Supplementary Materials for

Deficiency of microRNA *miR-34a* expands cell fate potential in pluripotent stem cells

Yong Jin Choi[1†], Chao-Po Lin[1*†], Davide Risso[2†], Sean Chen[1], Meng How Tan[3], Jin B Li[3], Yalei Wu[4], Caifu Chen[5], Zhenyu Xuan[6], Todd Macfarlan[7], Weiqun Peng[8], Sang Yong Kim[9], Terence P Speed[10] and Lin He[1]*

Correspondence to: newcplin@gmail.com (CPL) and lhe@berkeley.edu (LH)

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S6

Other Supplementary Materials for this manuscript includes the following:

Tables S1 to S6
Supplemental Information: the reannotate file (.GFF) of retrotransposons.

Materials and Methods

Mouse breeding and genotyping

The generation of *miR-34a*$^{-/-}$ mice was described previously (*16*). Both wild-type and *miR-34a*$^{-/-}$ mice were maintained on an isogenic C57BL/6J background, and housed in a non-barrier animal facility at UC-Berkeley. The following primers were used for genotyping, with parenthetical values indicating the size of the diagnostic PCR product: *mir-34a*-Common-R, ACTGCTGTACCCTGCTGCTT, with *mir-34a*-WT-F, GTACCCCGACATGCAAACTT (wild-type band, 400 bp), or *mir-34a*-KO-F, GCAGGACCACTGGATCATTT (knockout band, 263 bp) (*16*). NCr-*nu/nu* female athymic mice used for teratoma generation were purchased from Taconic (Taconic, Cat. # NCRNU). All the mouse work was done with approval of University of California, Berkeley's Animal Care and Use Committee. University of California, Berkeley's assurance number is A3084-01, and is on file at the National Institutes of Health Office of Laboratory Animal Welfare.

Derivation of embryonic stem cells (ESCs)

Mouse ESCs were isolated based on published protocols with slight modifications. Uteri containing E3.5 wild-type or *miR-34a*$^{-/-}$ embryos were isolated from timed pregnant females, and put in Knockout DMEM (Life Technologies, Cat. # 10829-018) supplemented with 10mM HEPES (Life Technologies, Cat. # 15630-080). E3.5 blastocysts were flushed with 1ml syringes with 18G needles and individually transferred to a 12-well plate with irradiated MEF (mouse embryonic fibroblasts) feeders in 1 ml N2B27 medium containing 100 U/ml LIF (EMD Millipore, Cat. # ESG1107), 1 µM PD0325901 (Sigma, Cat. # PZ0162) and 3 µM CHIR99021 (EMD Millipore, Cat. # 361559). After 5 days of incubation, embryo outgrowth was separated from the trophectoderm (TE) and picked up by a 10 µl pipette and transferred to 20 µl Accutase (Life Technologies, Cat. # A11105-01) and incubated at 37^{o}C for 20 min to dissociate cells. Dissociated cells were then cultured on irradiated MEF feeder cells with N2B27 medium containing LIF and two inhibitors for one passage. Subsequently, ESCs were passaged with 0.25% Trypsin-EDTA and maintained in regular mouse ES medium. ESCs were also derived in regular ES medium (see below) to test for variation among derivation protocols.

Generation of induced pluripotent stem cells (iPSCs)

Wild-type and *miR-34a*$^{-/-}$ iPSCs were generated from primary mouse embryonic fibroblasts (MEFs) by somatic reprogramming (*3*). Primary MEFs were isolated from littermate-controlled E13.5 wild-type and *miR-34a*$^{-/-}$ embryos, infected with pMX retroviral vectors that encode mouse Oct4, Sox2 and Klf4 (Addgene, Cat. # 13366, 13367 and 13370), and cultured on irradiated MEF feeder in ES medium containing Knockout DMEM (Life Technologies, Cat. # 10829-018), 15% ES-grade FBS (Omega scientific, Cat. # FB-01), 2mM L-glutamine (Life Technologies, Cat. # 25030-164), 1x10^{-4}M MEM non-essential amino acids (Life Technologies, Cat. # 11140-076), 1x10^{-4} M 2-mercaptoethanol (Sigma, Cat. # M3148) and 1X Penicillin-Streptomycin (Life Technologies, Cat. # 15140-122). Subsequently, single iPSC-like colonies were individually picked and expanded on irradiated MEF feeders to establish a stable line. At least three independent iPSC lines were generated for each genotype.

Embryoid body (EB) differentiation

For EB differentiation, ESCs or iPSCs were plated in 10cm petri dish (150,000 cells/ml) in ES cell medium without LIF and gently cultured on a rotator after removal of feeder cells by incubating on a gelatin-coated plate for 1 hour. Medium was changed daily. EBs were collected every 3 other days and analyzed by real-time qPCR and immunofluorescence (IF).

Generation of teratomas from pluripotent stem cells and histological analyses

$1x10^6$ of WT or *miR-34a*$^{-/-}$ iPSCs or ESCs were injected into the dorsal flanks of 6-7 week old immune-deficient NCr-*nu/nu* female mice (Taconic, Cat# NCRNU). After 4-5 weeks, resulting teratomas were collected by surgical removal, fixed overnight in 10% buffered formalin (Fisher Scientific, Cat. # SF100-4), dehydrated in a graded series of ethanol solutions, embedded in Paraplast X-TRA paraffin (Fisher Scientific, Cat. # 23-021-401), sectioned at 6 μm thickness, mounted on glass slides, and stained with hematoxylin and eosin (H&E) using standard procedures (*16*). Additionally, the paraffin sections were subjected to IF and immunohistochemistry.

Generation of chimeric blastocyst embryos and chimeric mice

To generate chimeric blastocysts by morula aggregation, we followed the method described by Nagy *et. al.* with minor modifications. One-cell stage, C57B6/J wild-type zygotes were collected at 0.5 day postcoitum (dpc), cultured in EmbryoMax KSOM Medium (Millipore, Cat. # MR-121-D) for 48h and only well-developed morula embryos were selected for aggregation. The ESCs or iPSCs were combined with morula embryos by sandwich method after removal of zona pellucida by acid Tyrode's solution (Sigma, Cat. # T1788) and then cultured overnight.

To generate chimeric blastocysts by microinjection, four ESCs or one ESC of the desired genotype were injected into E2.5 C57Bl/6N wild-type recipient morula embryos (16-32 cell stage) and then cultured *in vitro* overnight to obtain the chimeric blastocysts. To generate chimeric mice by microinjection, 10-15 ESCs of the desired genotype were injected into E3.5 recipient blastocyst embryos before implanted into pseudopregnant females. Chimeric embryos were collected at E9.5 and E12.5 and subjected to IF staining. In both experiments, the recipient embryos for microinjection were generated from 4 week old, super-ovulated C57Bl/6N female mice. These female mice were first treated with intraperitoneal injection with 5 IU of PMSG (Sigma, pregnant mare serum gonadotropin, Cat. # G4877-1000U) and 5 IU of HCG (Sigma, human chorionic gonadotropin, Cat. # CG5-1VL) and then mated with male mice from the same strain. Subsequently, one-cell embryos were collected from those carrying vaginal plugs 24 hour after intraperitoneal injection of HCG. Collected embryos were cultured for 2 or 3 days *in vitro* in EmbryoMax KSOM Medium (Millipore, Cat. # MR-121-D); and properly developed morulae or blastocysts were selected for microinjection.

Preimplantation embryo expression analysis

Superovulated wild-type and *miR-34a*$^{-/-}$ females were mated with males of matching genotype to generate 2C, 8C and blastocyst embryos at E1.5, E2.5, and E3.5, respectively. Oocytes were collected at E0.5 from unmated, superovulated females. 2C and 8C embryos were recovered by oviduct flushing with DMEM (Thermo Fisher, Cat. #11995-040), while blastocysts were recovered by uterine flushing. Embryos were washed in PBS and subject to real-time PCR analyses using a Single Cell-to-Ct qRT-PCR kit (Life Technologies, Cat. # 4458236). All primers used are listed in Table S6.

Immunofluorescence (IF) and Immunohistochemistry (IHC)

For IF staining of differentiated EBs or ESCs/iPSCs, samples were fixed with 4% paraformaldehyde for 10 min at room temperature and incubated with blocking solution (0.1%Triton X-100 and 5% normal goat serum in PBS) for 1 hour at room temperature. To detect the expression of pluripotent or lineage markers, EBs/cells were incubated overnight at 4°C with antibodies against MERVL-Gag (1:2000, a gift from T. Heidmann laboratory), Oct4 (1:100, Santa Cruz Biotechnology, Cat. # sc-5279), Gata4 (1:100, Santa Cruz Biotechnology, Cat. # sc-9053) or Cdx2 (1:100, Abcam, Cat. # ab76541 or #157524), followed by staining with goat anti-rabbit IgG (H+L) secondary antibody, Alexa Fluor 594-conjugated secondary antibody (1:500, Life Technologies, Cat. # A11037) for 1 hour at room temperature. EBs/cells were then stained with DAPI (300 nM, Sigma, Cat. # D9564) and subjected to imaging analyses using Laser scanning confocal microscopy (Zeiss LSM710) and Zeiss Observer.A1 microscope.

