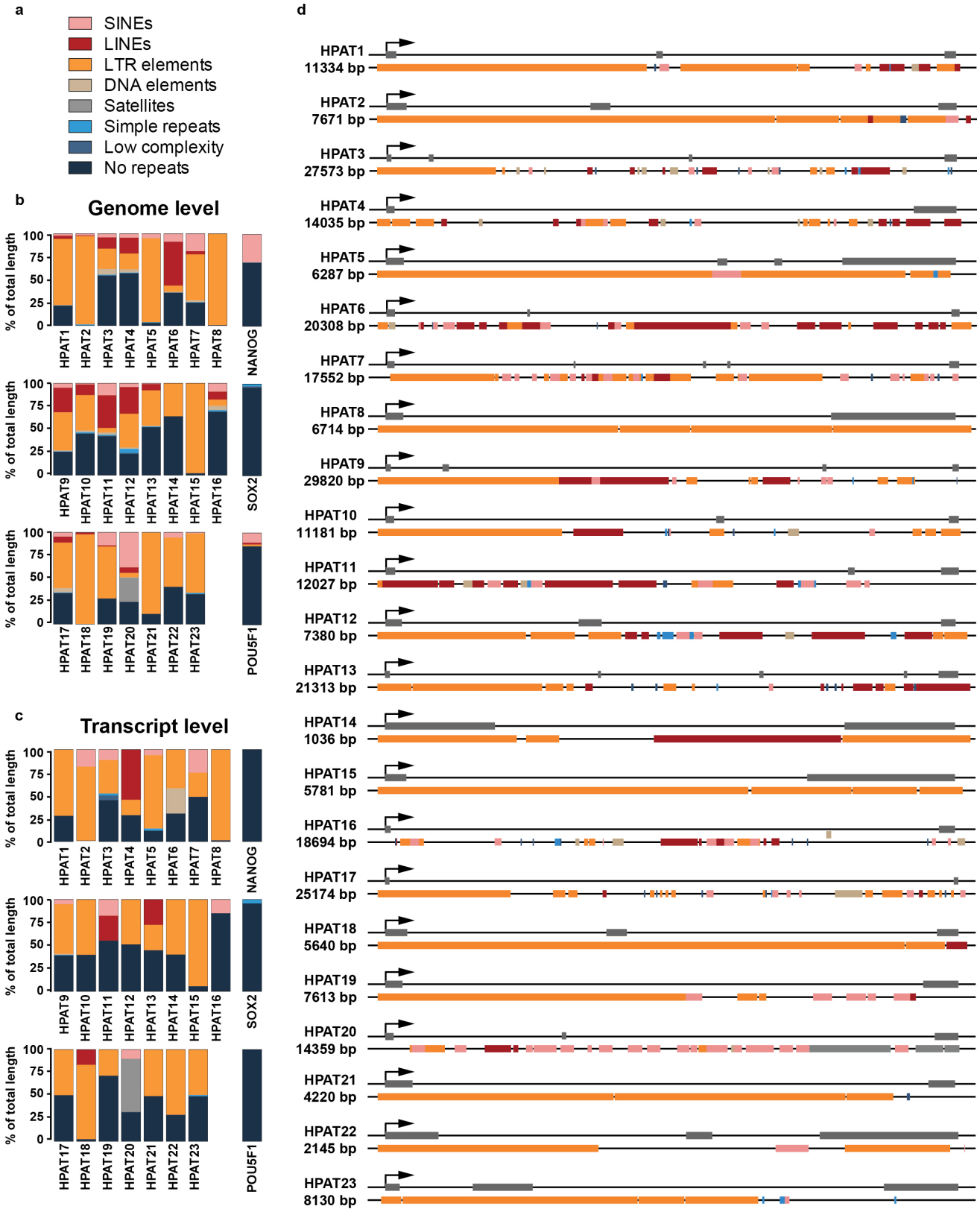
Supplementary Fig. 1



Nature Genetics: doi:10.1038/ng.3449

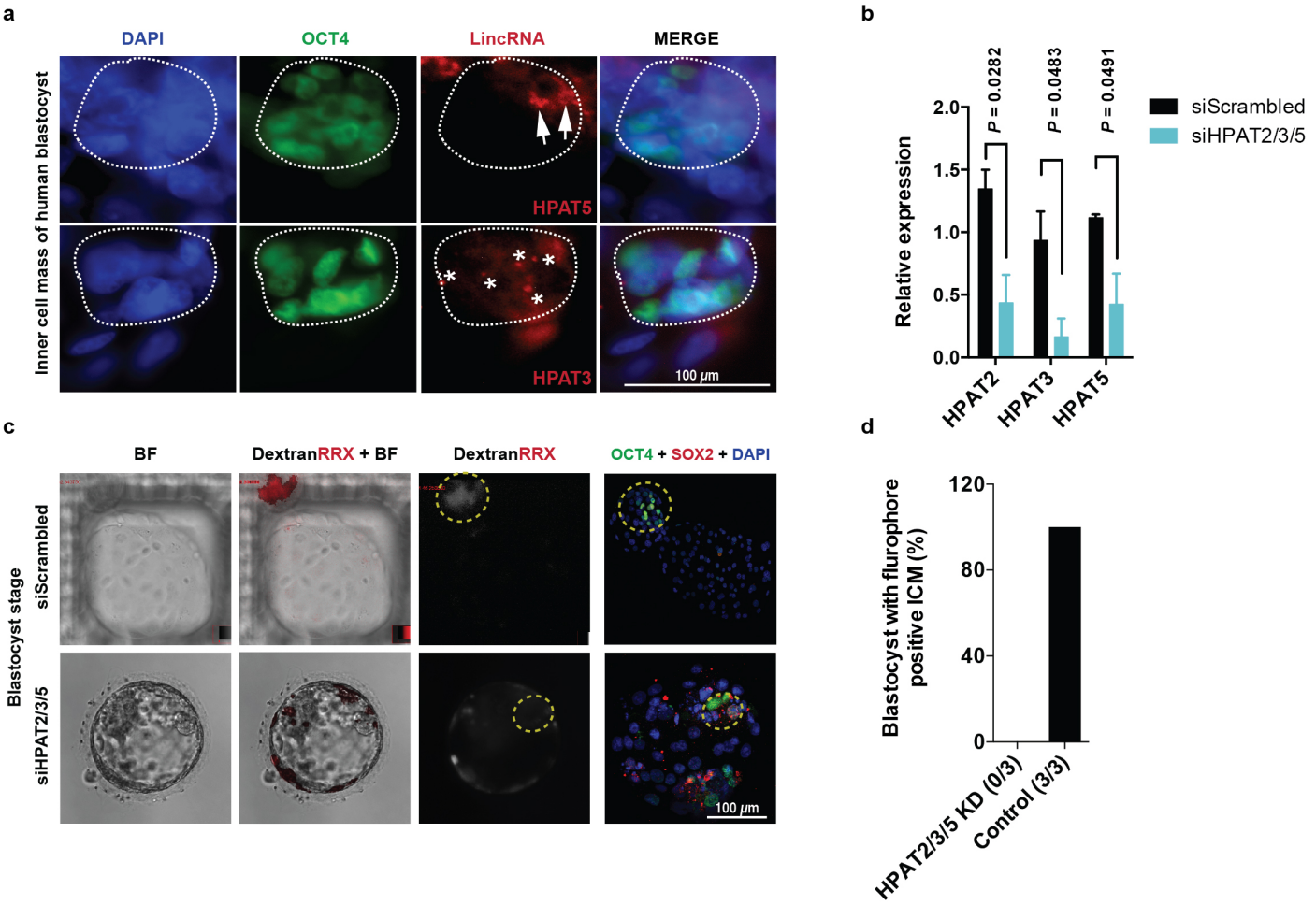
Supplementary Figure 1

All 23 HPATs are significantly enriched for TEs.

(**a**) The different classes of TEs are color-coded; corresponding colors are used in **b**–**e**. (**b**,**c**) Coverage of different TE classes on the genome (exons + introns) and transcript (only exons) levels. The percentage of total length for each TE is represented for all 23 HPATs. Three control genes are included. (**d**) The 23 HPATs with embedded TEs and genomic length. Displayed are the most highly expressed isoforms for each HPAT gene. Genomic DNA is represented as a black line, and exons are represented by gray boxes. TEs are represented by the colored boxes underneath. Each exon is exonized with TEs (exons overlap TEs). The length of each genomic locus is not to scale.

