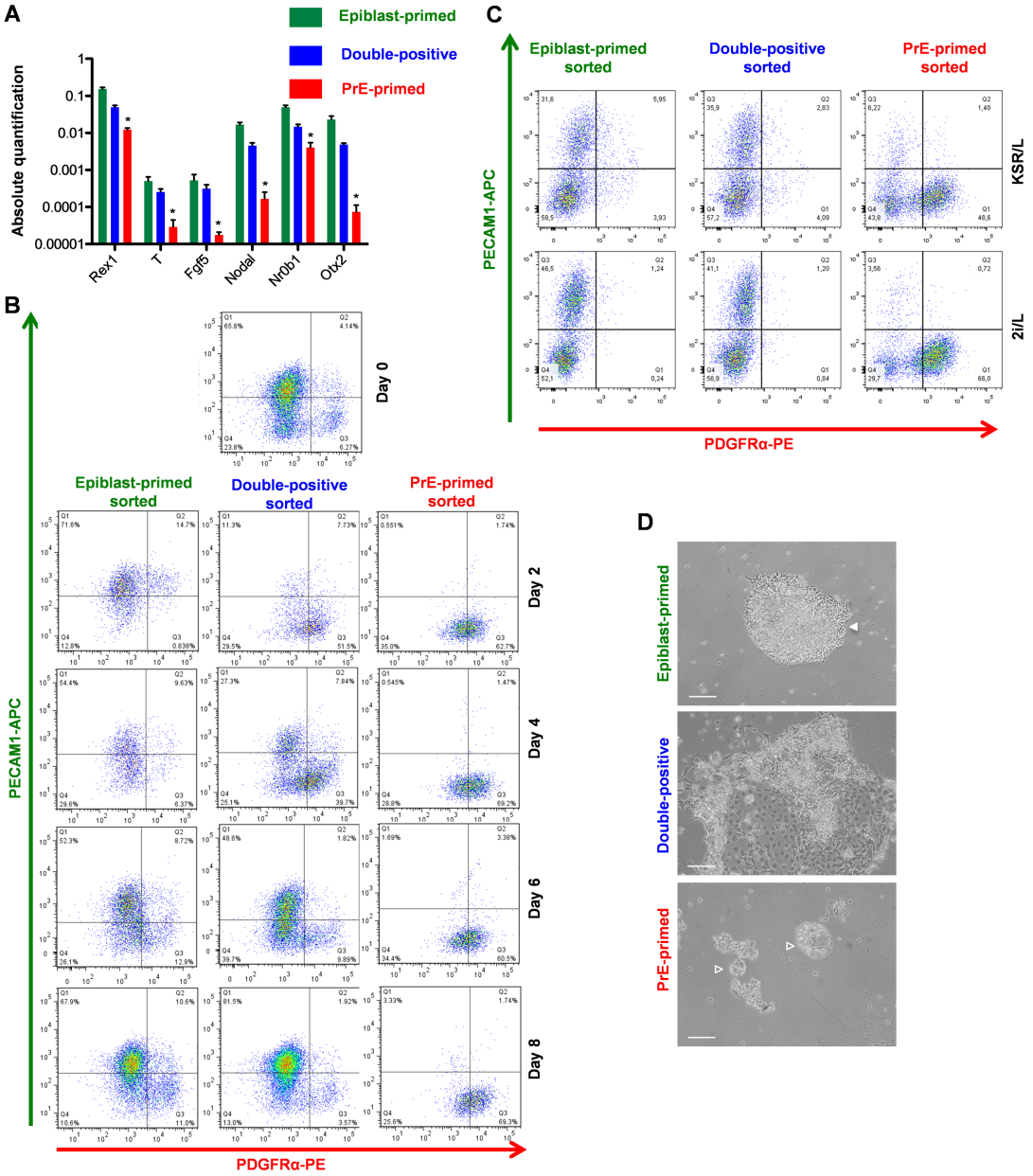

Figure S3. *In vitro* functional differences of the PrE-primed subpopulations; related to Figure 3.

(A) qRT-PCR analysis for post-implantation epiblast markers, after sorting. Data are represented as Mean ± SEM of each transcript from three independent experiments (normalized to *β-Actin*), *$p < 0.05$, t test.

(B) Representative time-course FACS analysis of the sorted subpopulations for PDGFRα and PECAM1 during re-establishment of the initial heterogeneity (E14 line). n=3. Gating strategy was based on isotype controls.

(C) Representative FACS analysis of the sorted subpopulations for PDGFRα and PECAM1 cultured in KSR/L or 2i/L for a week (E14 line). n=3. Gating strategy was based on isotype controls.

(D) Bright field pictures of sorted subpopulations cultured in neural-promoting condition. Full arrows indicate neuroectodermal precursors, while empty arrows indicate vacuolated structures. Scale bar=50μm.