For IF staining of chimeric blastocysts, samples were fixed in 4% paraformaldehyde (PFA) for 20 min at room temperature and then transferred to phosphate-buffered saline (PBS) containing 0.1% bovine serum albumin (BSA). Embryos were permeabilized using 0.1% Triton X-100 in PBS containing 0.1% BSA for 5 min, and then blocked for 1 hour at room temperature in blocking solution (10% goat serum diluted in PBS/0.1% BSA). Blastocysts were then incubated with anti-Cdx2 primary antibody (1:100, Abcam, Cat. # ab76541) in blocking solution at 4°C overnight and stained with goat anti-rabbit, Alexa Fluor 594-conjugated secondary antibody (1:300, Life Technologies, Cat. # A11037) in blocking solution for 1 hour at room temperature. Blastocysts were then placed individually into chamber slides (Lab-Tek, Cat. # 155411) in 400 ml PBS/0.1%BSA solution. Images were taken using an Olympus Revolution XD spinning disk confocal microscope.

For IF staining of chimeric mouse embryos, including placentas and yolk sacs were fixed with 4% PFA for 2 hours, incubated in 30% sucrose for overnight in 4°C, embedded in Tissue-Tek O.C.T. compound (VWR, Cat. #25608-930), and cryo-sectioned at 8 mm. These sections were either directly visualized for GFP expression or subjected to IF using mouse anti-GFP (1:100, Abcam, Cat. # ab38689), rabbit anti-Tpbpa (1:200, Abcam, Cat. # ab104401), or rabbit anti-MTP1 (1:150, Alpha Diagnostic, Cat. # MTP11-A) primary antibodies. Trophoblast giant cells were identified based on their location in the placenta and the morphology of enlarged nuclei. Spongiotrophoblasts were identified based on the staining of the molecular marker, Tpbpa. The bilaminar structure of the yolk sac is identified by DAPI staining, and the visceral endoderm part is identified by its columnar, epithelial morphology. The GFP signals in three embryonic germ layers of chimeric mouse embryos (E12.5) and yolk sacs were directly visualized without staining.

For immunohistochemistry (IHC) analyses on teratomas or placentas of chimeric embryos , 5 mm paraffin sections were deparaffinized, dehydrated, and subjected to heat-induced antigen retrieval in a pressure cooker using Target Retrieval solution (DAKO, Cat. # S1699). Slides were incubated for 10 minutes with 3% H_2O_2, blocked for 3 hours with PBS containing 5% BSA and 0.3% Triton X-100, and incubated with primary antibodies against PL-1 (1:75, Santa Cruz

Biotechnology, Cat. # sc-34713) or GFP (1: 100, Abcam, Cat. # ab38689) overnight in PBS buffer containing 1% BSA and 0.3% Triton X-100. Slides were then incubated with horseradish peroxidase (HRP)-conjugated secondary antibodies for 2 hours at room temperature, and then subjected to 3,3'-Diaminobenzidine (DAB) staining (Life Technologies, Cat. # 00-2014) followed by a counterstain with Mayer's hematoxylin (Electron Microscopy Sciences, Cat. # 26503-04). The sinusoidal trophoblast giant cells (s-TGCs) were identified by their enlarged nuclei and adjacent location in the maternal blood sinusoid space.

Real-time PCR analyses

The expression levels of the indicated genes and retrotransposons were determined by real-time PCR with SYBR FAST qPCR Master Mix (Kapa Biosystems, Cat. # KK4604) in triplicate. In all real-time PCR experiments, *actin* mRNA was used as an endogenous normalization control. Primer design information for all real-time PCR analyses was included in Table S6.

RNA sequencing (RNA-seq)

Total RNA was extracted from 3 pairs of wild-type and *miR-34a*$^{-/-}$ iPSCs by Trizol (Life Technologies, Cat. # 15996-026) and followed by a clean-up procedure using on-column DNase I treatment with RNeasy Mini Kit (Qiagen, Cat. # 74104). The integrity of RNA samples was determined by Bioanalyzer 2100 (Agilent). RNA-seq libraries were made from samples with a RNA Integrity Number (RIN) of 9.0 or greater. Strand-specific, poly(A) selected RNA-seq libraries were prepared as previously described (*38*). In brief, polyA+ RNAs were enriched by the purification with two rounds of Dynal Oligo (dT) beads (Life Technologies, Cat # 610-06). The RNAs were then fragmented in the first-strand buffer (Life Technologies, Cat. #11752-050) at 85°C for 7–8 min, followed by the first strand cDNA synthesis by random hexamers (Life Technologies, Cat. # 48190-011) and SuperScript III (Life Technologies, Cat. # 18080-04). Second strand cDNA synthesis was performed with dUTP in place of dTTP to mark the second strand to allow strand specific RNA-seq library construction. Ends of the cDNAs were repaired with End-It DNA End Repair Kit (Epicentre, Cat. # ER0720), and adenosines were added to the 3′ ends using the Klenow fragment (New England Biolabs, Cat. # M0212S). After the cDNAs were ligated to the standard Illumina adapters, DNA fragments in the size range of 300–600 bp were gel extracted and treated with uracil-DNA glycosylase (New England Biolabs, Cat. # M0280L) to remove the second cDNA strand. Finally, each library was amplified using Phusion DNA polymerase (New England Biolabs, Cat. # F-531), quantified using the Qubit dsDNA High Sensitivity Assay Kit (Life Technologies, Cat # Q32854) and sequenced on HiSeq 2000 (Illumina) to produce 153 million paired-end 100-bp reads (average 25.5 million reads per sample). The same libraries were also sequenced on the NextSeq platform (Illumina) to generate 192 million paired-end 150bp reads (average 32 million reads per sample). The sequencing data are publicly available through NCBI GEO with the accession number GSE69484.

Chromatin immunoprecipitation (ChIP)

For each ChIP experiment, 10^6 ESCs or iPSCs were fixed with 1% formaldehyde (VWR, Cat. # 5016-02) to extract chromatin for immunoprecipitation. Nuclei were isolated by Farnham lysis buffer (5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP-40 with Protease Inhibitor Cocktail Tablet (Roche, Cat. # 11836153001) and lysed in nuclear lysis buffer (50 mM Tris pH 8.0, 10 mM EDTA, and 1% SDS with the protease inhibitor cocktail). Chromatin was fragmented by a

Covaris S220 Focused ultrasonicator (peak power 175, duty factor 20, cycles/burst 200, duration 30s, with 35 treatments) and diluted in RIPA buffer (10 mM Tris pH 7.6, 1 mM EDTA, 0.1% sodium deoxycholate, and 1% Triton X-100 with protease inhibitor cocktail) in a 1:9 ratio. The pull-down was performed at 4°C overnight using 40 ml Dynabeads protein A (Life Technologies, Cat. # 10001D) and 2 mg antibodies against Gata2 (Santa Cruz, Cat. # sc-9008), H3K4Me (Abcam, Cat. # ab8895), H3K4Me3 (EMD Millipore, Cat. # 17-614), H3K9Me2 (Abcam, Cat. # ab1220), and H3 (Abcam, Cat. #1791). Washes were performed twice with the low-salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 150 mM NaCl), three times with the high-salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 500 mM NaCl), and four times with the LiCl wash buffer (0.25 M LiCl, 1% IGEPAL CA630, 1% sodium deoxycholate 1 mM EDTA, 10 mM Tris pH 8.0). Chromatin-immunoprecipitated DNA was analyzed by qPCR using SYBR FAST qPCR Master mix (KAPA biosystems, Cat. # KK4602) and a 7900HT Fast Real-Time PCR machine (Applied Biosystems). Percent input of immunoprecipitated samples was calculated according to the real-time PCR results of serially-diluted lysates. The enrichment of each individual locus is calculated as (the percentage of input of modified histone H3)/(the percentage of input of histone H3). The following primers were used for qPCR analysis: *mervl* forward, 5'-CTTCCATTCACAGCTGCGAC-3'; *mervl* reverse 5'-CTAGAACCACTCCTGGTACC-3'; *iap* forward; 5'-GCTCCTGAAGATGTAAGCAATAAAG-3'; *iap* reverse, 5'-CTTCCTTGCGCCAGTCCCGAG-3'; *mmervk10c* forward, 5'-TTCGCCTCTGCAATCAAGCTCTC; *mmervk10c* reverse, 5'-TCGCTCRTGCCTGAAGATGTTTC-3'; *tcstv1* forward, 5'-ATCTACTTGGGGTGCCTGGT-3'; *tcstv1* reverse, 5'-GAAGACCAGCTGAACCATCC-3'.