Nature Genetics: doi:10.1038/ng.3449

Supplementary Fig. 2



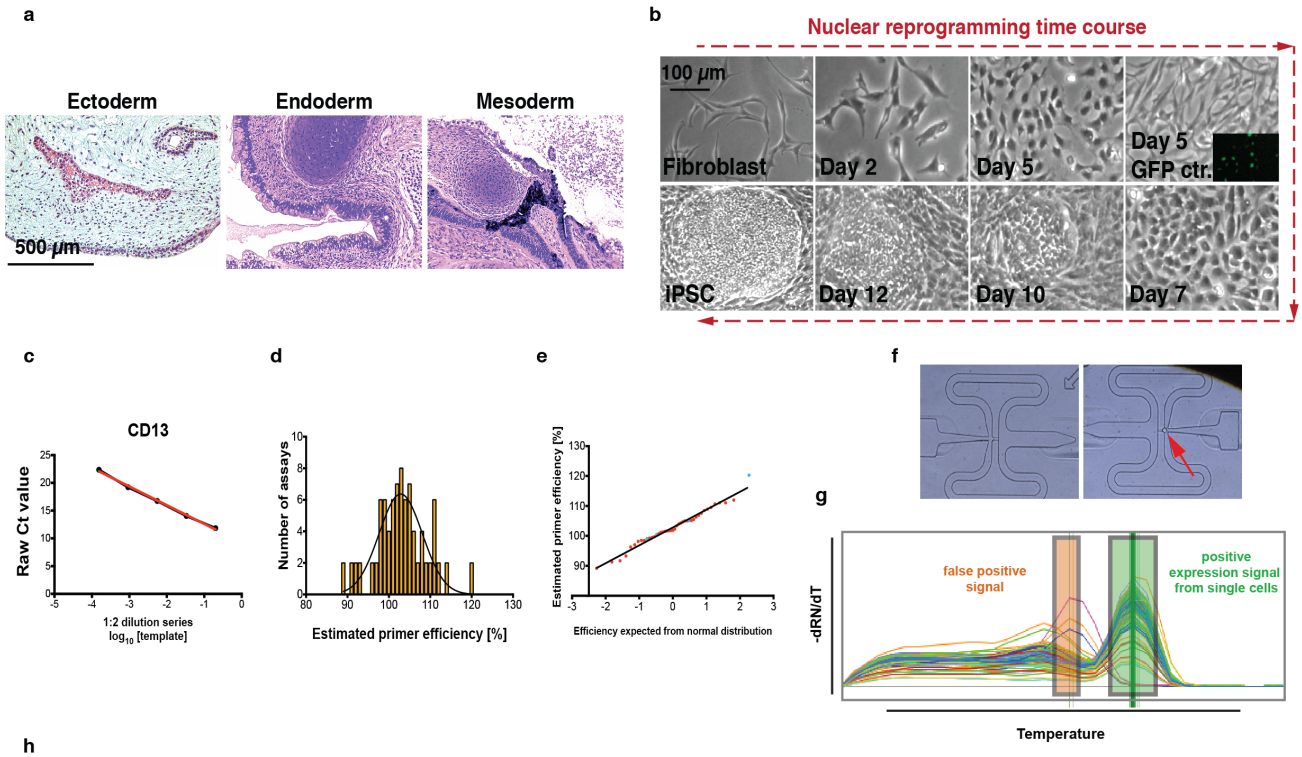
Supplementary Figure 2

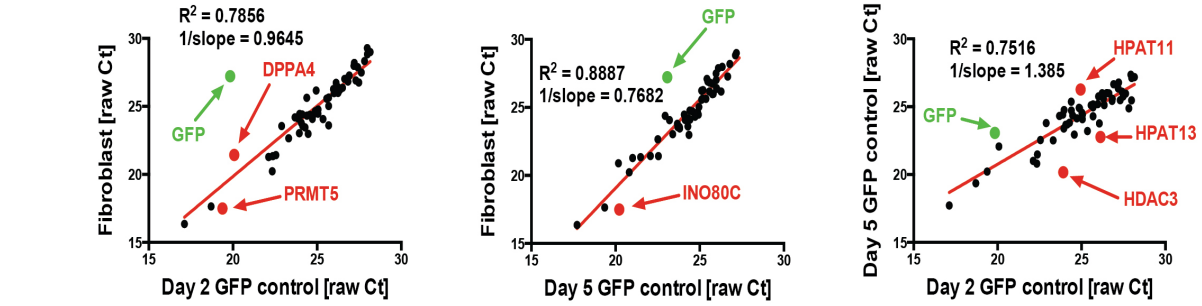
Molecular and functional analysis of *HPAT2*, *HPAT3* and *HPAT5* during preimplantation development.

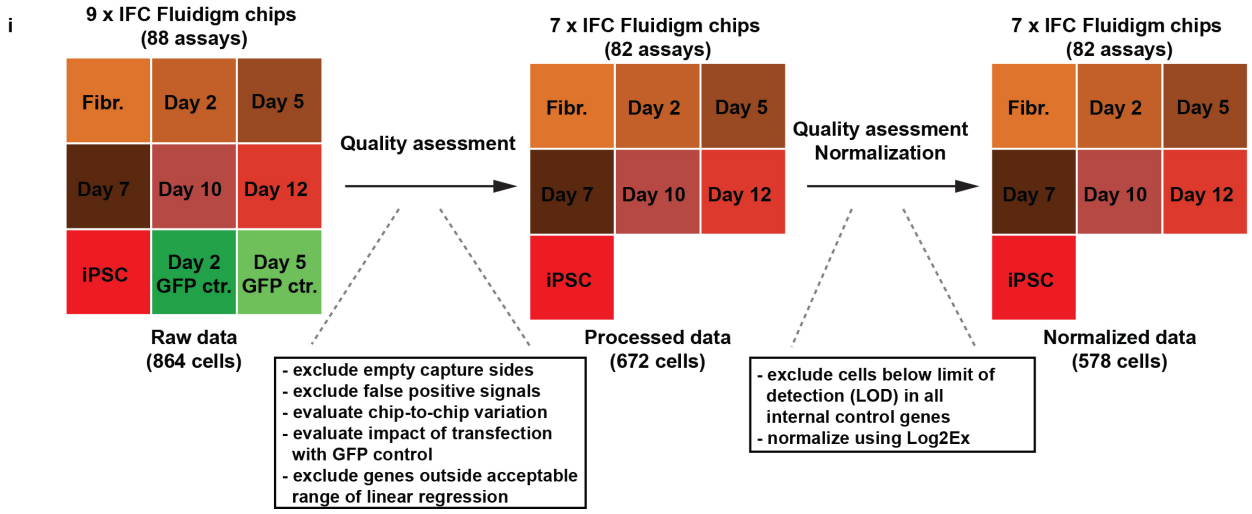
(**a**) Magnified view of the ICM in human blastocyst demonstrates a specific staining pattern in the ICM of human blastocysts. Stars depict *HPAT3* signal. Arrows depict *HPAT5* signal. (**b**) *HPAT2*, *HPAT3* and *HPAT5* are significantly downregulated in human blastocysts injected with siRNAs compared to siScrambled controls (n = 3 blastocysts; data are shown with s.e.m.). (**c**) Blastomeres with knockdown of *HPAT2*, *HPAT3* and *HPAT5* during human embryo development did not contribute to ICM. The presence of ICM was validated with OCT4 and SOX2 staining. The ICM is highlighted by a yellow dashed circle. (**d**) Fluorescence-positive ICM in blastocysts *with* HPAT knockdown and control blastocysts.

Nature Genetics: doi:10.1038/ng.3449

Supplementary Fig. 3







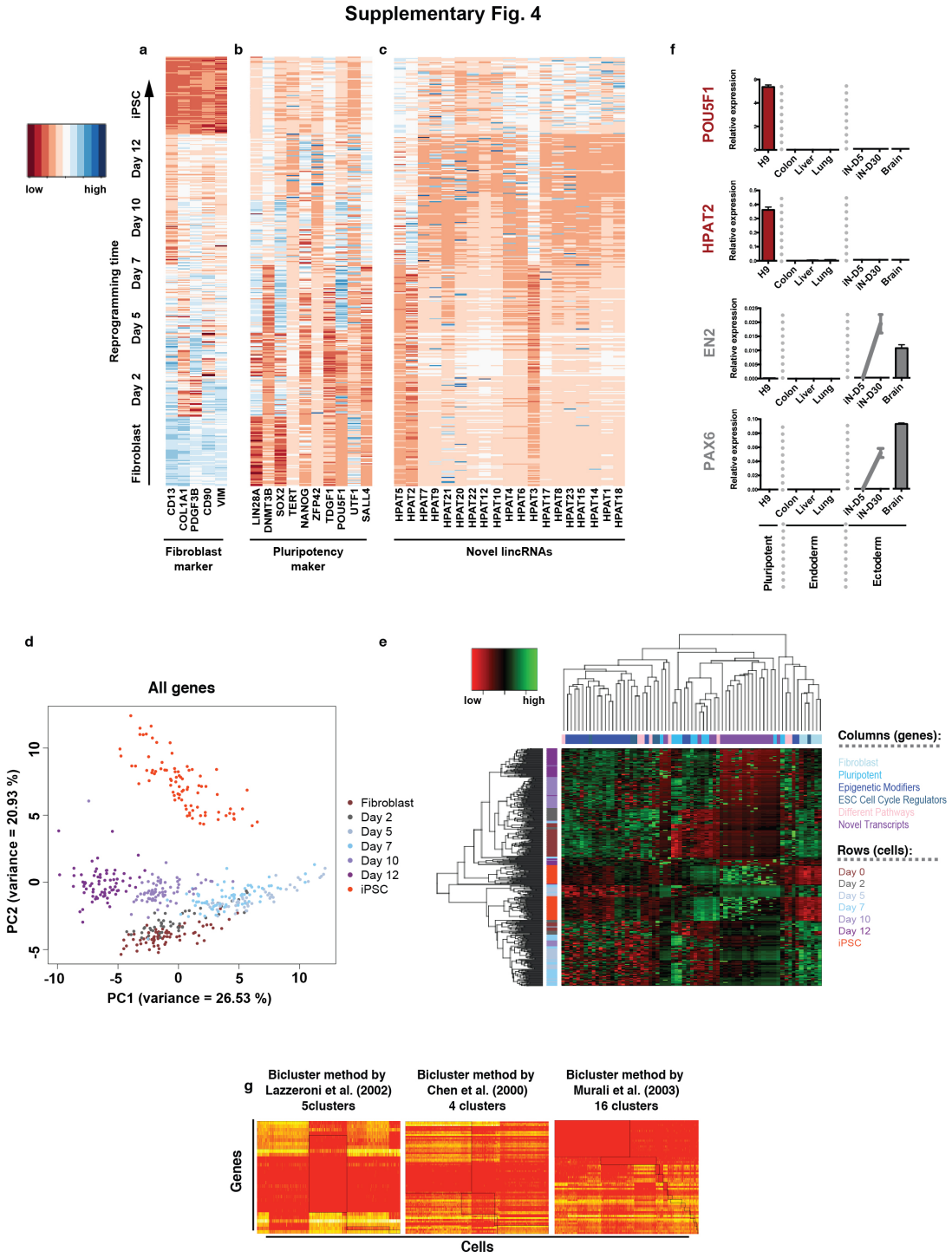
Nature Genetics: doi:10.1038/ng.3449

Supplementary Figure 3

Primer validation and quality control of single-cell gene expression data.