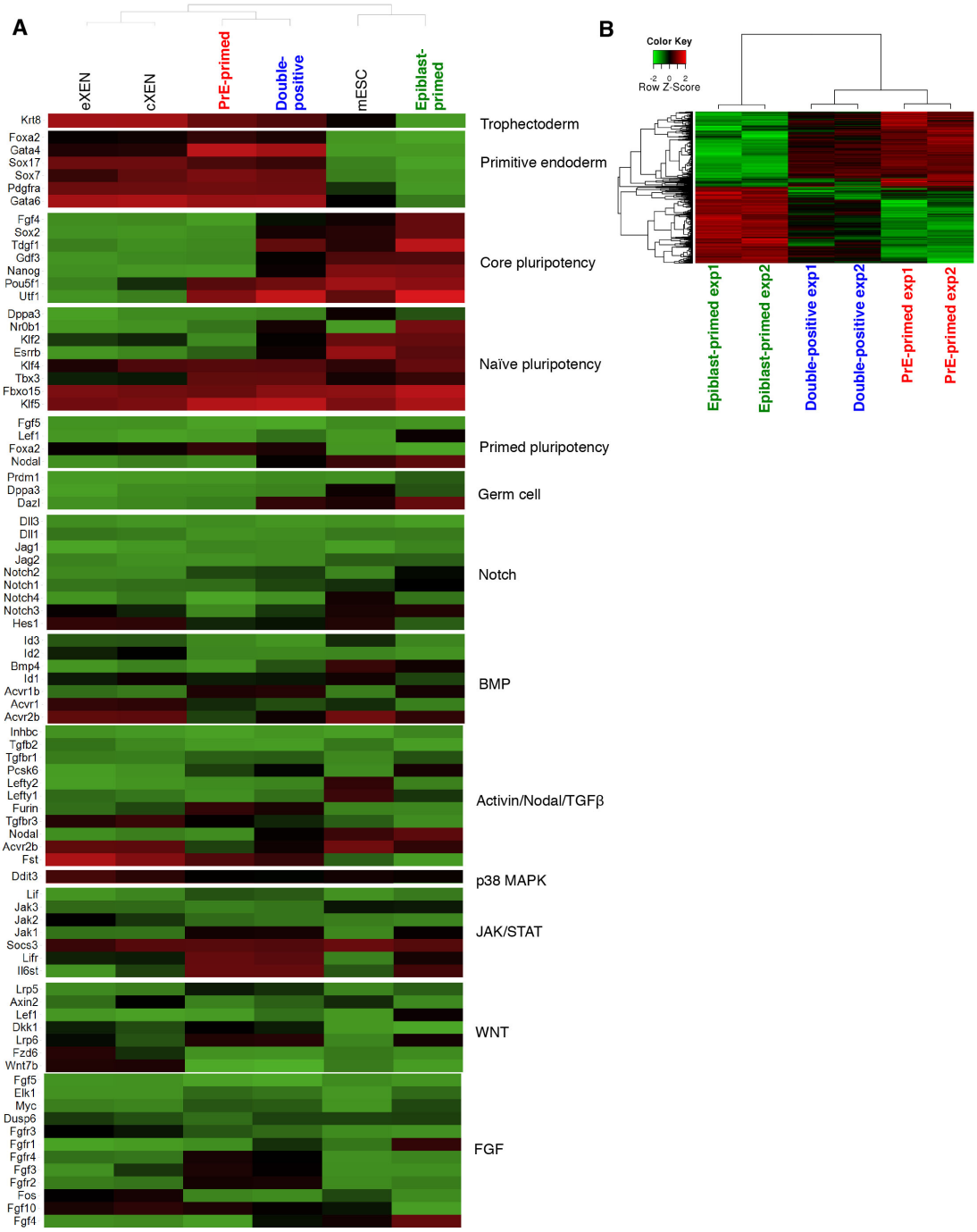

Figure S4. RNA-Seq analysis on sorted subpopulations and comparison to e/cXEN; related to Figure 6.

(A) Heat map of Naive/extraembryonic and pathway related genes on sorted subpopulations, eXEN, cXEN and unsorted ESC. Heat-map shows differentially expressed genes identified by pairwise comparison of sorted fractions. Genes are hierarchically clustered by average Euclidean distance.

(B) Heat map of sorted subpopulations, based on RNA-seq data. Heat-map shows differentially expressed genes identified by pairwise comparison of sorted fractions. Genes are hierarchically clustered by average Euclidean distance. Two biological replicates per sample are shown. Red represents upregulation while green downregulation.