Luciferase Assays

For MERVL-luciferase reporter assays, we used pGL3 luciferase reporter vectors (Promega, Cat. # E1751) harboring MERVL$_{1\text{-}1000}$, MERVL$_{1\text{-}493}$, and MERVL$_{500\text{-}1000}$ fragments, as described by Macfarlan *et. al.* (*25*). The MERVL$_{125\text{-}375}$-Luc reporter was constructed by truncating the MERVL$_{1\text{-}493}$ fragment using a QuikChange Site-Directed Mutagenesis Kit (Strategene, Cat. # 200518). The two fully conserved Gata2 binding sites (BS1 and BS3) were ablated in the MERVL$_{125\text{-}375}$-Luc reporter construct, either individually or in combination, using a QuikChange Site-Directed Mutagenesis Kit. The following primers were designed for mutagenesis:

1-375 Forward, TGGACTTCCATTCACCTCGAGATCTGCGATCTAAGTAAGC; 1-375 Reverse, ATCGCAGATCTCGAGGTGAATGGAAGTCCAAGGATCTAGC; 125-375 Forward, CTTACGCGTGCTAGCGATCTTGAGCCATAGTGGCTATGGA; 125-375 Reverse, CTATGGCTCAAGATCGCTAGCACGCGTAAGAGCTCGGTAC; Gata2ΔBS1 Forward, TCTCCGAGTTTAAGGAACACACCTTTGGGCTACGCCTTTC; Gata2ΔBS1 Reverse, AATCCCAGATGAAAGGCGTAGCCCAAAGGTGTGTTCCTTA; Gata2ΔBS3 Forward, TTAAAGGTGTGGTGGAACACACCTTTGGGCTACACCTTCT; and Gata2ΔBS3 Reverse, TGTCTCCAGCAGAAGGTGTAGCCCAAAGGTGTGTTCCACC. MERVL-Luc reporters and control Renilla luciferase reporter pRL-TK (Promega, Cat. # E2241) were co-transfected (600 ng and 150 ng per well of a 12-well plate, respectively) using Lipofectamine 2000 reagent (Life Technologies, Cat. # 11668027) into ESCs. Transfection complexes containing the reporter constructs were prepared in Opti-MEM Reduced-Serum Medium (Life Technologies, Cat. # 31985062) according to manufacturer's instructions. After trypsinization with 0.25% Trypsin + EDTA (Life Technologies, Cat. # 25200-056), 100,000 cells were resuspended in ES media lacking Pen Strep, incubated with transfection complexes for 10 minutes at 37° C, and then transferred to one well of a 12-well plate containing feeders. After 48 hours, transfected ESCs were trypsinized, plated onto gelatin-coated plates for 1 hour to remove feeders, and then assayed for luciferase activity by Dual-Luciferase® Reporter Assay System (Promega, Cat. # E1910) using a Glomax 20/20 Luminometer (Promega).

For *gata2* 3’UTR luciferase assays, a fragment that includes the 3’ portion of the ORF and the entire Gata2 3’UTR was amplified by PCR. The fragment was then cloned into a psiCheck-2 vector (Promega, Cat. # C8021) to generate the *gata2* 3’UTR-Luc reporter.*miR-34a* binding site mutants were generated using a QuikChange Site-Directed Mutagenesis Kit. The following primers were used: Gata2 3’UTR F XhoI, CTCGAGAGTCTCTCTTTTGGCCACCC; Gata2 3’UTR R NotI, GCGGCCGCCAAGGCCACCTGACAGCTTA; Gata2 3’UTRΔ34aBS F, CCGTCCAGCATGGTGATGGGCTAGGCAAGCCTCCCACTGG; Gata2 3’UTRΔ34aBS R, GCTTGCCTAGCCCATCACCATGCTGGACGGGTGGGGGTGG; Gata2 3’UTRΔ34aBS2 F AGAGACCCACTTCCTGCCTAGCCTGGCCGAAGCCACCTCT; Gata2 3’UTRΔ34aBS2 R, TCGGCCAGGCTAGGCAGGAAGTGGGTCTCTTGGGATGGGC; Gata2 3’UTRΔ34aBS3 F, CTTCTTTGGGACCTCCCAGTCAGGGCTCTCGGGGGCAGAC; Gata2 3’UTRΔ34aBS3 R, GAGAGCCCTGACTGGGAGGTCCCAAAGAAGGACCCCAAGA. The *gata2* 3’UTR-Luc reporters (2 ng per well of 12-well plate) were co-transfected with 100 nM siGFP or mature *miR-34a* RNA mimics using Lipofectamine 2000 reagent (Life Technologies, Cat. #11668027) into a feeder-free mouse ESC line (*39*). After 48 hours, cells were lysed and assayed for luciferase activity by Dual-Luciferase® Reporter Assay System (Promega, Cat. # E1910) using a Glomax 20/20 Luminometer (Promega).

Transfection and retrovirus/lentivirus transduction

To overexpress *miR-34a* in wild-type and *miR-34a*$^{-/-}$ iPSCs, cells were infected with MSCV (murine stem cell virus) retrovirus that encoded a LTR-*miR-34a* and a PGK-puromycin-IRES-GFP cassette (*40*). *MSCV* and *MSCV-miR-34a* transduced iPSCs were selected with 3 μg/ml puromycin for two days before collected for real-time PCR analyses and western blotting.

The ESCs and iPSCs used for microinjection were labeled by green fluorescence protein (GFP) using the *PiggyBac*-GFP plasmid. The *PiggyBac* vector contains an EF1a-driven GFP expression cassette and an ubiquitin-puromycin selection marker. The *PiggyBac*-GFP plasmid was mixed with the *PiggyBac* transposase plasmid in a 1:1 ratio (*41*), and subsequently transfected into ESCs or iPSCs using Lipofectamine 2000 (Life Technologies, Cat. # 12566014). Cells were selected with 3 μg/ml puromycin for two days and cultured in puromycin-free ES medium for following analyses.

To knock down *gata2* by RNAi, two Gata2 shRNAs were cloned into pLKO.1 lentiviral vector (Addgene, #10878) using the following oligos (shgata2#1 sense: 5’-CCGGGAGGTGGATGTCTTCTTCAACCACTCGAGTGGTTGAAGAAGACATCCACCTCTTTTTG-3’; sh*Gata2*#1 antisense: 5’-AATTCAAAAAGAGGTGGATGTCTTCTTCAACCACTCGAGTGGTTGAAGAAGACATCCACCTC-3’; sh*Gata2*#2 sense: 5’-CCGGGGACGAGGTGGATGTCTTCTTCAACTCGAGTTGAAGAAGACATCCACCTCGTCCTTTTTG-3’; sh*Gata2*#2 antisense: 5’-

AATTCAAAAAGGACGAGGTGGATGTCTTCTTCAACTCGAGTTGAAGAAGACATCCACCTCGTCC-3') (*42*); and the corresponding lentiviruses were produced by co-transfecting pLKO.1 shRNA vectors with pMD2.G and psPAX2 to HEK293T cells. After infection, iPSCs were selected in 3 μg/ml puromycin for two days and expanded for *in vitro* and *in vivo* analyses.

Western blotting

For ESC or iPSC collection, trypsinized cells were plated on a gelatin-coated plate for 1 hour to remove feeders. Cells separated from the feeders were then lysed in Laemmli sample buffer (60 mM Tris-Cl pH 6.8, 2% SDS, 100 mM DTT, 10% glycerol, 0.02% bromophenol blue) and subjected to western analyses. Antibodies against mouse Gata2 (Santa Cruz Biotechnology, Cat # CG2-96) was used at 1:500 dilution, and α-tubulin (Sigma, clone B-5-1-2) was used at a 1:4,000 dilution as a loading control. The quantitation of all western analyses was carried out with ImageJ (NIH).

Bisulfite sequencing

Genomic DNAs were purified by the standard phenol/chloroform method. 2 μg DNA was subjected to bisulfite conversion and subsequent purification using the EZ DNA Methylation-Gold Kit (Zymo Research, Cat. # D5005). Bisulfite-treated DNAs were amplified using Jumpstart REDTaq Readymix (Sigma, Cat. # P1107) with the following primers: MERVL forward, ATATGAATAAAGTGGTTATGGTGGT; MERVL reverse, AATTCCTAAACCCATAAATCCTAAC; IAP forward, TTGATAGTTGTGTTTTAAGTGG; IAP reverse, AAAACACCACAAACCAAAATC. The amplified DNA fragments were cloned to pGEM-T Easy vector (Promega, Cat. # A1360) for sequencing. The methylation patterns were analyzed by QUMA (Quantification tool for methylation analysis, http://quma.cdb.riken.jp).

Supplementary Text

RNA-seq data analysis

The paired-end RNA-seq reads were mapped to the GRCm38 (mm10) reference genome, downloaded from Ensembl (*43*) (v. 77), with TopHat (*36*) (v. 2.0.11). We quantified gene expression by counting the fragments overlapping with Ensembl genes, using FeatureCounts (*44*) (v. 1.4.5, options –p –B –C). Note that Ensembl annotation includes both non-coding and protein-coding genes. To obtain retrotransposon expression estimates, we counted only the fragments marked by TopHat as "primary alignments" that overlapped any annotated copy of a retrotransposon family (FeatureCounts options –p –B –C --primary). To avoid confounding between gene and retrotransposon expression, we excluded all the retrotransposon copies completely included in an annotated Ensembl exon. Although primary alignments are not meant to precisely indicate the "true" genomic origin of the repetitive fragment, TopHat guarantees only one primary alignment per fragment; this procedure let us count each fragment only once, and is similar in spirit to selecting one random mapping, if two or more mappings have the same number of mismatches. To obtain family-level expression estimates, we summed the expression of all the retrotransposon loci for each family. This procedure is preferable to quantifying the retrotransposon expression by mapping the reads to the RepBase consensus sequence for at least two reasons: (i) by mapping to the genome, we avoid spurious mapping to the consensus sequences (e.g., reads that would map to a gene with fewer mismatches, but end up mapped to a retrotransposon consensus sequence instead); (ii) it allows us to count towards the retrotransposon expression those reads for which one end map to the retrotransposon and the other end to a gene (i.e., *junction reads,* see below). Reassuringly, the derepression of MERVL in *miR-34a*$^{-/-}$ cells was confirmed when mapping the reads to the RepBase consensus sequence instead (data not shown).