(**a**) Histological sections stained with hematoxylin and eosin from teratomas derived from established iPSCs (iPSCs that resulted from derivation from BJ fibroblasts are termed fully established iPSCs and were used as the last time point for collection (see **b**). (**b**) Tracking of morphological changes of BJ fibroblasts during mRNA reprogramming with the Yamanaka factors. Depicted are the days at which cells were collected. Fibroblasts transfected with GFP only for five consecutive days are shown as well with GFP signal (images are representative). (**c**) Representative example of a dilution series for all 96 assays. C_t values were plotted as a function of the dilution factors (1:2) on a log scale. Linear regression analysis is depicted by the red line. Eight assays with R^2 <0.97 were excluded, thus leaving 88 assays. (**d**) Distribution histogram of calculated primer efficiencies for 88 DELTAgene assays estimated from the slopes of standard curve plots. The average efficiency is 1.02 with s.d. = 0.06. (**e**) Quantile-quantile plot with experimentally estimated efficiencies (*y* axis) and the values expected for a normal distribution with mean efficiency = 1.02 and s.d. = 0.06 (*x* axis). The black line indicates the values expected for a normal distribution ($y = x$). Efficiency values that were derived from plots with three points in the standard curve are depicted in blue. Values derived from plots with >3 points in the standard curve are depicted in red. (**f**) Microscopic view of two capture sides on the C1 Single-Cell Auto Prep System microfluidic chip. The left capture side has no cell, and the right capture side has one captured cell (red arrow). (**g**) Representative example of primer specificity evaluation using melting curve analysis (here with *HPAT2*). The graph shows the relative change in fluorescence signal (EvaGreen) over the temperature range for all 96 cells on a single array. The area in red depicts false positive signals with incorrect melting curve temperatures (determined with bulk RNA and based on the melting curve temperature provided by Fluidigm). The area in green depicts the correct melting curves. Data outside the correct melting curve were set to 0. (**h**) Correlation analysis of mean C_t values generated 96 cells of three dynamic IFC arrays (single cells of (i) BJ fibroblasts, (ii) BJ fibroblasts transfected with mRNA encoding GFP for 2 d, (iii) BJ fibroblasts transfected with mRNA encoding GFP for 5 d). Genes that were detected in at least 20% of the 96 cells for each dynamic IFC array are considered. Shown are all three comparisons. Outliers are shown in green (GFP) and red. The assays in red (total of six) were excluded from subsequent analysis due to a non-correlative pattern among the arrays, leaving the 82 assays that are listed in **Supplementary Table 2**. **i,** Schematic overview of the quality assessment before normalization of single-cell gene expression. Nine dynamic IFC arrays (96.96 Fluidigm chips) were used for gene expression analysis. Two GFP control chips along with one fibroblast chip were used for correlation analysis (**h**) followed by initial quality assessment. Processed chips were used for a second round of quality assessment, resulting in 578 normalized single cells.

Nature Genetics: doi:10.1038/ng.3449



Supplementary Figure 4

Single-cell gene expression analysis during nuclear reprogramming and reactivation of HPAT expression during *in vitro* transdifferentiation from fibroblasts into neurons.

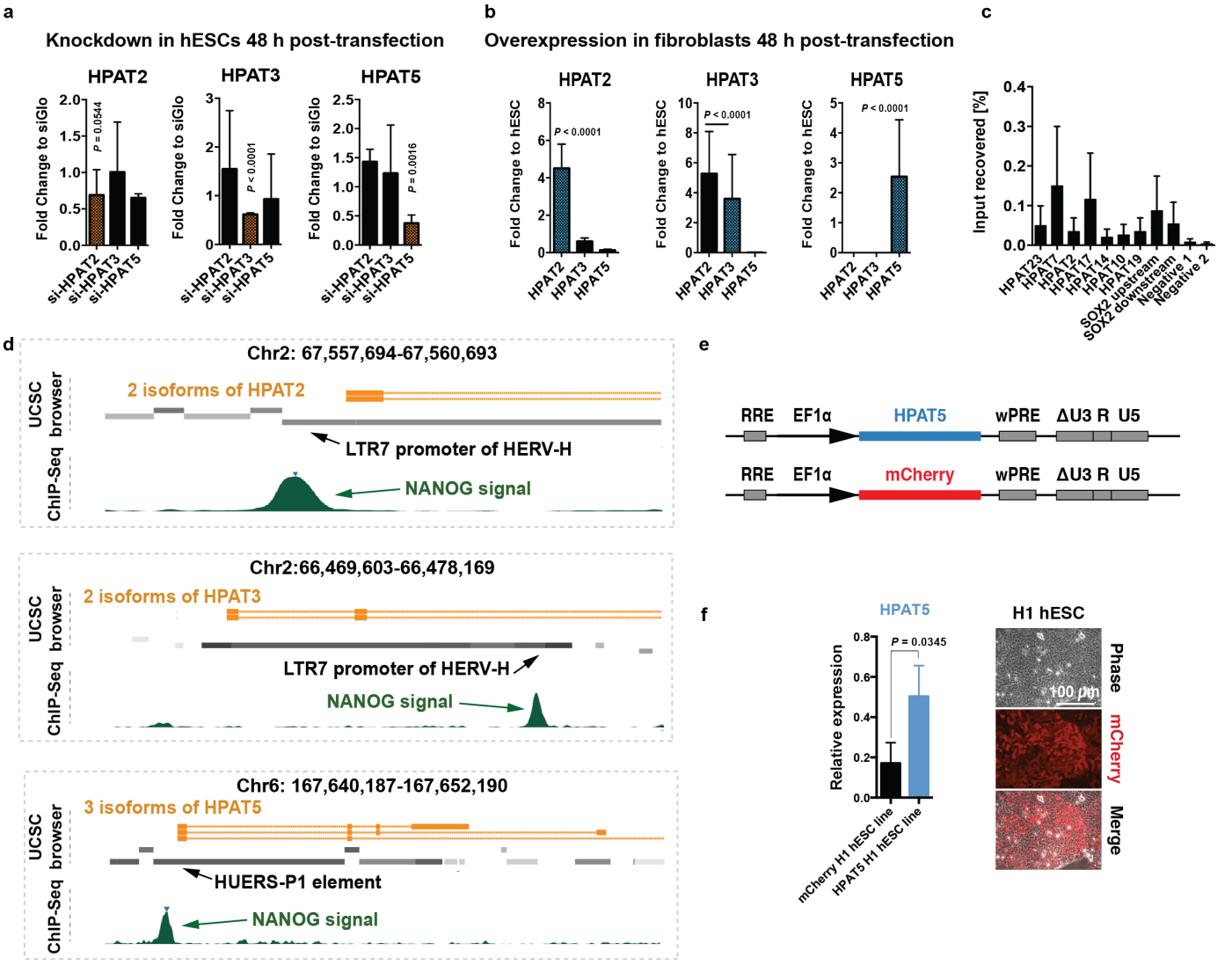
(**a**–**c**) Heat map plot of single-cell gene expression of different markers during nuclear reprogramming. Single cells are in rows. Genes are in columns. Fibroblast markers decrease over time, and pluripotency-specific markers, including selected HPATs, increase over

Nature Genetics: doi:10.1038/ng.3449

time as cell progress toward iPSCs (*n* = 87, 85, 72, 70, 86, 83 and 95 for fibroblasts, day 2, day 5, day 7, day 10, day 12 and iPSCs, respectively). Normalization was performed accordingly (the **Supplementary Note** provides details). White color indicates no expression. **(d**) PCA of 578 single cells collected at different time points during nuclear reprogramming. (**e**) Heat map and unsupervised clustering for 578 single-cell gene expression values resulted in clustering of novel genes implicating a similar biological context during reprogramming. Samples are color-coded according to the specific gene groups (horizontal) and the day at which single cells were collected (vertical). (**f**) The pluripotency marker *POU5F1* (red) and *HPAT2* (red; representative of all HPATs in this study) were exclusively expressed in H9 cells (hESCs) but not in (i) cDNA from colon, liver and lung (endoderm) and (ii) during neuronal transdifferentiation from fibroblasts (gray; samples collected at day 5 and day 30 are labeled iN-D5 and iN-D30, respectively) or cDNA from brain) (all ectoderm). *EN2* and *PAX6*, included as ectoderm control markers, were detected during neuron differentiation and in brain samples (*n* = 3; data shown with s.e.m.). **(g**) Heat map of bicluster analysis illustrating a different bicluster within each plot (**Supplementary Table 3**). Three different algorithms for bicluster calculation were applied, resulting in the identification of five clusters, four clusters and 16 clusters.