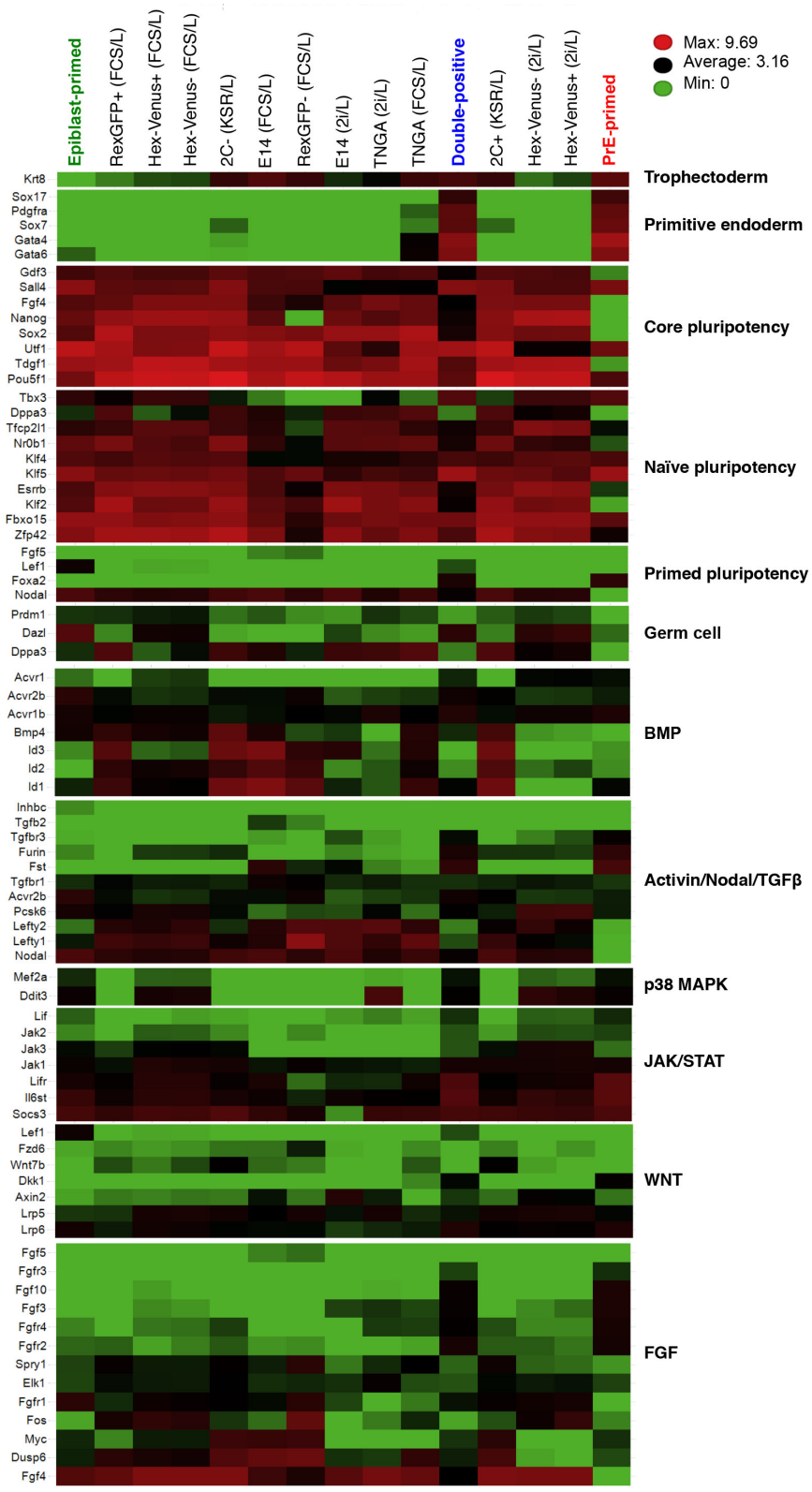

Figure S5. Heat-map comparison with other cell lines; related to Figure 6.

Heat map comparison of Naive/extraembryonic and pathway related genes on sorted subpopulations and previously published cell lines.