MERVL-gene junctions were defined as those junctions, identified by TopHat, which overlap on one side with an annotated Ensembl gene (including protein coding genes, long ncRNAs, pseudogenes and antisense transcripts) and on the other side with an annotated element of MERVL (including both complete, truncated and solo LTR copies). Due to the repetitive nature of retrotransposon sequences, this procedure is not entirely accurate, especially in the presence of gene families and/or pseudogenes. For this reason, we adopted a stringent filtering procedure in which we retained only the junctions observed independently in at least three biological replicates. This procedure identified 87 high confidence MERVL-gene junctions, 42 of which are differentially expressed between *miR-34a*$^{-/-}$ and wild-type iPSCs (False Discovery Rate of 0.05; Table S5).

MERVL proximal genes were defined as those genes whose genomic coordinates are within a distance of 10 kb (either upstream or downstream) from an annotated copy of MERVL (either complete or solo LTR). In particular, a gene was annotated as "upstream" if it has an annotated MERVL element within 10 kb upstream of it; "downstream" if it has an annotated MERVL element within 10Kb downstream of it; "intronic" if a MERVL copy resides in its introns; and "overlap" if a MERVL copy is annotated as overlapping with its annotated exons. We further distinguished the genes in sense ("+") or antisense ("-") by comparing the gene's and the retrotransposon's strands.

Differential expression analysis was performed within the R/Bioconductor statistical framework (*45*). We combined the expression of the Ensembl genes and the retrotransposon families into a single matrix of expression estimates. We retained only the genes with at least 5 read pairs in at least 3 samples. We used edgeR to test for differential expression between *miR-34a*$^{-/-}$ and wild-type iPSCs (exact negative binomial test with upper-quartile normalization) (*37*). Genes with an adjusted p-value (Benjamini-Hochberg) of 0.05 or less and with an absolute $\log_2$ fold-change of 2 or more were defined as differentially expressed (DE). This procedure led to a total of 437 DE genes and 4 DE retrotransposon families (Table S2). We repeated the same procedure to identify individual DE retrotransposon elements, and we found 949 DE elements (600 of which belonged to MERVL) (Table S3). The expression quantitation of the retrotransposons at the individual locus level is a challenging problem, due to the repetitive nature of retrotransposons. By only counting "primary alignments" from TopHat, we are guaranteed to count each read only once towards the retrotransposon expression estimates (i.e., each read will contribute to the expression of only one locus). This is similar to randomly assigning each multi-mapping read to only one locus. An alternative approach is to allow all possible mappings and probabilistically assign each read, using tools like RSEM (*46*). Using RSEM on our data led to similar results (765 DE elements, 495 of which belong to MERVL). The advantage of considering genomic primary alignments is that the retrotransposon-junction reads are counted in the retrotransposon expression estimation (this is not possible with RSEM, since it aligns the reads to the transcript sequences rather than to the genome).

To ensure that we are not missing a large fraction of MERVL expressed loci due to mappability issues, we re-sequenced our RNA-seq libraries with the Illumina NextSeq 500 platform, generating additional 192 million 150 paired-end reads (average of 32 million reads per sample). We repeated the analyses described above on two additional datasets: (i) the NextSeq 150 bp paired-end reads, and (ii) a combined dataset obtained by pooling the reads from the two platforms. Reassuringly, these analyses largely recapitulated the results obtained on our previous dataset, confirming that short reads are sufficient to support our conclusions (Fig. S4). Using the longer reads, we identified 110 high confidence MERVL-gene junctions, 46 of which are differentially expressed between *miR-34a*$^{-/-}$ and wild-type iPSCs (False Discovery Rate of 0.05; Table S5); 520 DE genes and 4 DE retrotransposon families (Table S2 and S4); 1,349 DE elements (705 of which belonged to MERVL) (Table S2). With the combined dataset, we identified 134 high confidence MERVL-gene junctions, 57 of which are differentially expressed between *miR-34a*$^{-/-}$ and wild-type iPSCs (False Discovery Rate of 0.05; Table S4 and S5); 656 DE genes and 4 DE retrotransposon families (Table S2 and S4); 2,065 DE elements (772 of which belonged to MERVL) (Table S3).

To show that our *miR-34a*$^{-/-}$ samples are transcriptionally similar to reported totipotent-like ESCs, we conducted a meta-analysis of several publicly available RNA-seq datasets. Specifically, we re-analyzed 2C+ and 2C- ES cell samples (*7*) (GEO accession GSE33920); *kdm1a*$^{-/-}$ and *kdm1a*$^{+/+}$ ES cell samples (*25*); si-p60, si-p150 and si-control ES cell samples (*9*) (ArrayExpress accession E-MTAB-2684). The raw data were downloaded and analyzed following the same pipeline used for our samples. After mapping and gene quantitation, the read counts were normalized using upper-quartile normalization (*47*). To remove batch effects, we applied ComBat (*48*) as implemented in the sva Bioconductor package (*49*) to the log-transformed read counts, using the ID of the original submission as the “batch” variable. We then applied Principal Component Analysis (PCA) to the normalized and batch-adjusted data and selected the first two components for visualization (*50*). Hierarchical clustering was based on the 1,000 most variable genes (Fig. S3A).

To analyze ESC small RNA sequencing data, we trimmed the adapter sequence from each reads with in-house script, and mapped the trimmed reads to mouse reference genome (Build mm9) with program ELAND by using varied seed length (15-32 bp). The mapping results were merged based on the unique mapping locations of each read. We calculated the relative expression level of each microRNA annotated in MiRBase (version R14) by counting the number of reads overlapping the genomic location of the microRNA, and then normalized the count with the total count of uniquely mapped reads from each sample. Small RNA sequencing data from two biological replica were analyzed as described above.

Retrotransposons annotation

The annotation of retrotransposon coordinates for the mouse reference genome (mm10) was downloaded from Repeat Masker (v. rm403-db20130422). Repeat Masker annotated sequences often correspond to fragments of repetitive elements, leading to multiple disconnected fragments being annotated for each retrotransposon element at such loci. This issue particularly affects endogenous retrovirus (ERV) annotation. We used REAnnotate to merge fragmented Repeat Masker annotations that belong to a single retrotransposons element (*51*). This allows us to distinguish between ERVs with a complete or truncated gene structure, as well as ERVs that only contain a solo LTR (Supplemental Information).

Prediction of transcription factor binding sites

We used MATCH (version 8.6) program to search transcription factor binding motifs of TRANSFAC database in the 251-bp central portion of the MERVL LTR (MERVL$_{125\text{-}375}$). MinSUM profile was used as cutoff to balance both the false positives and false negatives in detecting binding sites. For highlighting the GATA binding sites in all MERVL LTR sequences, we used CLUSTAL-W (version 1.83) to perform multiple alignment with default setting.

Fig. S1

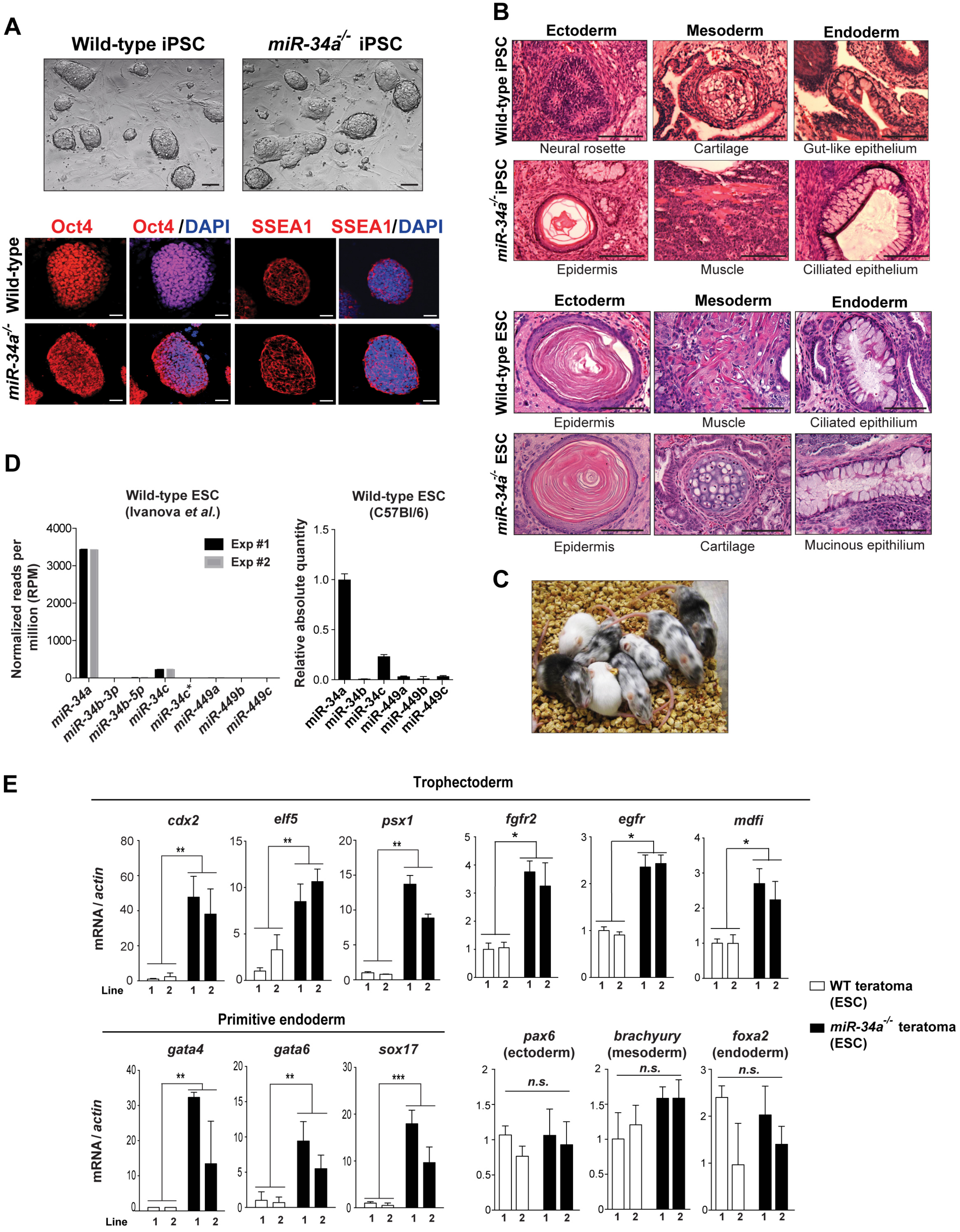
A
Wild-type iPSC
miR-34a-/- iPSC
Oct4
Oct4 /DAPI
SSEA1
SSEA1/DAPI
Wild-type
miR-34a-/-
B
Ectoderm
Mesoderm
Endoderm
Wild-type iPSC
Neural rosette
Cartilage
Gut-like epithelium
miR-34a-/- iPSC
Epidermis
Muscle
Cilliated epithelium
Ectoderm
Mesoderm
Endoderm
Wild-type ESC
Epidermis
Muscle
Ciliated epithilium
miR-34a-/- ESC
Epidermis
Cartilage
Mucinous epithilium
C
D
Wild-type ESC (Ivanova et al.)
Normalized reads per million (RPM)
0
1000
2000
3000
4000
Exp #1
Exp #2
miR-34a
miR-34b-3p
miR-34b-5p
miR-34c
miR-34c*
miR-449a
miR-449b
miR-449c
Wild-type ESC (C57Bl/6)
Relative absolute quantity
0.0
0.5
1.0
1.5
miR-34a
miR-34b
miR-34c
miR-449a
miR-449b
miR-449c
E
Trophectoderm
cdx2
elf5
psx1
fgfr2
egfr
mdfi
mRNA / actin
**
**
**
*
*
*
Line
1
2
1
2
Primitive endoderm
gata4
gata6
sox17
**
**
***
pax6 (ectoderm)
brachyury (mesoderm)
foxa2 (endoderm)
n.s.
n.s.
n.s.
WT teratoma (ESC)
miR-34a-/- teratoma (ESC)