Nature Genetics: doi:10.1038/ng.3449

Supplementary Fig. 5



Supplementary Figure 5

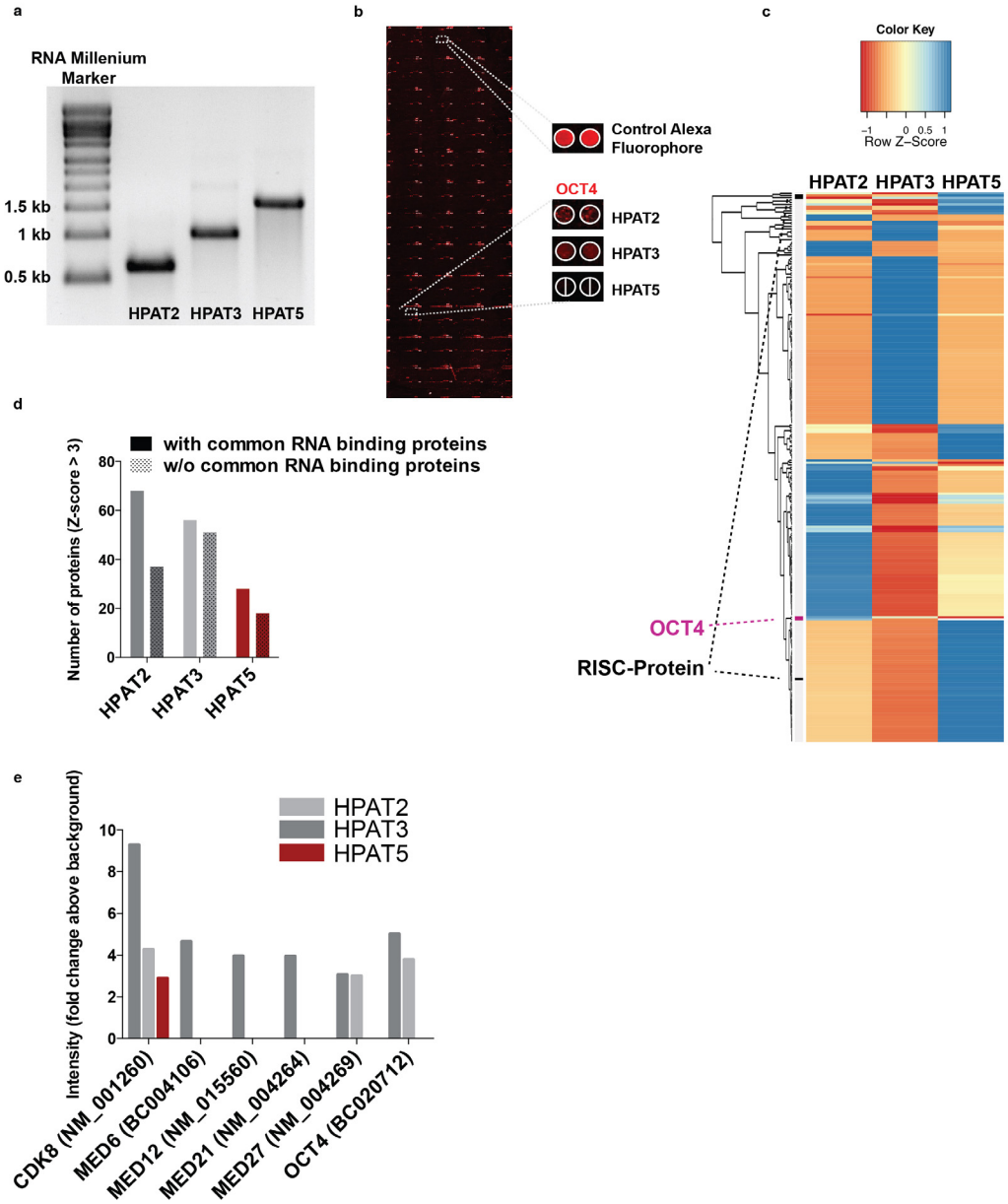
Nature Genetics: doi:10.1038/ng.3449

Overexpression and silencing constructs for *HPAT2*, *HPAT3* and *HPAT5*, NANOG ChIP-seq and regulation of *HPAT5* during hESC differentiation.

(**a**) Validation of siRNAs targeting *HPAT2*, *HPAT3* and *HPAT5*, respectively, in hESCs. Gene expression and *P* values were measured relative to siGlo control 48 h after transfection (*n* = 9). Orange color depicts expected gene downregulation. (**b**) Validation of the overexpression vectors. BJ fibroblasts were transfected with *HPAT2*, *HPAT3* and *HPAT5*. Gene expression and *P* values were measured 48 h after transfection relative to those in GFP-transfected fibroblasts (*n* = 9). Blue color depicts expected gene upregulation. (**c**) ChIP-qPCR analysis in H9 cells (hESCs) using NANOG. Signals were quantified using primer sets specific to a subset of HPATs or two 'negative' intergenic, non-repetitive regions. Two enhancers around *SOX2* are included as positive controls (*n* = 3; data are shown with s.e.m.)**.** (**d**) Three snapshots of the UCSC browser (genome location indicated) aligned with the NANOG-binding region for *HPAT2*, *HPAT3* and *HPAT5* from ChIP-seq analysis. (**e**,**f**) Overexpression constructs and validation of the *HPAT5*-OE and mCherry-OE lines. *HPAT5* was significantly upregulated in hESC-OE cells compared to control cells. mCherry protein expression was also confirmed. *n* = 3; data are sgiwb with s.e.m. (**g**) Increase in differentiation markers representing all three germ layers significantly repressed in *HPAT5*-OE cells. *P* values are calculated for comparison of the mCherry and *HPAT5*-OE lines on the same days.

Nature Genetics: doi:10.1038/ng.3449

Supplementary Fig. 6



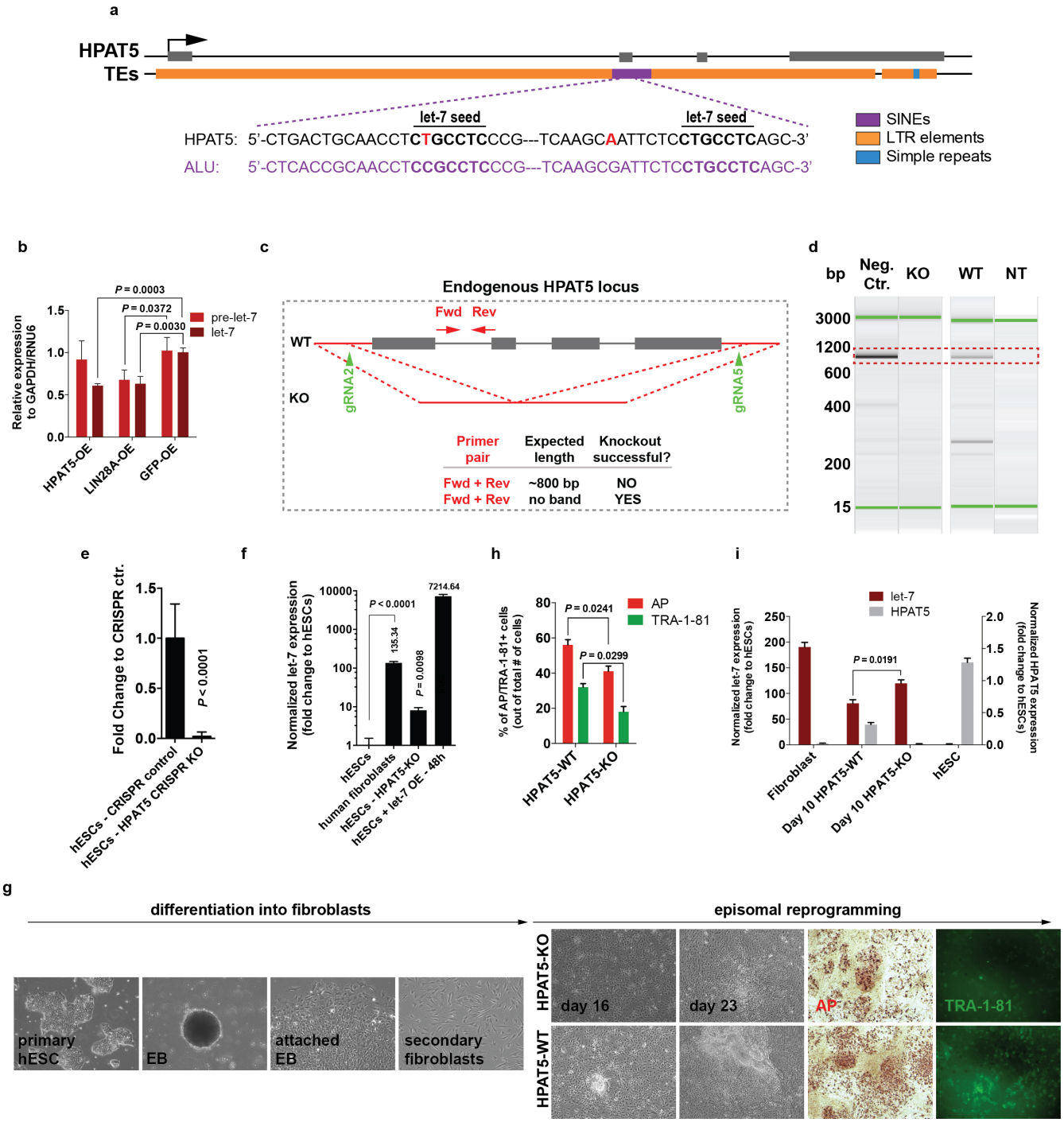
Supplementary Figure 6

Protein microarray with *HPAT2*, *HPAT3* and *HPAT5*.

(**a**) Formaldehyde agarose RNA gel of the Cy5-labeled lincRNAs before hybridization to the protein array. (**b**) Representative image of a ProtoArray and fluorescence intensity for *HPAT2* and *HPAT3* (positive) and *HPAT5* (negative) on OCT4 protein in duplicate. (**c**) Heat map of HPAT2-, HPAT3- and HPAT5-binding proteins with RISC proteins and OCT4 highlighted (*z* score > 2.5). (**d**) Total number of candidate proteins identified with the three HPATs (with and without common RNA-binding proteins). (**e**) Validation of the findings by Lu *et al.* that HERV-H–derived lincRNAs (*HPAT2* and *HPAT3*) bind to specific OCT4, coactivators and mediators.

Nature Genetics: doi:10.1038/ng.3449

Supplementary Fig. 7



Supplementary Figure 7

Loss-of-function analyses in hESCs.