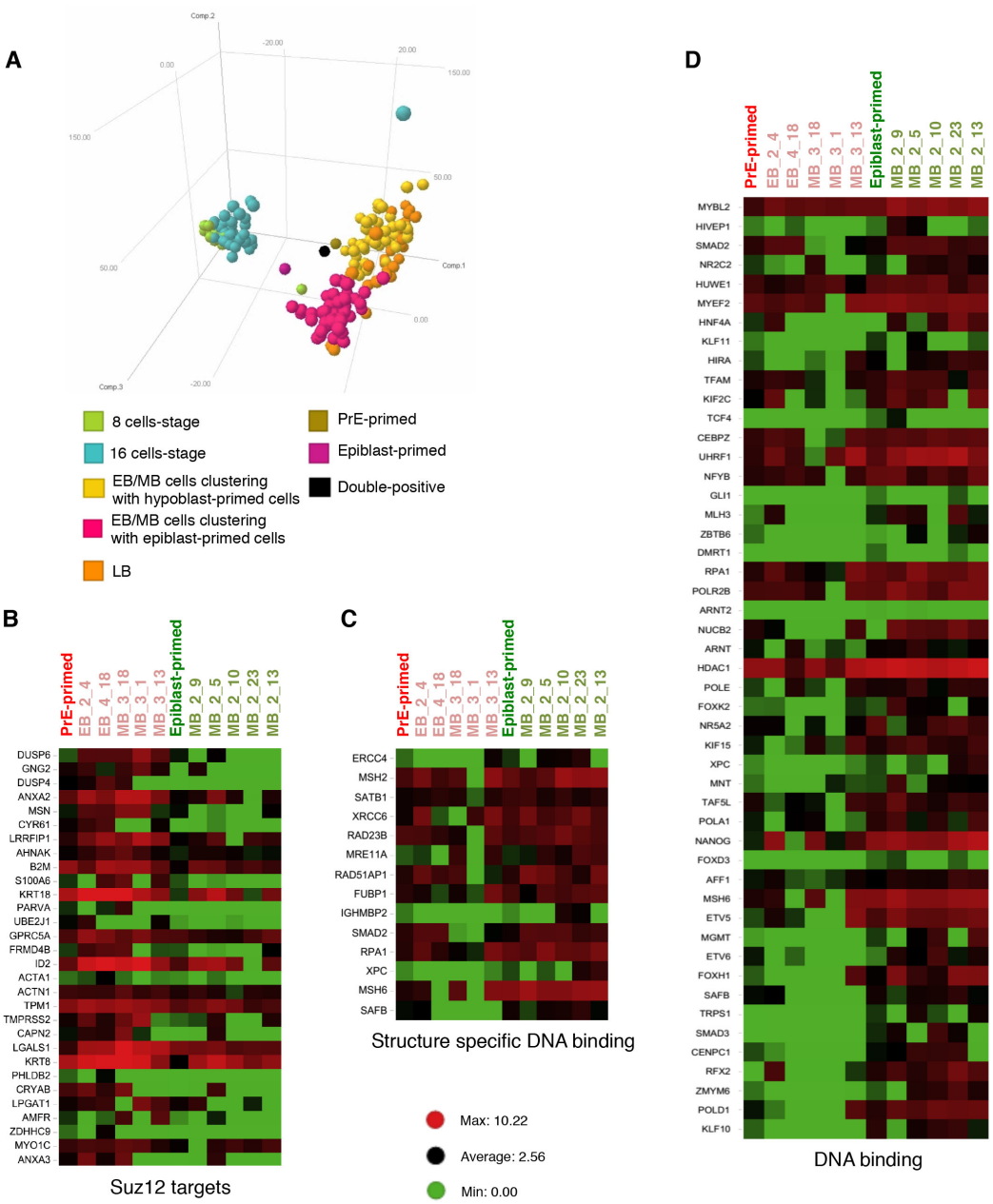

Figure S6. RNA-Seq comparison with *in vivo* cells; related to Figure 6.

(A) PCA analysis and explained variance with in vivo single cells showing EB/MB cells clustering with PrE-primed (in yellow) and epiblast-primed (in pink) subpopulations.

(B) Heat map of Suz12 targets, after GSEA of PrE- and epiblast-primed cells with their five most similar *in vivo* cells.

(C) Heat map of structure specific DNA binding-related genes, after GSEA of PrE- and epiblast-primed cells with their five most similar *in vivo* cells.

(D) Heat map of DNA binding-related genes, after GSEA of PrE- and epiblast-primed cells with their five most similar *in vivo* cells.