Fig. S1 (Cont'd)

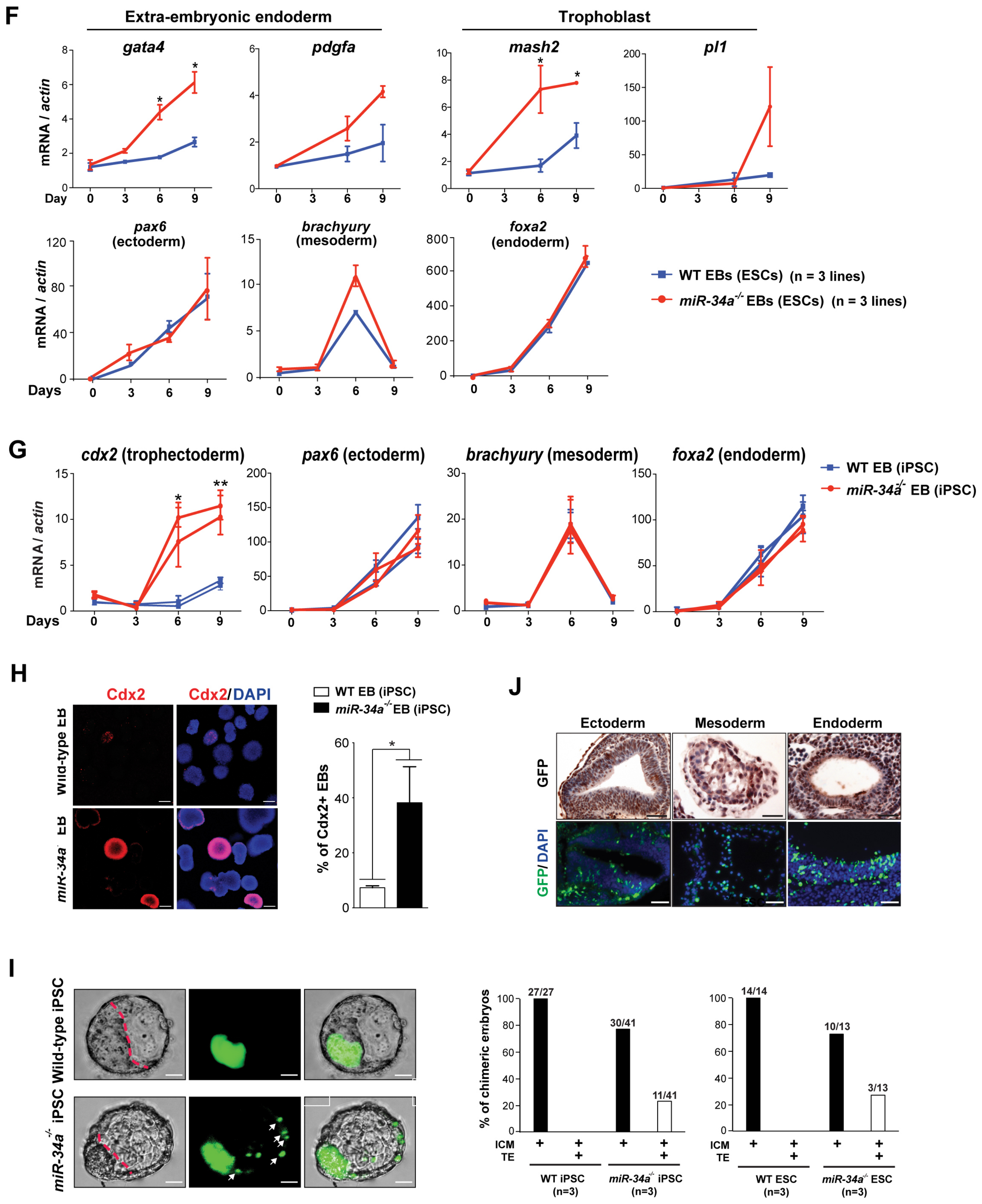
F
Extra-embryonic endoderm
Trophoblast
gata4
pdgfa
mash2
pl1
mRNA / actin
Day
0
3
6
9
*
pax6
(ectoderm)
brachyury
(mesoderm)
foxa2
(endoderm)
Days
WT EBs (ESCs) (n = 3 lines)
miR-34a-/- EBs (ESCs) (n = 3 lines)
G
cdx2 (trophectoderm)
pax6 (ectoderm)
brachyury (mesoderm)
foxa2 (endoderm)
mRNA / actin
Days
**
WT EB (iPSC)
miR-34a-/- EB (iPSC)
H
Cdx2
Cdx2/DAPI
Wild-type EB
miR-34a-/- EB
WT EB (iPSC)
miR-34a-/- EB (iPSC)
% of Cdx2+ EBs
J
Ectoderm
Mesoderm
Endoderm
GFP
GFP/DAPI
I
Wild-type iPSC
miR-34a-/- iPSC
% of chimeric embryos
27/27
30/41
11/41
14/14
10/13
3/13
ICM
TE
+
WT iPSC
(n=3)
miR-34a-/- iPSC
(n=3)
WT ESC
(n=3)
miR-34a-/- ESC
(n=3)

Fig. S1. *miR-34a$^{-/-}$* pluripotent stem cells exhibit an expanded cell fate potential *in vitro* and *in vivo*.