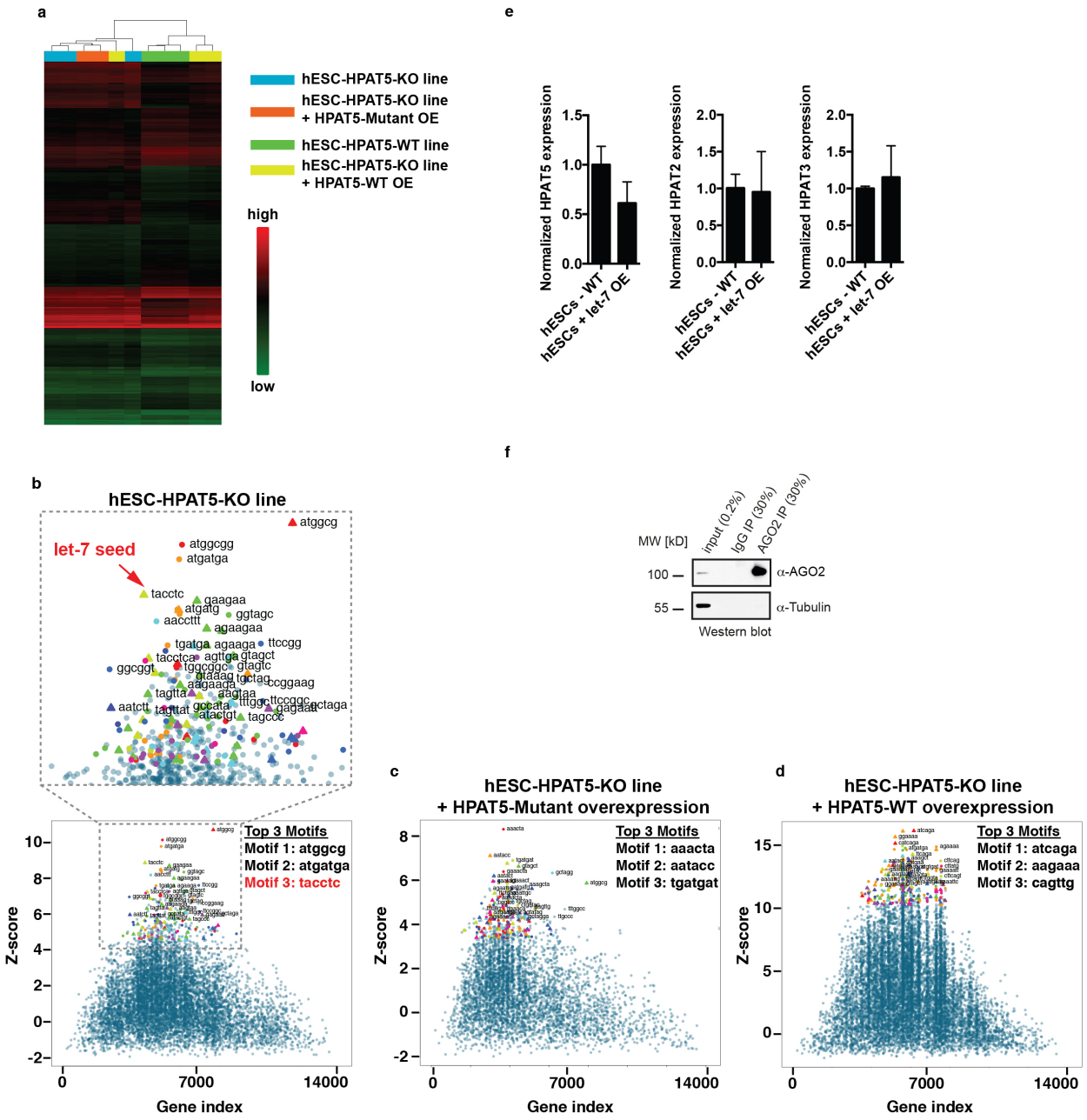
(**a**) Predicted let-7 binding sites in *HPAT5* transcript. Shown is *HPAT5* with embedded TEs along the genomic length (black line). Exons are shown as gray boxes. TEs are shown as colored boxes underneath. let-7 binding sites are within a SINE element (Alu). Bases in red are point mutations and confer HPAT5 specificity. (**b**) Gene expression analysis of endogenous pre-let-7 and mature let-7 in

Nature Genetics: doi:10.1038/ng.3449

fibroblasts. *n* = 3; data are shown with s.e.m. (**c**) Schematic overview of the *HPAT5* locus in genomic DNA from subcloned hESCs that were treated with CRISPR pairs 2 and 5 (gRNA2/5). Forward and reverse primers (in red) were designed to amplify a region of genomic DNA that is inside the deleted *HPAT5* locus. (**d**) Agarose gel illustrating successful derivation of the *HPAT5*-knockout hESC line. Genomic DNA from hESCs (passage 4 after subcloning) did not result in specific amplification. The controls included negative control (treatment only with one CRISPR arm, gRNA2), wild-type hESCs and no-template control (NT). (**e**) Gene expression analysis of endogenous *HPAT5* in hESCs. *n* = 3; data are shown with s.e.m. (**f**) Endogenous let-7 levels do not reach the levels in differentiated cells during 48 h of hESC differentiation. Endogenous let-7 levels are significantly increased 48 h after differentiation with bFGF removal (tenfold). The levels of endogenous let-7 are still significantly higher in human fibroblasts (100-fold) compared to differentiated hESCs. *HPAT5* knockout increases endogenous let-7 levels to ones similar to those found in hESCs differentiated for 24 h. Overexpression of let-7 in hESCs results in a ~50-fold increase compared to human fibroblasts. let-7 levels were normalized to Hs-RNU6-2. *n* = 3; data are shown with s.e.m. (**g**) Differentiation of hESCs into secondary fibroblasts followed by episomal reprogramming into iPSCs. (**h**) Percentage of AP- and TRA-1-81–positive cells in *HPAT5*-WT and *HPAT5*-KO cells 25 d after reprogramming. *n* = 3; data are shown with s.e.m. (**i**) Endogenous let-7 and *HPAT5* levels during nuclear reprogramming at day 10. *n* = 3; data are shwon with s.e.m.

Nature Genetics: doi:10.1038/ng.3449

Supplementary Fig. 8



Supplementary Figure 8

***HPAT5* regulates let-7 in hESCs during differentiation.**

(**a**) Heat map of differentially expressed genes ($P < 0.05$) after let-7 overexpression in four different samples). (**b–d**) Enrichment of let-7 seed sites in transcripts that were downregulated in hESC-*HPAT5*-KO cells. Overexpression from *HPAT5*-WT transcript rescued let-7–mediated differentiation. The Word cluster plot shows sequences in genes ranked by differential expression, after let-7 transfection. Each dot represents a word, summarizing *z* scores, and enrichment specificity indices of the enrichment profiles of negatively correlated 6- and 7-mer words. Triangles annotate known seed sites of human miRNAs. (i) A zoomed-in view (top) from the cluster plot. (**e**) Endogenous *HPAT2*, *HPAT3* and *HPAT5* expression in hESCs with let-7 overexpressed. *n* = 3; data are shown with s.e.m. (**f**)

Nature Genetics: doi:10.1038/ng.3449

Immunoblot confirming specific AGO2 pulldown. OE, overexpression. *n* = 3 samples; data are shown with s.e.m.

Nature Genetics: doi:10.1038/ng.3449

The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming

Jens Durruthy-Durruthy, Vittorio Sebastiano, Mark Wossidlo, Diana Cepeda, Jun Cui, Edward J. Grow, Jonathan Davila, Moritz Mall, Wing Wong, Joanna Wysocka, Kin Fai Au, and Renee A. Reijo Pera

Supplementary Note inventory (as PDF)

The following Supplementary Note is merged in one PDF file:

1) Supplementary Tables 1-4

2) Supplementary Note

3) Supplementary References

Nature Genetics: doi:10.1038/ng.3449

Supplementary Tables 1-8

Supplementary Table 1 - Sequence features of 23 *HPATs* on a genomic level.

HPAT	LTR/ERV elements [%]	Fraction of HERV-H elements [%]*	Bases masked [%]	Length of genomic locus of HPAT [bp]
HPAT1	72.24	66.02	78.33	11334
HPAT2	96.07	100	93.42	7671
HPAT3	22.05	100	45.71	27573
HPAT4	17.21	100	43.14	14035
HPAT5†	91.36	0	97.01	6287
HPAT6	6.70	100	71.55	20308
HPAT7	50.16	55.09	75.12	17553
HPAT8	99.64	92.44	99.64	6714
HPAT9	41.41	100	74.57	29820
HPAT10	38.51	100	54.74	11181
HPAT11	4.93	100	57.75	12027
HPAT12	36.73	100	76.90	7380
HPAT13	37.69	92.83	52.01	21313
HPAT14	36.00	63.64	36.00	1036
HPAT15	97.94	100	97.94	5781
HPAT16	6.88	0	31.38	18694
HPAT17	49.38	51.71	66.23	25174
HPAT18	97.62	100	99.89	5640
HPAT19	56.60	100	72.02	7613
HPAT20	4.63	0	75.79	14359
HPAT21	88.84	100	88.84	4220
HPAT22	53.66	0	59.39	2145
HPAT23	65.84	100	67.34	8130

Controls	Found ERV class 1 repeats [%]	Fraction of HERV-H elements [%]*	Bases masked [%]	Length of genomic locus of HPAT [bp]
NANOG	0.00	0	30.91	6661
POU5F1	1.56	0	29.82	6337
SOX2	0.00	0	4.06	2512

Supplementary Table 1. Sequence features of 23 *HPATs* on a genomic level. All *HPATs* share repeats from the LTR/ERV class with the majority being HERV-H elements. Control genes are included for comparison. * Fraction extracted from column 2. † Contained large fraction of HUERS-P1 elements. Analysis was performed with RepeatMasker (www.repeatmasker.org).

Nature Genetics: doi:10.1038/ng.3449

Supplementary Table 2 – Assays that passed quality tests and used for single-cell gene expression analysis; Primer sequences used in this study.

Gene	Category	Gene	Category	Gene	Category	Gene	Category
RPLP0 GAPDH HSP90AB1 ACTB HPRT1 POU5F1* SOX2* LIN28A* EGFP*	Internal/ External Control	LIN28A DNMT3B SOX2 TERT NANOG ZFP42 TDGF1 POU5F1 UTF1 SALL4	Pluripotency Marker†	BPTF CBX7 DNMT1 EED GLP G9A P300 EZH2 JARID2 KDM3B MBD3 MCRS1 MLL2 RING1B BRG1 SNF2H HP1 TAF1 TET1 THAP11 WDR5 YY1 PRMT1	Epigenetic Regulator	HPAT1 HPAT2 HPAT3 HPAT4 HPAT5 HPAT6 HPAT7 HPAT8 HPAT9 HPAT10 HPAT12 HPAT14 HPAT15 HPAT16 HPAT17 HPAT18 HPAT19 HPAT20 HPAT21 HPAT22 HPAT23	Novel TE-derived lincRNAs**
CD13 COL1A1 PDGFRB CD90 VIM	Fibroblast Marker	CDH1 CDKN2A GRB2 LEFTY2 LMNB1 MAPK1 MAPK3 P53	Differentiation Pathways				
BUB1 CDC20 CDKN1A LATS2 MAD2L1 RBL1	Pluripotent Cell Cycle Regulator	CDX2 EpCAM TROP2 KRT7 TEAD4	Trophectoderm†				

Supplementary Table 2. Assays that passed quality tests and used for single-cell gene expression analysis. Genes were selected based on literature data and to aid to examine putative functions of novel genes related to pluripotency establishment and/or maintenance. *Assays designed in order to exclusively detect exogenous transcripts used during iPSC derivation. **Assays are the 21 most abundant expressed previously identified[13]. † Assays used for single-cell analysis in human embryos.