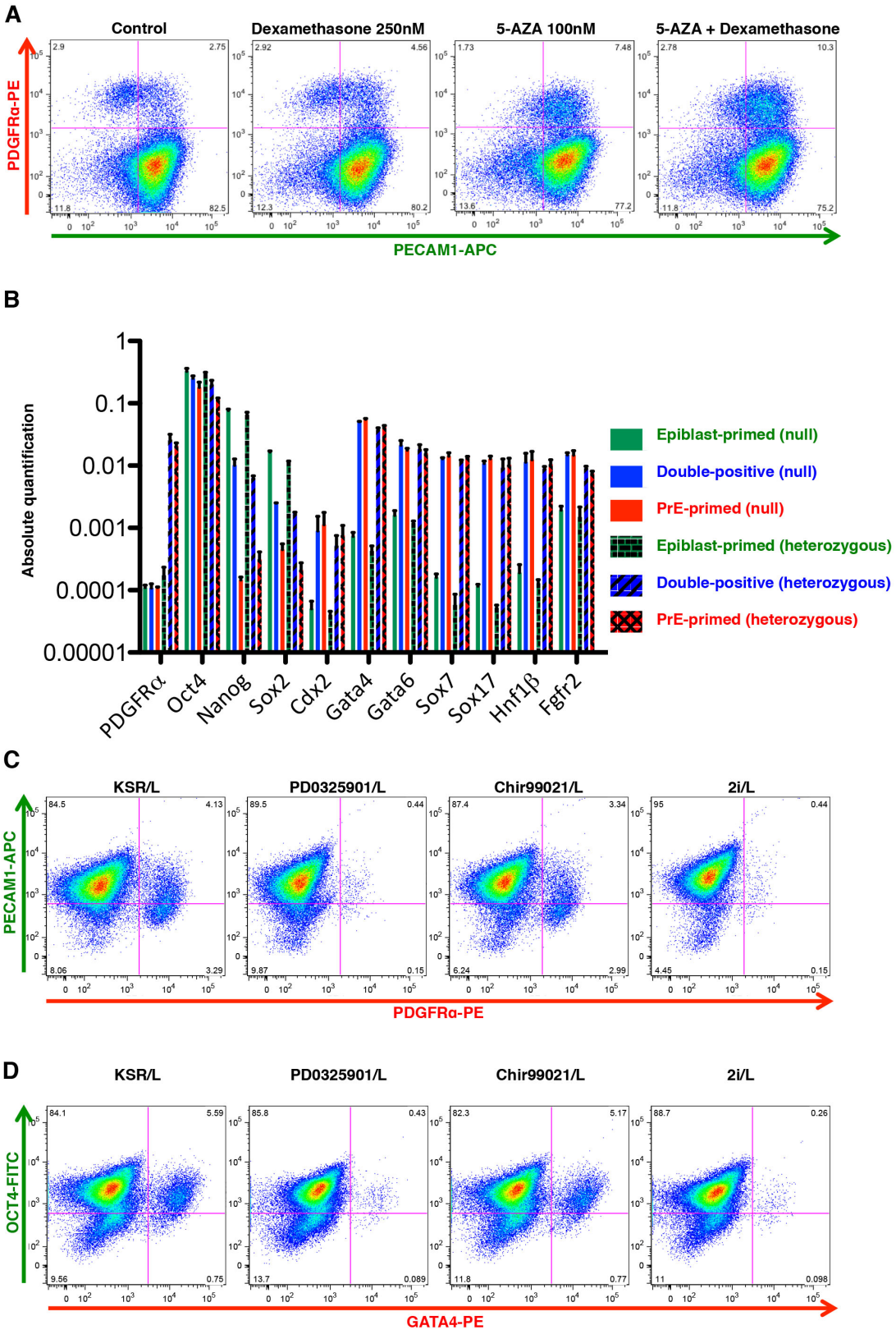

Figure S7. Epigenetic modifications and signaling involved in the regulation of the PrE-primed subpopulations; related to Figure 7.

(A) Representative FACS analysis for PDGFRα and PECAM1 on ESC cultured with Dexamethasone, 5-AZA and their combination, n=3. Gating strategy was based on isotype controls.

(B) qRT-PCR analysis for embryonic and extraembryonic markers in *Pdgfrα*$^{H2B\text{-}GFP/H2B\text{-}GFP}$ (null) and *Pdgfrα*$^{H2B\text{-}GFP/+}$ (heterozygous) ESC line. Data are represented as Mean ± SEM of each transcript from three independent experiments (normalized to *β-Actin*), *p < 0.05, t test.

(C) Representative FACS analysis for PDGFRα and PECAM1 on ESC cultured with MEK inhibitor, GSK3 inhibitor and 2i, n=3. Gating strategy was based on isotype controls.

(D) Representative intracellular FACS analysis for OCT4 and GATA4 on ESC cultured with MEK inhibitor, GSK3 inhibitor and 2i, n=3. Gating strategy was based on isotype controls.

SUPPLEMENTAL TABLE LEGENDS

Table S1: Summary of blastocyst injections.

Table S2: RNA sequencing on sorted samples. The excel file contains RNA-seq data of the tree different subpopulations

Table S3: GSEA on PrE-primed cells. The excel file contains gene set enrichment analysis for PrE-primed cells. Terms highlighted in yellow have been reported for 2i/L ESC (ground-state pluripotency).

Table S4: GSEA on double-positive cells. The excel file contains gene set enrichment analysis for double-positive cells. Terms highlighted in yellow have been reported for FBS/L ESC (primed pluripotency).

Table S5: GSEA on epiblast-primed cells. The excel file contains gene set enrichment analysis for epiblast-primed cells. Terms highlighted in yellow have been reported for FBS/L ESC (primed pluripotency).

Table S6: GSEA between *in vitro/in vivo* cells. The excel file contains gene set enrichment analysis for PrE/epiblast-primed cells and their five most similar cells from early/mid blastocyst stages.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Immunostaining cellular material

Cells were washed with PBS and fixed for 10 min with a 4% formaldehyde solution, permeabilised in 0.2% Triton-X-100 (Sigma) in PBS, incubated with a blocking solution (5% donkey serum in PBS) and stained overnight at 4°C with primary antibodies diluted in DAKO antibody diluent. Samples were incubated for 30 min at RT with the secondary antibodies in DAKO antibody diluent together with Hoechst 33258, diluted to 1:2000, for nuclear staining. In between the incubation steps, cells were washed with PBS containing 0.2% Triton-X-100. The immunostained cells were examined with a Nikon Eclipse Ti microscope using the 10x, 20x and 40x objectives, and using the Image pro plus software. Antibodies are listed in the dedicated table.

RNA-Sequencing

Library preparation. We extracted RNA from 10^5 sorted events for each subpopulation, using the Mirneasy RNA micro kit (QUIAGEN). RNA was amplified by using the Ovation RNA Amplification System V2 (NUGEN). cDNA profile was checked by running the samples on a BioAnalyzer 2100 instrument (Agilent Technologies, Santa Clara, USA) using a High Sensitivity DNA chip. cDNA quantification was done with the Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, USA) using the Qubit dsDNA HS Assay Kit. The concentrations ranged from 69.4 - 99.8 ng/µl.For all samples 1000 ng cDNA was sheared to 300 bp using the Covaris M220 Focused Ultrasonicator (SonoLab 7 Software) and screw cap micro tubes of 50 µl. Result was checked by running the sheared samples on a BioAnalyzer 2100 instrument on a High Sensitivity DNA chip. The concentration of the sheared samples was measured on the Qubit 2.0 Fluorometer. Measured concentrations were all between 17.0 and 19.8 ng/µl. Finally, library preparation was done following the manufacturer's NEBNext Ultra DNA Library preparation protocol (NEB#E7370L version 1.0, New England Biolabs) with the following minor modification: only halve of the adaptor ligated DNA fragments are subjected to a 10 cycle PCR in the final library amplification step. As input for the library preparation, 100 ng of sheared cDNA in a volume of 55.5 µl was used. Size selection was done following the recommendation of the