A. Wild-type and *miR-34a$^{-/-}$* iPSCs exhibit ESC-like morphology and express pluripotency markers Oct4 and SSEA1. Scale bars, 50 μm for differential interference contrast (DIC) images and 20 μm for immunofluorescence (IF) images. **B.** Wild-type and *miR-34a$^{-/-}$* iPSCs (top) and ESCs (bottom) generate differentiated teratomas containing tissues derived from three germ layers (ectoderm, mesoderm and endoderm) as shown by hematoxylin and eosin (H&E) staining. Scale bars, 25 μm. Two independent pairs of passage- and littermate-controlled wild-type and *miR-34a$^{-/-}$* iPSCs and ESCs are compared. **C.** *miR-34a$^{-/-}$* iPSCs efficiently contribute to adult chimeric mice as determined by coat color pigmentation. Shown here is a representative image of three independent experiments, in which three independent lines of *Oct4-Gfp/+; a/a; miR-34a$^{-/-}$* iPSCs were microinjected into albino-C57BL/6/cBrd/cBrd/cr blastocysts. **D.** The expression level of *miR-34/449* family miRNAs in wild-type ESCs measured by small RNA sequencing (left) and real-time PCR analyses (right). *miR-34a* is the most highly expressed *miR-34/449* miRNA in two independent wild-type ESC lines measured. **E.** The trophectoderm (TE) marker *cdx2*, *elf5*, *psx1*, *fgfr2*, *egfr* and *mdfi*, as well as the primitive endoderm marker *gata4*, *gata6*, and *sox17*, were highly induced in *miR-34a$^{-/-}$* teratomas, as determined by real-time PCR. In contrast, wild-type and *miR-34a$^{-/-}$* teratomas similarly induced the expression of *pax6* (an ectoderm marker), *brachyury* (a mesoderm marker) and *foxa2* (an endoderm marker). Teratomas were generated from two independent pairs of passage- and littermate- controlled wild-type and *miR-34a$^{-/-}$* ESC lines. Error bars: standard deviation (*s.d.*), n=3. **F.** Compared to wild-type EBs, *miR-34a$^{-/-}$* EBs derived from ESCs yield a greater expression of extra-embryonic endoderm marker *gata4* and *pdgfra,* and trophoblast lineage marker *mash2 (ascl2)* and *pl1 (prl3d1).* The real-time PCR analyses were performed using data collected from three independent pairs of passage- and littermate-controlled wild-type *and miR-34a$^{-/-}$* ESC lines. Error bars: *s.e.m.*, n=3. **G.** During EB differentiation, *miR-34a$^{-/-}$* iPSCs, but not wild-type iPSCs, exhibited strong induction of the TE marker *cdx2*. In contrast, wild-type and *miR-34a$^{-/-}$* EBs similarly induced the expression of *pax6* (an ectoderm marker), *brachyury* (a mesoderm marker) and *foxa2* (an endoderm marker). Two independent pairs of passage- and littermate-controlled wild-type *and miR-34a$^{-/-}$* iPSC lines are compared. Error bars: *s.d.*, n=3. **H**. Compared to wild-type EBs, *miR-34a$^{-/-}$* EBs derived from iPSCs yield a greater percentage of Cdx2-positive EBs, as well as an increased percentage of Cdx2+ cells in each EB. Scale bars, 100 μm. Shown here are representative images and quantitation from one pair of passage- and littermate-controlled wild-type and *miR-34a$^{-/-}$* iPSC and ESC lines. Error bars, *s.d.*, n=6 random fields measured. **I.** *miR-34a$^{-/-}$* iPSCs contribute to both ICM and TE when aggregated with recipient wild-type morulae. The ICM versus the TE contributions from the GFP-labeled wild-type or *miR-34a$^{-/-}$* iPSCs were determined by the localization of GFP positive cells in the chimeric blastocysts. While wild-type iPSCs exclusively contribute to the ICM, a fraction of *miR-34a$^{-/-}$* iPSCs contributes to both ICM and TE (white arrows) (top). Scale bar, 20 μm. The percentage of chimeric embryos with iPSC (left) or ESC (right) contribution to the ICM, the TE and ICM+TE in the chimeric blastocysts was quantified for both wild-type and *miR-34a$^{-/-}$* iPSCs from six independent aggregation experiments (bottom). Three independent pairs of passage- and littermate-controlled wild-type and *miR-34a$^{-/-}$* iPSC and ESC lines were compared. **J.** *miR-34a$^{-/-}$*ESCs contributed to all three lineage germ layers. Scale bars, 50 μm for IHC images and 100 μm for IF images. All *P*-values were calculated on a basis of a two-tailed Student's *t*-test. * $P < 0.05$, ** $P < 0.01$.

Fig. S2

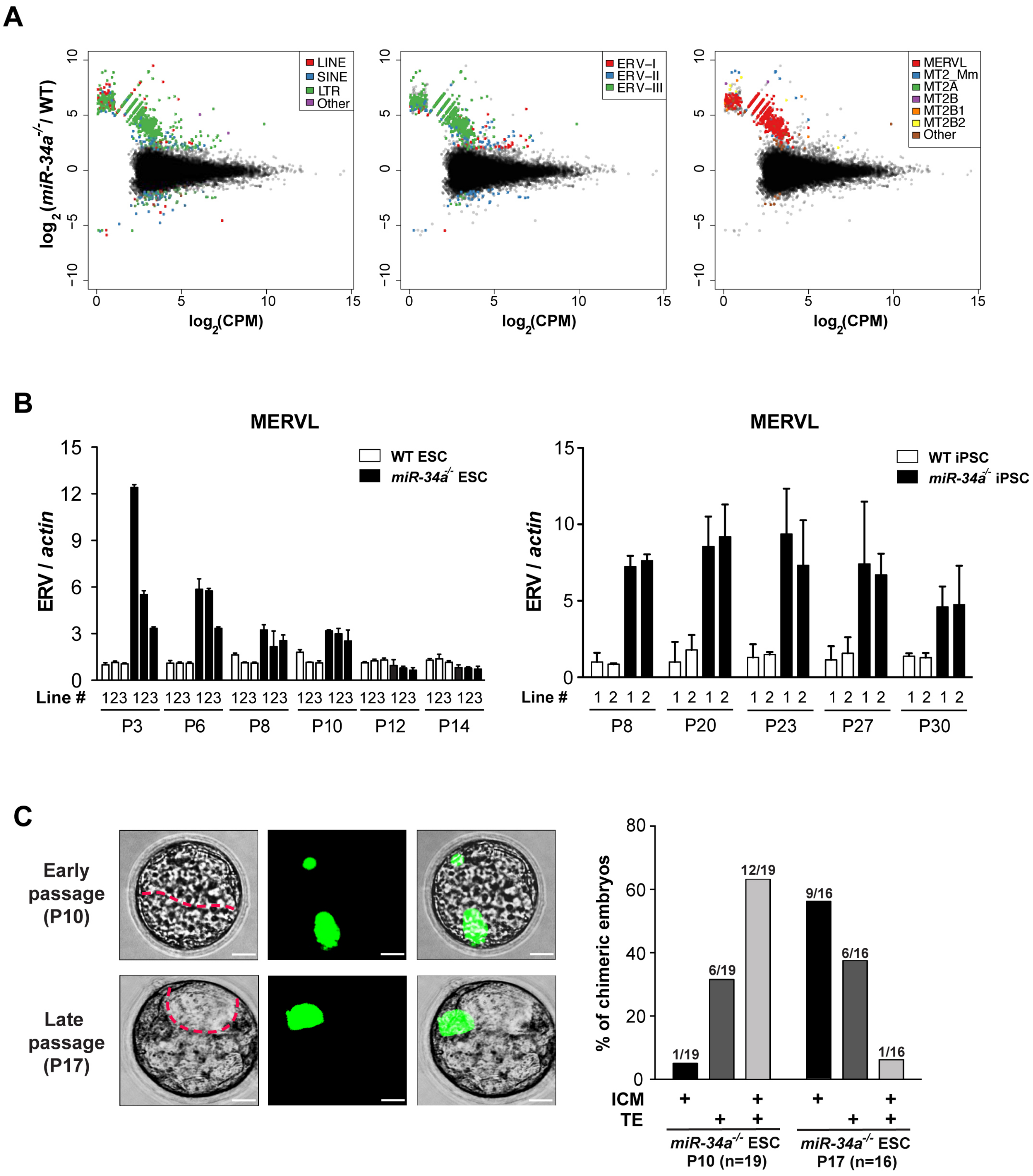
A
LINE
SINE
LTR
Other
ERV−I
ERV−II
ERV−III
MERVL
MT2_Mm
MT2A
MT2B
MT2B1
MT2B2
Other
log$_2$(*miR-34a*$^{-/-}$/ WT)
log$_2$(CPM)
−10
−5
0
5
10
15
B
MERVL
WT ESC
miR-34a$^{-/-}$ ESC
ERV / *actin*
0
3
6
9
12
15
Line #
123 123
P3
P6
P8
P10
P12
P14
MERVL
WT iPSC
miR-34a$^{-/-}$ iPSC
1 2 1 2
P8
P20
P23
P27
P30
C
Early passage (P10)
Late passage (P17)
% of chimeric embryos
0
20
40
60
80
1/19
6/19
12/19
9/16
6/16
1/16
ICM
TE
+
miR-34a$^{-/-}$ ESC P10 (n=19)
miR-34a$^{-/-}$ ESC P17 (n=16)

Fig. S2 (Cont'd)