Nature Genetics: doi:10.1038/ng.3449

Supplementary Table 3 - Bi-cluster and correlation analysis on single cells during reprogramming

Bicluster method by Cheng *et al.* (2000)	Genes	# of cells
Bicluster C1 (43 x 96)	BPTF, CBX7, CD90, CDKN1A, DNMT1, EED, G9A, GLP, GRB2, HP1, HPAT1, HPAT6-8, HPAT10, HPAT12, HPAT14-15, HPAT17-18, HPAT21-23, JARID2, KDM3B, LATS2, MAD2L1, MAPK1, MCRS1, MLL2, NANOG, NPM1, P300, P53, PDGFRB, SALL4, SNF2H, TAF1, TDGF1, TERT, VIM, WDR5, ZFP42	96
Bicluster C2 (13 x 48)	BRG1, BUB1, CDC20, CDKN2A, EZH1, **HPAT5**, LEFTY2, LMBN1, MBD3, RBL1, TET1, THAP11, UTF1	48
Bicluster C3 (7 x 33)	CD13, COLA1A, **HPAT2**, HPAT4, HPAT9, LIN28A, POU5F1	33
Bicluster C4 (5 x 87)	CDH1, DNMT3B, **HPAT3**, MAPK3, SOX2	87

Bicluster method by Murali *et al.* (2003)	Genes	# of cells
Bicluster X1 (18 x 242)	CBX7, HPAT1, HPAT4, HPAT6, HPAT7, HPAT8, HPAT9, HPAT10, HPAT12, HPAT14, HPAT15, HPAT17, HPAT18, HPAT22, HPAT23, P53, TERT, ZFP42	242
Bicluster X2 (4 x 178)	MAD2L1, MAPK1, RBL1, UTF1	178
Bicluster X3 (5 x 104	CDH1, DNMT3B, **HPAT3**, **HPAT5**, HPAT21	104
Bicluster X4 (4 x 78)	**HPAT2**, NANOG, POU5F1, SALL4	78
Bicluster X5 (4 x 75)	CD13, CDKN1A, COL1A1, EED	75
Bicluster X6 (3 x 82)	BRG1, GRB2, SNF2H	82
Bicluster X7 (2 x 88)	HP1, WDR5	88
Bicluster X8 (2 x 76)	CD90, PDGFRB	76
Bicluster X9 (2 x 72)	MLL2, P300	72
Bicluster X10 (2 x 71)	LMNB1, MCRS1	71
Bicluster X11 (2 x 71)	BUB1, EZH2	71
Bicluster X12 (2 x 69)	LIN28A, NPM1	69
Bicluster X13 (2 x 69)	BPTF, TET1	69
Bicluster X14 (2 x 61)	DNMT1, GLP	61
Bicluster X15 (2 x 56)	TAF1, THPA11	56
Bicluster X16 (2 x 56)	CDC20, G9A	56

Supplementary Table 3-1. Bi-cluster analysis with CC and Xmotif method on single-cells. Two methods are described by Cheng *et al.*[57] and Murali *et al.*[58], respectively. Number of bi-clusters (first column) and their respective collection of genes (second column) and number of cells (third column). Novel genes HPAT2, 3, and 5, highlighted in orange, are found in bi-clusters correlated to key pluripotency markers.

Fibroblast			
Gene 1	**Gene 2**	**Pearson cor. Coef. R^2**	
CD13	PDGFRB	0.7244301	
CDKN2A	LEFTY2	0.7206901	
DNMT1	WDR5	0.7192214	
VIM	SNF2H	0.7161610	
PDGFRB	CD90	0.6758941	
BUB1	CDC20	0.6609234	
BUB1	LMNB1	0.6589819	
MAD2L1	LMNB1	0.6563781	
NANOG	TDGF1	0.6545420	
BUB1	MAD2L1	0.6498815	
SNF2H	GRB2	0.6392659	
COL1A1	UTF1	0.6242778	
GRB2	MAPK1	0.6165660	
CD13	COL1A1	0.6118486	
VIM	GRB2	0.6112345	
Day 2			
Gene 1	**Gene 2**	**Pearson cor. Coef. R^2**	
GRB2	NPM1	0.7715967	
VIM	NPM1	0.7477494	
CD90	DNMT1	0.7264469	
MBD3	NPM1	0.7262332	
CD13	VIM	0.7181578	
BUB1	LMNB1	0.7110734	
VIM	WDR5	0.7108301	
LIN28A	SOX2	0.7063501	
MBD3	SNF2H	0.7032002	
SNF2H	NPM1	0.6956402	
VIM	DNMT1	0.6793156	

Nature Genetics: doi:10.1038/ng.3449

EZH2	RBL1	0.6708873	
COL1A1	PDGFRB	0.6670295	
SNF2H	GRB2	0.6662780	
BRG1	NPM1	0.6581776	
		Day 5	
Gene 1	**Gene 2**	**Pearson cor. coef. R^2**	
LIN28A	SOX2	0.9535118	
BPTF	P300	0.8615977	
CD90	P300	0.8471875	
BPTF	JARID2	0.8333670	
CD13	PDGFRB	0.8260145	
TAF1	CDKN2A	0.8194652	
GRB2	NPM1	0.8088202	
CD90	BPTF	0.8003200	
PDGFRB	BPTF	0.7912487	
P300	JAIRD2	0.7670918	
P300	MLL2	0.7628003	
VIM	BPTF	0.7594642	
SNF2H	NPM1	0.7588008	
CDKN2A	LEFTY2	0.7530967	
G9A	KDM3B	0.7528663	
		Day 7	
Gene 1	**Gene 2**	**Pearson cor. coef. R^2**	
HPAT12	HPAT17	0.9390666	
SOX2	POU5F1	0.9104285	
LIN28A	POU5F1	0.8665226	
HPAT12	HPAT4	0.7564206	
MAD2L1	HPAT12	-0.7285055	
HPAT4	HPAT17	0.7028677	
MAD2L1	NPM1	0.7017917	
MAD2L1	HPAT4	-0.6962021	
CD13	PDGFRB	0.6899112	
NANOG	UTF1	0.6736726	
VIM	P300	0.6410589	
BRG1	THAP11	0.6333262	
COL1A1	PDGFRB	0.6072682	
BPTF	P300	0.6015609	
SNF2H	NPM1	0.5990415	
		Day 10	
Gene 1	**Gene 2**	**Pearson cor. coef. R^2**	**References**
SOX2	POUF51	0.9652009	Rizzino (2013)[59], Fong *et al.* (2011)[60], Chew *et al.* (2005)[61], Wang *et al.* (2012)[62], Chambers *et al.* (2009)[63]
HPAT23	HPAT14	0.8416919	Novel
THAP11	CDKN2A	0.6589637	Novel
HPAT6	HPAT14	0.6489457	Novel
P300	MLL2	0.6221004	Jiang et al. (2013)[64]
TDGF1	CDH1	0.6203397	Novel
CDKN2A	LEFTY2	0.6125083	Kim *et al.* (2014)[65,*]
NANOG	CDC20	0.6118823	Novel
THAP11	GRB2	0.6087451	Novel
DNMT1	EED	0.6057788	Jin et al. (2009)[66,*]
EED	THAP11	0.6048661	Dejosez et al. (2008)[67,*]
DNMT1	CDKN2A	0.6024900	Robert et al. (2002)[68]
GRB2	NPM1	0.6013000	Zhao et al. (2013)[69], Fujimoto et al. (1996)[70]
NANOG	EED	0.6010966	Denholtz et al. (2013)[71], Villasante et al. (2011)[72]
HPAT6	HPAT23	0.5551024	Novel
		Day 12	
Gene 1	**Gene 2**	**Pearson cor. coef. R^2**	**References**
PDGFRB	CD90	0.8149248	Hewitt *et al.* (2012)[73]
LIN28A	POU5F1	0.7726876	Yu *et al.* (2007)[74], Qiu *et al.* (2010)[75]
LIN28A	SOX2	0.7512026	Qiu *et al.* (2010)[74], Cimadamore *et al.* (2013)[76]
SOX2	POU5F1	0.7510157	Rizzino (2013)[59], Fong *et al.* (2011)[60], Chew *et al.* (2005)[61], Wang *et al.* (2012)[62], Chambers *et al.* (2009)[63]
COL1A1	PDGFRB	0.6942359	Takahira *et al.* (2007)[77], Hewitt *et al.* (2011)[78]
COL1A1	CD90	0.6309951	Hewitt *et al.* (2012)[73]
UTF1	THAP11	0.6214731	Novel
CD90	VIM	0.6047015	Hewitt *et al.* (2012)[73]
BPTF	JARID2	0.5917150	Landry *et al.* (2008)[79,*], Gaspar-Maia *et al.* (2011)[80,*]
CDKN2A	LEFTY2	0.5746193	Kim *et al.* (2014)[65,*]
VIM	SNF2H	0.5745164	Gaspar-Maia *et al.* (2011)[80,*]
CD90	SNF2H	0.5614012	Gaspar-Maia *et al.* (2011)[80,*]
CD13	PDGFRB	0.5598867	Hewitt *et al.* (2011)[78]
TDGF1	TET1	0.5589469	Novel
NANOG	SNF2H	0.5288758	Gaspar-Maia *et al.* (2011)[80,*]