protocol with insert size 250 bp and total Library Size of 400 bp. The used barcoded adaptors (NEBNext Multiplex Oligos for Illumina, NEB#E7335L (Set A) + NEB#7500L (Set B)) were A001, A003, A005, A006, A012 and A019. Adaptors were checked with Illumina Experiment Manager. Final libraries were quantified using the Qubit High Sensitivity assay (Life Technologies, Carlsbad, USA). Size distribution and average length were determined by running the libraries on a BioAnalyzer 2100 using the DNA High Sensitivity chip. Because of the presence of primer dimer and over-amplification in library 2 and 4, these libraries were subjected to a Double Side Size Selection with SPRI Select Beads (Beckman Coulter). Concentration and BA profile were checked again. Molarity of each library was calculated from the concentration and the average insert length and ranged between 16.12 and 65.8 nM. Individual libraries were equimolar pooled to a 10 nM pool in a total volume of 60 µl. This pool was mixed with another 10 nM pool (60 µl) using different adaptors.

Sequencing and statistical data analysis. Sequencing was performed on 1/2 lane on an Illumina HiSeq 2000 using the 50 bp single read recipe at the Genomics Core (Leuven, Belgium).

Preprocessing. Low quality ends and adapter sequences were trimmed off from the Illumina reads with FastX 0.0.13 and Cutadapt 1.2.1 (http://hannonlab.cshl.edu/fastx_toolkit/index.html; Martin, 2011). Using FastX 0.0.13 and ShortRead 1.20.0, we filtered subsequently small reads (length < 35 bp), polyA-reads (more than 90% of the bases equal A), ambiguous reads (containing N) and low quality reads (more than 50% of the bases < Q25) (M. Morgan, 2009). With Bowtie2 v2.1.0 we identified and removed reads that mapped to the spiked-in PhiX (Langmead and Salzberg, 2012). The number of processed reads per sample then varied between 11,247,851 and 14,399,346.

Mapping. Processed reads were aligned with Tophat v2.0.8b to the reference genome of Mus musculus (GRCm38.73), as downloaded from the Genome Reference Consortium (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/; Trapnell et al., 2009). Default Tophat parameter settings were used, except for 'min-intron-length=50', 'max-intron-length=500,000', 'no-coverage-search' and 'read-realign-edit-dist=3'. Using Samtools 0.1.19, reads with a mapping quality smaller than 20 were removed from the alignments(Li et al., 2009).

Counting - Transcript coordinates were extracted from the GRC reference annotation with Gffread from the Cufflinks v2.1.1 suite and merged to gene coordinates with mergeBed from the Bedtools v2.17.0 toolkit (Quinlan and Hall, 2010; Trapnell et al., 2010). GC content and gene length were derived from the gene coordinates. The numbers of aligned reads per gene were summarized using HTSeq-count v0.5.4p3 with parameters 'm=union', 'stranded=no', 'a=0', 't=exon' and 'i=gene id' (http://www-huber.embl.de/users/anders/HTSeq). We removed 22,456 genes for which all samples had less than 1 count-per-million. As such, we continued with raw counts for 16,105 genes. Raw counts were further corrected within samples for GC-content and between samples using full quantile normalization, as implemented in the EDASeq 1.8.0 package from Bioconductor (Risso et al., 2011).

Identifying differential expression. With the EdgeR 3.4.0 package, a negative binomial generalized linear model (GLM) was fitted against the normalized counts (Robinson and Smyth, 2007). We did not use the normalized counts directly, but worked with offsets. Differential expression was tested for with a GLM likelihood ratio test, also implemented in the EdgeR package. The resulting p-values were corrected for multiple testing with Benjamini-Hochberg to control the false discovery rate. Statistical analysis was done at the Nucleomics Core (Leuven, Belgium).