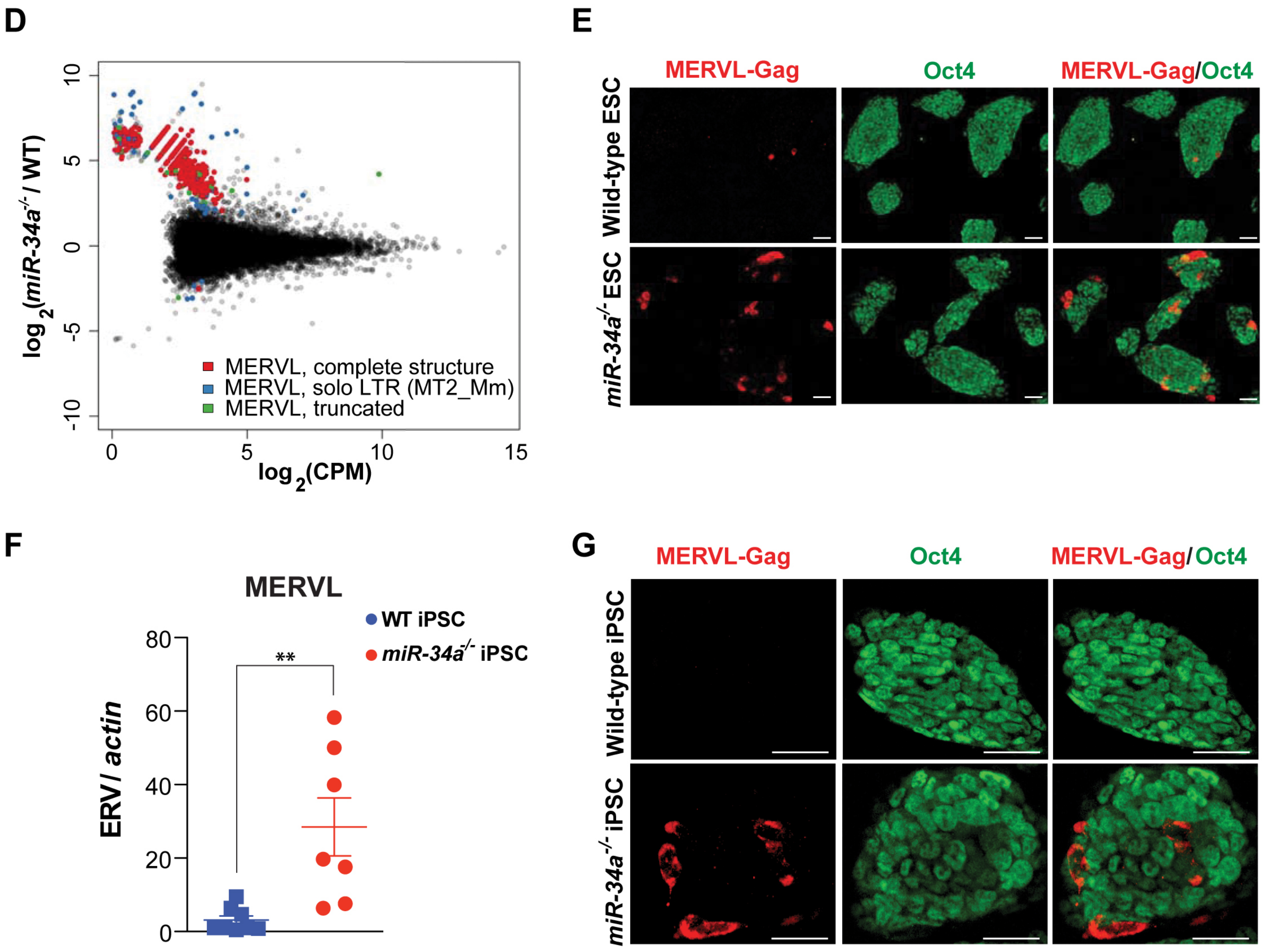
D
log2(miR-34a-/- / WT)
10
5
0
-5
-10
0
5
10
15
log2(CPM)
🟥 MERVL, complete structure
🟦 MERVL, solo LTR (MT2_Mm)
🟩 MERVL, truncated
E
MERVL-Gag
Oct4
MERVL-Gag/Oct4
Wild-type ESC
miR-34a-/- ESC
F
MERVL
● WT iPSC
● miR-34a-/- iPSC
ERV / actin
80
60
40
20
0
**
G
MERVL-Gag
Oct4
MERVL-Gag/Oct4
Wild-type iPSC
miR-34a-/- iPSC

Fig. S2. MERVL ERVs are specifically induced in *mir-34a*$^{-/-}$ pluripotent stem cells.

A. MA plots illustrate the comparison of the transcription profiles of all retrotransposon loci between wild-type and *miR-34a*$^{-/-}$ iPSCs using the RNA-seq data. We found 949 differentially expressed (DE) retrotransposon loci (FDR of 5%, absolute $\log_2$ fold-change ≥ 2). (Left) The DE loci are color-coded by retrotransposon class, with 69 LINEs (red, 7%), 101 SINEs (blue, 11%), 777 ERVs (green, 82%) and 2 others (purple, <1%). (Middle) Loci from the ERV-III class are preferentially derepressed in *miR-34a*$^{-/-}$ iPSCs. When only the ERV loci are highlighted for DE retrotransposons in the MA plot, 46 belong to the ERV-I class (red, 6%), 97 belong to the ERV-II class (blue, 12%) and 634 belong to the ERV-III glass (green, 82%). (Right) Loci from the MERVL family of ERVs are preferentially derepressed in *miR-34a*$^{-/-}$ iPSCs. When only ERV-III loci are highlighted for DE retrotransposons in the MA plot, we identified 552 MERVL (red, 87%), 27 MT2_Mm (blue, 4%), 1 MT2A (green, <1%), 3 MT2B (purple, <1%), 6 MT2B1 (orange, 1%), 11 MT2B2 (yellow, 2%) and 34 other ERV-III loci (brown, 5%). MT2_Mm is designated as the solo LTR of the canonical MERVL ERV; MT2A, MT2B, MT2B1, and MT2B2 refer to solo-LTRs that are related to MT2_Mm. CPM, counts per million. **B, C.** The extent of MERVL derepression (**B**) and the expanded cell fate potential (**C**) is decreased in late passages of *miR-34a*$^{-/-}$ ESCs. In contrast, MERVL derepression is more stable in *miR-34a*$^{-/-}$ iPSCs. The level of MERVL expression was measured using real-time PCR analyses for three (ESC) or two (iPSC) independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ cells. Error bars: *s.d.*, n=3. **C**. When four GFP-labeled ESCs were microinjected into each recipient morula, 63% of early passage (P10) *miR-34a*$^{-/-}$ ESCs contribute to both ICM and TE (n=12/19), while only 6% of late passage (P17) *miR-34a*$^{-/-}$ ESCs exhibit the same phenotype (n=1/16). Scale bar, 20 μm. **D.** An MA-plot compares the transcription profiles of all retrotransposon loci between wild-type and *miR-34a*$^{-/-}$ iPSCs using the RNA-seq data. Among the 949 DE retrotransposon loci (FDR of 5%, absolute $\log_2$ fold-change of 2 or more); 600 belong to the MERVL family (highlighted in color, 63%), revealing a specific derepression of this ERV family in *miR-34a*$^{-/-}$ iPSCs. Interestingly, 86.5% (519) are MERVL elements with a complete structure (red); 7.5% (45) are solo LTRs (blue); and 6% (36) are truncated copies (green). CPM, counts per million. **E**. *miR-34a*$^{-/-}$ ESC cultures exhibit an increase of cells expressing MERVL-Gag but lacking Oct4, as determined by IF staining. Scale bars, 20 μm. Images shown are representative from two pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ iPSCs. **F**. The MERVL expression level is heterogeneous among *miR-34a*$^{-/-}$ iPSC colonies. Using single-cell real-time PCR technology, we measured the MERVL expression level in individual colonies of wild-type and *miR-34a*$^{-/-}$ iPSCs, demonstrating the existence of colonial heterogeneity. Error bars: *s.e.m.*, n=7-8. All *P*-values were calculated on a basis of a two-tailed Student's *t*-test. ** $P < 0.01$. **G**. A representative *miR-34a*$^{-/-}$ iPSC colony contains cells with strong MERVL-Gag expression, yet no Oct4 expression. Scale bars, 50 μm.

Fig. S3

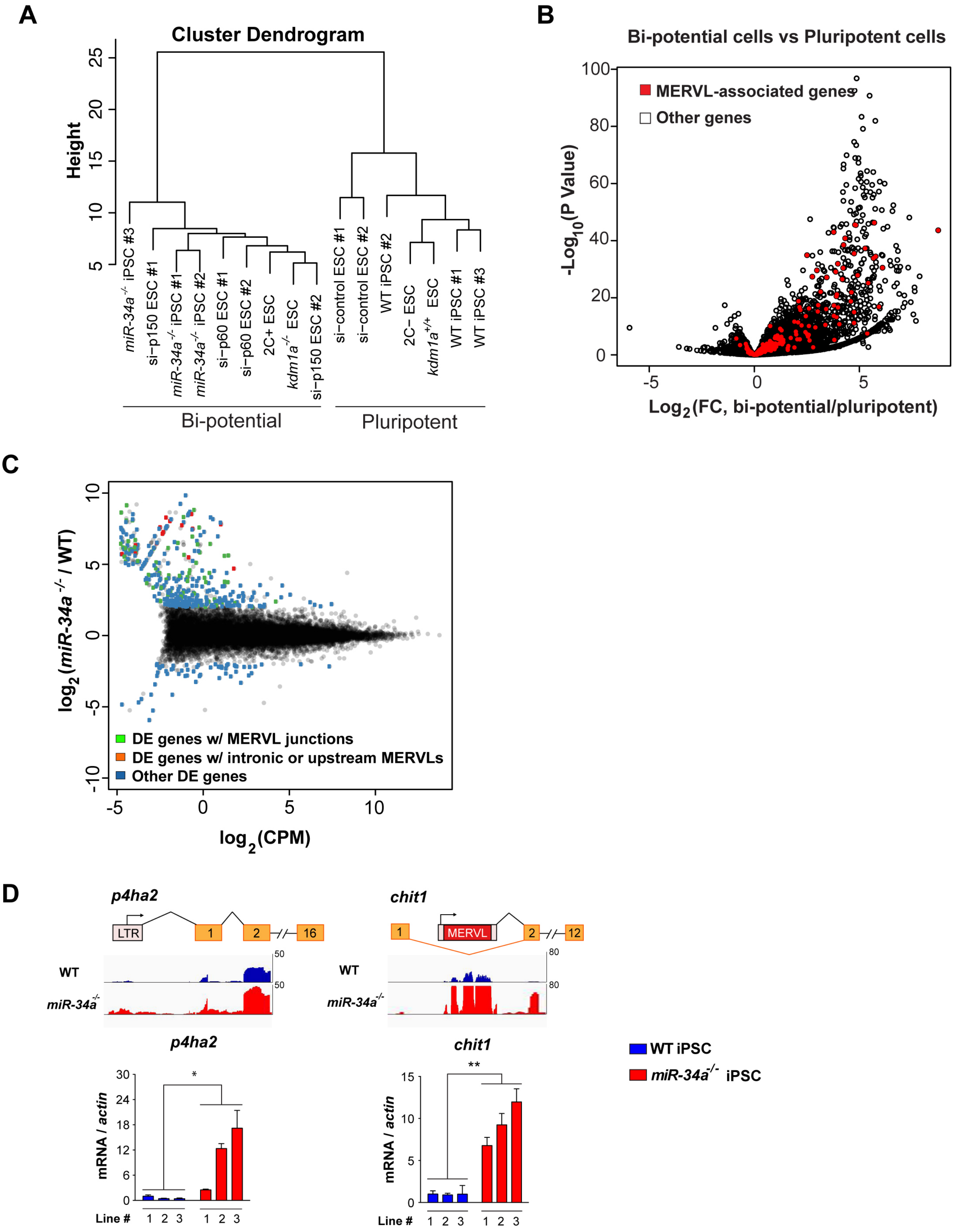
A
Cluster Dendrogram
Height
25
20
15
10
5
miR-34a-/- iPSC #3
si-p150 ESC #1
miR-34a-/- iPSC #1
miR-34a-/- iPSC #2
si-p60 ESC #1
si-p60 ESC #2
2C+ ESC
kdm1a-/- ESC
si-p150 ESC #2
si-control ESC #1
si-control ESC #2
WT iPSC #2
2C- ESC
kdm1a+/+ ESC
WT iPSC #1
WT iPSC #3
Bi-potential
Pluripotent
B
Bi-potential cells vs Pluripotent cells
🟥 MERVL-associated genes
□ Other genes
-Log10(P Value)
100
80
60
40
20
0
-5
0
5
Log2(FC, bi-potential/pluripotent)
C
log2(miR-34a-/- / WT)
10
5
0
-5
-10
🟩 DE genes w/ MERVL junctions
🟧 DE genes w/ intronic or upstream MERVLs
🟦 Other DE genes
-5
0
5
10
log2(CPM)
D
p4ha2
LTR
1
2
16
WT
miR-34a-/-
50
50
chit1
1
MERVL
2
12
WT
miR-34a-/-
80
80
p4ha2
*
mRNA / actin
30
24
18
12
6
0
Line #
1 2 3
1 2 3
chit1
**
mRNA / actin
15
10
5
0
Line #
1 2 3
1 2 3
🟦 WT iPSC
🟥 miR-34a-/- iPSC