iPSCs

Nature Genetics: doi:10.1038/ng.3449

Gene 1	Gene 2	Pearson cor. coef. R^2	References
HPAT2	HPAT3	0.8851197	Novel
DNMT3B	JARID2	0.8550930	Assou *et al.* (2009)[81]; Pasini *et al.* (2010)[82], Peng *et al.* (2009)[83]
TDGF1	SALL4	0.8530899	Kidder *et al.* (2008)[84]
SALL4	GRB2	0.8383334	Hamazaki *et al.* (2006)[85], Lanner *et al.* (2010)[86]
SALL4	SNF2H	0.8313401	Kuijk *et al.* (2011)[87]
SALL4	P300	0.8312009	Chitilian *et al.* (2014)[88], Zhong *et al.* (2009)[89]
WDR5	GRB2	0.8280270	Ang *et al.* (2011)[90], Yang *et al.* (2014)[91],†, Lanner *et al.* (2010)[86]
TDGF1	DNMT1	0.8266580	Vassena *et al.* (2011)[92], Hochedlinger *et al.* (2009)[93]
SOX2	SALL4	0.8266580	Tanimura et a. (2013)[94], Yang *et al.* (2008)[95]
DNMT3B	SALL4	0.8210343	Yang *et al.* (2012)[96], Tan *et al.* (2012)[97]
SNF2H	GRB2	0.8208503	Kuijk *et al.* (2011)[87]
SALL4	G9A	0.8198704	Kuijk *et al.* (2011)[87]
GRB2	NPM1	0.8194229	Johansson *et al.* (2010)[98],*
TDGF1	GRB2	0.8184948	Kuijk *et al.* (2011)[87]
NANOG	POU5F1	0.8141164	Loh *et al.* (2006)[29], Wang *et al.* (2012)[62]
		All 578 cells	
Gene 1	**Gene 2**	**Pearson cor. coef. R^2**	**References**
DNMT3B	HPAT3	0.9168379	Novel
DNMT3B	HPAT5	0.8865738	Novel
TDGF1	HPAT3	0.8760132	Novel
VIM	CDKN1A	0.8667649	Li *et al.* (2014)[99], Mergui *et al.* (2010)[100]
DNMT3B	SALL4	0.8661937	Yang *et al.* (2012)[96], Tan *et al.* (2012)[97]
SALL4	HPAT3	0.8618967	Novel
HPAT5	HPAT3	0.8558339	Novel
P300	MLL2	0.8554477	Jiang *et al.* (2013)[64]
HPAT2	HPAT3	0.8403645	Novel
DNMT3B	TDGF1	0.8390348	Tan *et al.* (2013)[97]
CDH1	HPAT3	0.8378528	Novel
TDGF1	SALL4	0.8301321	Kidder *et al.* (2008)[84], Vassena *et al.* (2011)[92]
LIN28A	SOX2	0.8284545	Cimadamore *et al.* (2012, 2013)[76,101]
DNTM3B	HPAT2	0.8235213	Novel
DNMT3B	CDH1	0.8229433	Rahnama *et al.* (2009)[102], Kwon *et al.* (2010)[103]

Supplementary Table 3-2. Correlation analysis on single-cell populations. Top 15 correlated gene pairs (first two columns) for each collected single cell population during nuclear reprogramming and their correlation coefficient (third column). Novel genes HPAT2, 3 and 5, highlighted in orange, are found be among the top 15 correlated gene pairs late during reprogramming. Fourth column (day 10 and later stage) validates correlation analysis and shows previously reported (in)direct interactions between gene pairs. * are not directly associating both genes, though suggest putative interactions. † involve reported lincRNAs.

Supplementary Table 4 – Target-specific siRNA sequences

Target transcript	Sense siRNA sequence
HPAT2	CCCAGAUCUUCUCGGCUUAUU
	UCCCAAGGUCAUACCGCAAUU
	CAUCACGGACGCCGAGCUUUU
	GGAGGCAGGAGGAUCGCUUUU
HPAT3	GAAAUUUGGUGCCGUGACUUU
	GAGAAAGAUCCACCUACGAUU
HPAT5	UAACAAUAACCACGAGAUAUU
	GAGAAAGGGAGCCCGGAAAUU
	CAGGAAGACGCCAGAGCGAUU

Supplementary Table 4. Target-specific siRNA sequences used for knockdown experiments.

Supplementary Table 5 – refer to Excel file – ChIP analysis with NANOG

Supplementary Table 6 – refer to Excel file – Protein microarray analysis

Supplementary Table 7 – refer to Excel file – Prediction of miRNA binding sites

Supplementary Table 8 – refer to Excel file – Microarray and cWords analysis

Nature Genetics: doi:10.1038/ng.3449

Supplementary Note

Bi-cluster analysis on single cells

In an effort to retrieve subgroups within each cell population, we applied bi-cluster analysis on our high dimensional dataset. Bi-cluster analysis resolves local rather than global gene association patterns and identifies gene sets with related expression motifs across subsets of cells[104]. We explored 3 different algorithms (Methods section for detail) for bi-cluster identification, as existing methods are extremely sensitive to parameter and data variation, thus constituting a challenge to rely solely on one method. The Plaid method defined by Lazzeroni and Owen[52] assembled the data into five clusters (P1-P5) (Fig. 2d); the algorithm by Cheng and Church[57] identified four clusters (C1-C4, Supplementary Fig. 4j, Supplementary Table 3) and the Xmotifs algorithm[58], generated 16 clusters (X1-X16, Supplementary Fig. 4j, Supplementary Table 3). The latter searches for genes with constant values over a set of different single cells, thus revealing "conserved gene expression motifs". We found that *HPAT2*, *HPAT3* and *HPAT5* were identified in all three algorithms with expression patterns highly correlative to key pluripotency markers including *SALL4, POU5F1, SOX2, DNMT3B* and *NANOG*. For instance, cluster X3 and X4 consisted of 104 and 78 cells, respectively, and correlated *HPAT3/HPAT5* with *CDH1/DNMT3B* and *HPAT2* with *SALL4/POU5F1/NANOG*, respectively. Cluster P1, P2, P4 and P5 included 111, 81, 94 and 24 cells, respectively, and collectively identified subpopulations that uniquely express *HPAT2, HPAT3* and *HPAT5* in a correlative manner to above mentioned key pluripotency markers, including *SALL4*. As internal validation, 83 cells grouped in cluster P3 were of fibroblast origin as they expressed fibroblast specific marker genes including *CD13, COL1A1, VIM* and *PDGFRB*.

Correlation analysis on single cells

We then reasoned that single-cell expression data could be used to identify pairs of genes with correlated expression and reveal regulatory linkages, as recently shown[25]. Correlation analysis across all 578 cells revealed one group of genes (C1), that included five positively correlated

Nature Genetics: doi:10.1038/ng.3449

genes (*CD13, CD90, COL1A1, VIM* and *PDGFRB*) consistent with previous reports[73] (Fig. 2d). C2, a second group included *POU5F1, SOX2* and *LIN28A,* known to be crucial for pluripotency establishment and maintenance[59,63]; *NANOG* and *TET1*, crucial for pluripotency establishment and shown to physically interact with each other[105] positively correlated in group C3. In contrast, negative correlation was observed between *POU5F1/SOX2/LIN28A* (that maintain self-renewal and pluripotency in ESCs/iPSCs) and *MAPK3* (C4), which upon activation, triggers differentiation [86,106]. Most interestingly, the majority of novel lincRNAs including *HPAT2, HPAT3*, and *HPAT5* correlated positively with each other (group C5) and negatively with fibroblast specific markers (group C6) suggesting a common role during reprogramming. We located *SALL4, CDH1* and *DNMT3B* within that same group validating our bi-cluster analysis. Among the top 15 most correlated gene pairs, *HPAT3* correlated with *DNMT3B, TDGF1, SALL4, HPAT5, HPAT2* and *CDH1* (Supplementary Table 3) at late stages during reprograming, indicating coordinated expression and a central role of this particular novel transcript. See Supplementary Table 3 for top 15 correlated gene pairs for different time points of single-cell collection. Correlations that were found late during reprogramming (day 10 – iPSCs) were consistent with previous reports (references in fourth column).

Protein microarray assays

Protein microarrays are chips containing more than 9,400 human recombinant proteins spotted in duplicate. *HPAT-2, -3* and *-5* were *in vitro* transcribed (Supplementary Fig. 6a), labeled with Cy5 and independently probed on two human protein microarrays. We included *HPAT-2* and *-3*, both HERV-H-containing lincRNAs, to validate previous findings[17] as well as to identify differences with *HPAT5*, We adopted a previously published protocol[35] for optimal labeling conditions such that 3 pmol dye per μg RNA with an average efficacy of 1 dye molecule per 850 bp RNA was achieved to minimize modification of native RNA structures while yielding signal intensities that are readily visualized. Each RNA-species was probed in two technical replicates (denatured and non-denatured). By selecting for the most stringent RNA-protein binding events (minimum signal

Nature Genetics: doi:10.1038/ng.3449

above background of 2.5 fold, Z-score > 2), 68, 56 and 28 binding events for *HPAT-2, -3* and -*5*, respectively were identified (Supplementary Fig. 6c). To further reduce the list of candidate RNA-protein interactions we removed common RNA binding-proteins that contained known RNA binding motifs (Additional Table 1 in Siprashvili *et al*.[35]). Neither known pluripotency-associated proteins nor proteins related to chromatin modification complexes (such as PCR2) were among the remaining list (Supplementary Table 6) which confirms a recent report[17] and possibly explains that no significant binding sites genome wide were found during ChIRP analysis. We then asked whether *HPAT-2* and -*3*, both HERV-H derived lincRNAs, bind to OCT4 and other mediators as previously described. 6 out of 10 proteins that were used in this study were printed onto the protein array (Supplementary Fig. 6d). Notably, *HPAT2* bound to all 6 proteins (OCT4, CDK8, MED6, MED12, MED21 and MED27), validating our approach. *HPAT3* specifically bound to 3 out of 6 proteins including OCT4. This is in contrast to *HPAT5*, a non-HERVH derived lincRNA, that only bound to 1 out of 6 (CDK8) proteins, underlining that it represents a different class of lincRNA. Taken together our findings confirm previously published results, represent a large resource to study specific *HPAT*-protein interactions and suggest that each retroviral-derived lincRNA is likely to be implicated in numerous physiological pathways and belong to many phenotypic classes not necessarily related to pluripotency maintenance[20]. Moreover our data indicate that *HPAT5* that bound to the fewest number of proteins, of which most were common RNA binders, is distinct from *HPAT2* and *HPAT3*.