RNA-seq analysis and comparison

RNA-seq datasets compared were downloaded from Gene Expression Omnibus (GEO) and Sequence Read Archive. RPKM values of single cells from mouse embryos at different preimplantation stage(Deng et al., 2014) was obtained from GSE45719. We downloaded RNA-seq data of 2C::tdTomato+, 2C::tdTomato-(Macfarlan et al., 2012), TNGA-2i, TNGA-serum, E14-serum, E14-2i, Rex1GFP-negative, and Rex1GFP-positive ES cells(Marks et al., 2012) from SRR385620, SRR385621, SRR064969, SRR064970, SRR064971, SRR064972, SRR392299, and SRR392300, respectively. Hex-Venus+ and Hex-Venus- ES cells in serum/LIF or 2i/LIF (Morgani et al., 2013) were obtained from GSE45719 and GSE45182, respectively. Illumina microarray data for mESCs, cXEN cells, and embryo-derived XEN cells (Cho et al., 2012) was obtained from GSE38477

They were aligned by Tophat2, and RPKM values are computed by Cufflink. Data across different publications was combined based on matching their ENSEMBL identifiers. Hierarchical clustering showed that data from each batch clustered independently, therefore empirical Bayes methods(Johnson et al., 2007) were used to eliminate these batch dependent effects and allow for data comparison.mPrincipal component analysis (PCA) and hierarchical clustering were performed using

OmicsOffice built in TIBCO Spotfire. Expression intensities were clustered by complete linkage method and their similarity was measured as Euclidean distance for hierarchical clustering.
Gene set enrichment analysis (GSEA) was done by GSEA software (Subramanian et al., 2005). The analysis was performed based on KEGG gene sets (http://www.broadinstitute.org/gsea/msigdb/collections.jsp#C2) downloaded from Broad Institute with 1000 permutations. GO and KEGG analysis between different sorted subpopulations was performed with DAVID (http://david.abcc.ncifcrf.gov/)

Table of primers used for this study

GENE	FORWARD	REVERSE	APPLICATION
Afp	CATGCTGCAAAGCTGACAA	CTTTGCAATGGATGCTCTCTT	qRT-PCR
B-Actin	TGTTACCAACTGGGACGACA	GGGGTGTTGAAGGTCTCAAA	qRT-PCR
Cdx2	AAGACAAATACCGGGTGGTG	CCAGCTCACTTTTCCTCCTG	qRT-PCR
Cxcr4	GCTCACCCTCATTACATACA	TAGAACTCAACAGGAGGCGG	qRT-PCR
Dnmt1	GGGTCTCGTTCAGAGCTG	GCAGGAATTCATGCAGTAAG	qRT-PCR
Dnmt3a	CCTGCAATGACCTCTCCATT	CAGGAGGCGGTAGAACTCAA	qRT-PCR
Dnmt3b	CCAAGGACACCAGGACGCGC	TCCGAGACCTGGTAGCCGGAA	qRT-PCR
Dnmt3l	CTGCTGACTGAGGATGACCA	GCTTGCTCCTGCTTCTGACT	qRT-PCR
Eomes	AGAACCGTGCCACAGACCAA	TGGTCACAGGTTGCTGGACA	qRT-PCR
Esrrb	AACCATTCAAGGCAACATCG	TTTGAGGCATTTCATGAATCGG	qRT-PCR
Fgf5	CGGACGGTGAACGACTACAC	CGTTGGAGAACCTCACTTGAC	qRT-PCR
Fgfr2	TCGATAAAGACAAACCCAAGGAG	AGATCAGACAGGTCCTTCTCTG	qRT-PCR
Foxa2	CGAGTTAAAGTATGCTGGGAG	TATGTGTTCATGCCATTCATCC	qRT-PCR
Gata3	CACAACGCAGAGCTAAGCAA	TTGTAGTTGGGGTGGTCCTG	qRT-PCR
Gata4	CCCAATCTCGATATGTTTGATGAC	ATTACATACAGGCTCACCCTC	qRT-PCR
Gata6	AGGATGTGACTTCGGCAGG	GCATCAGTGATGTCTGCAGT	qRT-PCR
Gooseco id	TGCACCTTCGGGAGGAGAAG	CCGAGGAGGATCGCTTCTGT	qRT-PCR
Hex	CGGACGGTGAACGACTACAC	CGTTGGAGAACCTCACTTGAC	qRT-PCR
Hnf1b	CCCCTCACCATCAGCCAAG	GGTTCTGAGATTGCTGGGGATT	qRT-PCR
Hnf4a	GGTCAAGCTACGAGGACAGC	ATGTACTTGGCCCACTCGAC	qRT-PCR
IAP 1st	TTGATAGTTGTGTTTTAAGTGGTAAAA	AAAACACCACAAACCAAAATCTTC	Bisulphite
IAP 2st	TTGTGTTTTAAGTGGTAAATAAATAAG	CAAAAAAAACACACAAACCAAA	Bisulphite
Krt7	ACCCTCAACAACAAATTCGCGTCC	TGCTCTTGGCTGACTTCTGTTCCT	qRT-PCR
Mixl1	ACCACCAGGCCTGACAACCT	TGGGTGCACACCATACCACA	qRT-PCR
MuErvL	GGCGCATCTGCGCACCTAAA	TAGGGTTAGACACCGGGGTT	qRT-PCR
Nanog	GAGTGTGGGTCTTCCTGGTC	GAGGCAGGTCTTCAGAGGAA	qRT-PCR
Nodal	CAGAATTGCGCCGGGATT	CCGGGATTACCAGAATTGCG	qRT-PCR
Nr0b1	TACCATCTCCTCCCTAGGG	CACCAAATCCCGGCCGGTTT	qRT-PCR
Oct4	CCAGGCAGGAGCACGAGTGG	CCACGTCGGCCTGGGTGTAC	qRT-PCR
Otx2	GGGCCACCAGGACTTACGGT	TATACCGCCGGGCCACCA	qRT-PCR
Pdgfra	GCAGCCCACACCGGATGGTA	TCCGGATCTGTGGTGCGGCA	qRT-PCR
Pecam1	CAAAGTGGAATCAAACCGTATCT	CTACAGGTGTGCCCGAG	qRT-PCR
Rex1	GCTCCTGCACACAGAAGAAA	GTCTTAGCTGCTTCCTTCTTGA	qRT-PCR
Sox17	CACAACGCAGAGCTAAGCAA	TTGTAGTTGGGGTGGTCCTG	qRT-PCR
Sox2	CTGTTTTTTCATCCCAATTGCA	CGGAGATCTGGCGGAGAATA	qRT-PCR
Sox7	CTTCAGGGGACAAGAGTTCG	GCTTGCCTTGTTTCTTCCTG	qRT-PCR
T-Bra	GTCAGACCAAGATCGCTTCT	GATCGCTTCTGTCAGACCAA	qRT-PCR
Tet1	GAAGCTGCACCCTGTGACTG	GACAGCAGCCACACTTGGTC	qRT-PCR
Tet2	AAGCTGATGGAAAATGCAAGC	GCTGAAGGTGCCTCTGGAGT	qRT-PCR
Tet3	TCACAGCCTGCATGGACTTC	ACGCAGCGATTGTCTTCCTT	qRT-PCR
Thbd	CTTCAGGGGACAAGAGTTCG	GCTTGCCTTGTTTCTTCCTG	qRT-PCR
Ttr	TTTCTTCCTGGCTTGCCTTG	AGGATGACCA CTGCTGACTG	qRT-PCR

Table of antibodies used for FACS, WB, immunofluorescence (IF) and dot blot.

ANTIBODY	CATALOGUE N°	COMPANY	DILUITION	APPLICATION
PDGFRα-APC	17-1401-81	ebioscience	1µl/10^6 cells	FACS , IF
PECAM1-FITC	11-0311-85	eBioscience	1µl/10^6 cells	FACS, IF
rIgG2a-APC	17-4321-81	ebioscience	1µl/10^6 cells	FACS
rIgG2a-FITC	11-4321-82	eBioscience	1µl/10^6 cells	FACS
Mouse anti-Oct3/4-488	560217	BD	1µl/10^6 cells	FACS
Active Casp3-647	560626	BD	1µl/10^6 cells	FACS
Ki67-647	558615	BD	5µl/10^6 cells	FACS
Gata4-PE	560328	BD	20µl/10^6 cells	FACS
Anti-Mouse Nanog -APC	50-5761-82	ebioscience	5µl/10^6 cells	FACS,WB
mIgG1-PE	550617	BD	1µl/^{106} cells	FACS
B-tubulin	05-661	Millipore	1/1,000	WB
Oct3/4 (N-19)	SC-8628	Santa Cruz B.	1/1,000	WB
Sox2	AB5603	Millipore	1/1,000	WB
Gata4 (C-20)	SC-1237	Santa Cruz B.	1/1,000	WB
Gata6	AF1700	R&D systems	1/1,000, 1/400	WB, IF
Donkey anti-Ms IgG-HRP	SC-2314	Santa Cruz B.	1/3,000	WB
Donkey anti-Rb IgG-HRP	SC-2313	Santa Cruz B.	1/3,000	WB
Donkey anti-Gt IgG-HRP	SC-2056	Santa Cruz B.	1/3,000	WB
Rb anti-5hmC	39791	Active Motif	1/5,000	Dot Blot
Ms anti-5mC	39649	Active Motif	1/500	Dot Blot
Lamininβ2	DSHB -D18	Hybridoma Bank	1/400	IF
Cdx2	Cdx2-88	Biogenex	1/200	IF
Foxa2	AB40874	Abcam	1/100	IF
Mixl1	ABS232	Millipore	1/100	IF
AlexaFluor 488 Donkey anti-Gt	A11055	Invitrogen	1/500	IF
AlexaFluor 555 Donkey anti-Gt	A21432	Invitrogen	1/500	IF
AlexaFluor 488 Donkey anti-Ms	A21202	Invitrogen	1/500	IF
AlexaFluor 555 Donkey anti-Rb	A21206	Invitrogen	1/500	IF

SUPPLEMENTAL REFERENCES

Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193-196.

http://hannonlab.cshl.edu/fastx_toolkit/index.html.

http://www-huber.embl.de/users/anders/HTSeq.

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

M. Morgan, M.L., and S. Anders. (2009). ShortRead: Base classes and methods for high-throughput short-read sequencing data.

Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57-63.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Stewart, A.F., Smith, A.*, et al.* (2012). The transcriptional and epigenomic foundations of ground state pluripotency. Cell *149*, 590-604.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1).

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. BMC bioinformatics *12*, 480.

Robinson, M.D., and Smyth, G.K. (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinformatics *23*, 2881-2887.

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. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545-15550.

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. Nature biotechnology *28*, 511-515.