Nature Genetics: doi:10.1038/ng.3449

Supplementary References

53 Bengtsson, M., Stahlberg, A., Rorsman, P. & Kubista, M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res.* **15**, 1388-1392 (2005).

54 Pang, Z. P. *et al.* Induction of human neuronal cells by defined transcription factors. *Nature* **476**, 220-223 (2011).

55 Goff, L. A. *et al.* Ago2 immunoprecipitation identifies predicted microRNAs in human embryonic stem cells and neural precursors. *PLoS One* **4**, e7192 (2009).

56 Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5**, 877-879 (2008).

57 Cheng, Y. & Church, G. M. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* **8**, 93-103 (2000).

58 Murali, T. M. & Kasif, S. Extracting conserved gene expression motifs from gene expression data. *Pac Symp Biocomput*, 77-88 (2003).

59 Rizzino, A. Concise review: The Sox2-Oct4 connection: critical players in a much larger interdependent network integrated at multiple levels. *Stem Cells* **31**, 1033-1039 (2013).

60 Fong, Y. W. *et al.* A DNA repair complex functions as an Oct4/Sox2 coactivator in embryonic stem cells. *Cell* **147**, 120-131 (2011).

61 Chew, J. L. *et al.* Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol. Cell. Biol.* **25**, 6031-6046 (2005).

62 Wang, Z., Oron, E., Nelson, B., Razis, S. & Ivanova, N. Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell Stem Cell* **10**, 440-454 (2012).

63 Chambers, I. & Tomlinson, S. R. The transcriptional foundation of pluripotency. *Development* **136**, 2311-2322 (2009).

64 Jiang, H. *et al.* Regulation of transcription by the MLL2 complex and MLL complex-associated AKAP95. *Nat. Struct. Mol. Biol.* **20**, 1156-1163 (2013).

65 Kim, D. K., Cha, Y., Ahn, H. J., Kim, G. & Park, K. S. Lefty1 and lefty2 control the balance between self-renewal and pluripotent differentiation of mouse embryonic stem cells. *Stem Cells Dev* **23**, 457-466 (2014).

66 Jin, B. *et al.* DNMT1 and DNMT3B modulate distinct polycomb-mediated histone modifications in colon cancer. *Cancer Res.* **69**, 7412-7421 (2009).

67 Dejosez, M. *et al.* Ronin is essential for embryogenesis and the pluripotency of mouse embryonic stem cells. *Cell* **133**, 1162-1174 (2008).

68 Robert, M. F. *et al.* DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. *Nat. Genet.* **33**, 61-65 (2003).

69 Zhao, H. *et al.* Novel modeling of cancer cell signaling pathways enables systematic drug repositioning for distinct breast cancer metastases. *Cancer Res.* **73**, 6149-6163 (2013).

70 Fujimoto, J. *et al.* Characterization of the transforming activity of p80, a hyperphosphorylated protein in a Ki-1 lymphoma cell line with chromosomal translocation t(2;5). *Proc. Natl. Acad. Sci. U. S. A.* **93**, 4181-4186 (1996).

71 Denholtz, M. *et al.* Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell Stem Cell* **13**, 602-616 (2013).

72 Villasante, A. *et al.* Epigenetic regulation of Nanog expression by Ezh2 in pluripotent stem cells. *Cell Cycle* **10**, 1488-1498 (2011).

73 Hewitt, K. J. *et al.* PDGFRbeta expression and function in fibroblasts derived from pluripotent cells is linked to DNA demethylation. *J. Cell Sci.* **125**, 2276-2287 (2012).

74 Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917-1920 (2007).

Nature Genetics: doi:10.1038/ng.3449

75 Qiu, C., Ma, Y., Wang, J., Peng, S. & Huang, Y. Lin28-mediated post-transcriptional regulation of Oct4 expression in human embryonic stem cells. *Nucleic Acids Res.* **38**, 1240-1248 (2010).
76 Cimadamore, F., Amador-Arjona, A., Chen, C., Huang, C. T. & Terskikh, A. V. SOX2-LIN28/let-7 pathway regulates proliferation and neurogenesis in neural precursors. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E3017-3026 (2013).
77 Takahira, T. *et al.* Detection of COL1A1-PDGFB fusion transcripts and PDGFB/PDGFRB mRNA expression in dermatofibrosarcoma protuberans. *Mod. Pathol.* **20**, 668-675 (2007).
78 Hewitt, K. J. *et al.* Epigenetic and phenotypic profile of fibroblasts derived from induced pluripotent stem cells. *PLoS One* **6**, e17128 (2011).
79 Landry, J. *et al.* Essential role of chromatin remodeling protein Bptf in early mouse embryos and embryonic stem cells. *PLoS Genet* **4**, e1000241 (2008).
80 Gaspar-Maia, A., Alajem, A., Meshorer, E. & Ramalho-Santos, M. Open chromatin in pluripotency and reprogramming. *Nat. Rev. Mol. Cell Biol.* **12**, 36-47 (2011).
81 Assou, S. *et al.* A gene expression signature shared by human mature oocytes and embryonic stem cells. *BMC Genomics* **10**, 10 (2009).
82 Pasini, D. *et al.* JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature* **464**, 306-310 (2010).
83 Peng, J. C. *et al.* Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* **139**, 1290-1302 (2009).
84 Kidder, B. L., Yang, J. & Palmer, S. Stat3 and c-Myc genome-wide promoter occupancy in embryonic stem cells. *PLoS One* **3**, e3932 (2008).
85 Hamazaki, T., Kehoe, S. M., Nakano, T. & Terada, N. The Grb2/Mek pathway represses Nanog in murine embryonic stem cells. *Mol. Cell. Biol.* **26**, 7539-7549 (2006).
86 Lanner, F. & Rossant, J. The role of FGF/Erk signaling in pluripotent cells. *Development* **137**, 3351-3360 (2010).
87 Kuijk, E. W., Chuva de Sousa Lopes, S. M., Geijsen, N., Macklon, N. & Roelen, B. A. The different shades of mammalian pluripotent stem cells. *Hum. Reprod. Update* **17**, 254-271 (2011).
88 Chitilian, J. M. *et al.* Critical components of the pluripotency network are targets for the p300/CBP interacting protein (p/CIP) in embryonic stem cells. *Stem Cells* **32**, 204-215 (2014).
89 Zhong, X. & Jin, Y. Critical roles of coactivator p300 in mouse embryonic stem cell differentiation and Nanog expression. *J. Biol. Chem.* **284**, 9168-9175 (2009).
90 Ang, Y. S. *et al.* Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell* **145**, 183-197 (2011).
91 Yang, Y. W. *et al.* Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. *Elife* **3**, e02046 (2014).
92 Vassena, R. *et al.* Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Development* **138**, 3699-3709 (2011).
93 Hochedlinger, K. & Plath, K. Epigenetic reprogramming and induced pluripotency. *Development* **136**, 509-523 (2009).
94 Tanimura, N., Saito, M., Ebisuya, M., Nishida, E. & Ishikawa, F. Stemness-related factor Sall4 interacts with transcription factors Oct-3/4 and Sox2 and occupies Oct-Sox elements in mouse embryonic stem cells. *J. Biol. Chem.* **288**, 5027-5038 (2013).
95 Yang, J. *et al.* Genome-wide analysis reveals Sall4 to be a major regulator of pluripotency in murine-embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 19756-19761 (2008).
96 Yang, J., Corsello, T. R. & Ma, Y. Stem cell gene SALL4 suppresses transcription through recruitment of DNA methyltransferases. *J. Biol. Chem.* **287**, 1996-2005 (2012).
97 Tan, M. H. *et al.* An Oct4-Sall4-Nanog network controls developmental progression in the pre-implantation mouse embryo. *Mol Syst Biol* **9**, 632 (2013).

Nature Genetics: doi:10.1038/ng.3449

98 Johansson, H. & Simonsson, S. Core transcription factors, Oct4, Sox2 and Nanog, individually form complexes with nucleophosmin (Npm1) to control embryonic stem (ES) cell fate determination. *Aging (Albany NY)* **2**, 815-822 (2010).
99 Li, X. L. *et al.* A p21-ZEB1 complex inhibits epithelial-mesenchymal transition through the microRNA 183-96-182 cluster. *Mol. Cell. Biol.* **34**, 533-550 (2014).
100 Mergui, X. *et al.* p21Waf1 expression is regulated by nuclear intermediate filament vimentin in neuroblastoma. *BMC Cancer* **10**, 473 (2010).
101 Cimadamore, F. *et al.* SOX2 modulates levels of MITF in normal human melanocytes, and melanoma lines in vitro. *Pigment Cell Melanoma Res* **25**, 533-536 (2012).
102 Rahnama, F. *et al.* Epigenetic regulation of E-cadherin controls endometrial receptivity. *Endocrinology* **150**, 1466-1472 (2009).
103 Kwon, O. *et al.* Modulation of E-cadherin expression by K-Ras; involvement of DNA methyltransferase-3b. *Carcinogenesis* **31**, 1194-1201 (2010).
104 Dhillon, I. S. Co-clustering documents and words using bipartite spectral graph partitioning. . *In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (San Francisco, California: ACM)*, pp. 269-274 (2001).
105 Costa, Y. *et al.* NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. *Nature* **495**, 370-374 (2013).
106 Kim, M. O. *et al.* ERK1 and ERK2 regulate embryonic stem cell self-renewal through phosphorylation of Klf4. *Nat. Struct. Mol. Biol.* **19**, 283-290 (2012).

Nature Genetics: doi:10.1038/ng.3449