Fig. S3 (Cont'd)

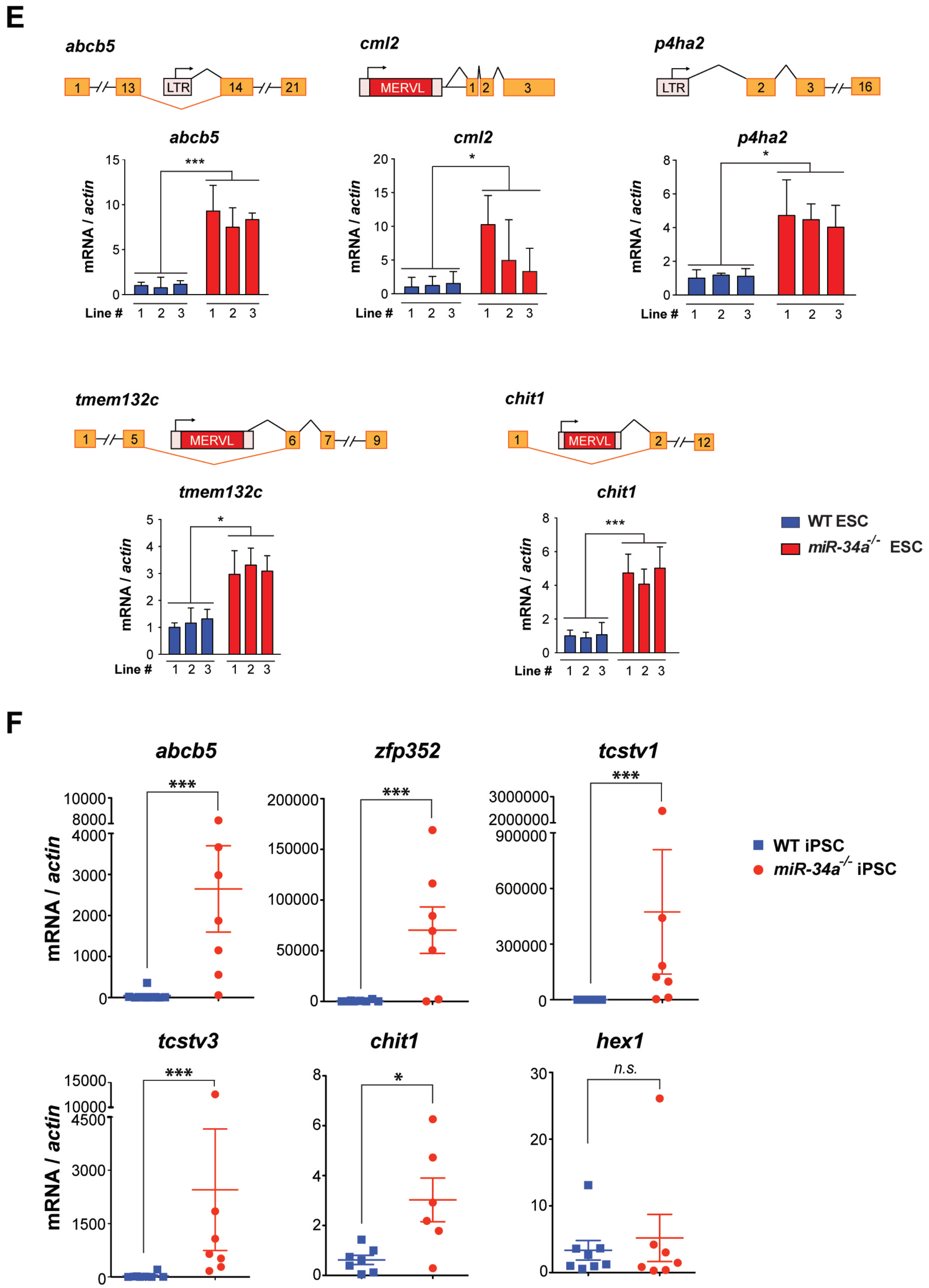
E
abcb5
1 13 LTR 14 21
cml2
MERVL 1 2 3
p4ha2
LTR 2 3 16
abcb5
mRNA / actin
0 5 10 15
***
Line # 1 2 3 1 2 3
cml2
mRNA / actin
0 5 10 15 20
*
Line # 1 2 3 1 2 3
p4ha2
mRNA / actin
0 2 4 6 8
*
Line # 1 2 3 1 2 3
tmem132c
1 5 MERVL 6 7 9
chit1
1 MERVL 2 12
tmem132c
mRNA / actin
0 1 2 3 4 5
*
Line # 1 2 3 1 2 3
chit1
mRNA / actin
0 2 4 6 8
***
Line # 1 2 3 1 2 3
WT ESC
miR-34a-/- ESC
F
abcb5
mRNA / actin
0 1000 2000 3000 4000 8000 10000
***
zfp352
0 50000 100000 150000 200000
***
tcstv1
0 300000 600000 900000 2000000 3000000
***
WT iPSC
miR-34a-/- iPSC
tcstv3
mRNA / actin
0 1500 3000 4500 10000 15000
***
chit1
0 2 4 6 8
*
hex1
0 10 20 30
n.s.

Fig. S3. The MERVL induction in *miR-34a*$^{-/-}$ pluripotent stem cells alters the expression and structure of a subset of MERVL proximal genes.

A-B. *miR-34a*$^{-/-}$ pluripotent stem cells and other reported totipotent-like ESCs share a similar expression profile. We reanalyzed several publicly available datasets (see Supplementary Text), and revealed a similar transcriptional profile between *miR-34a*$^{-/-}$ iPSC and other published totipotent-like ESCs, such as 2C+ ESCs (*7*), *kdm1a*$^{-/-}$ ESCs (*25*), and chromatin assembly factor-1 (CAF-1)-deficient (si-p60 and si-150) ESCs (*9*). **A.** Hierarchical clustering of published datasets (*7*, *9*, *25*) effectively clusters the expression data in two main branches: bi-potential vs. pluripotent. **B**. Differential expression analysis of bi-potential cells versus pluripotent stem cells using published datasets (*7*, *9*, *25*). We excluded our *miR-34a*$^{-/-}$ and wild-type iPSC data from this analysis to avoid the whole dataset driven by the *miR-34a*$^{-/-}$ iPSC data. The volcano plot shows that a large fraction of MERVL-associated genes (58/224) is differential upregulated in totipotent-like cells. FC: fold-change (bi-potential/pluripotent); CPM: counts per million. **C**. The differentially upregulated protein-coding genes in *miR-34a*$^{-/-}$ iPSCs are enriched for genes with an observed MERVL-gene junction or genes proximal to MERVL (n=50/242, Fisher's exact test, $P < 10^{-15}$). An MA-plot is shown to compare the transcription profiles of protein-coding genes between *miR-34a*$^{-/-}$ and wild-type iPSCs using the RNA-seq data. The DE protein-coding genes are color-coded by their relation to MERVL. Green: DE genes with an observed junction read with an adjacent MERVL element (see Supplementary Text); orange: DE genes with an upstream or intronic MERVL element; blue: all other DE genes. CPM: counts per million. **D, E.** Examples are shown for induced expression and altered transcript structure of MERVL proximal genes in *miR-34a*$^{-/-}$ iPSCs and ESCs. MERVL associated genes are induced in *miR-34a*$^{-/-}$ iPSCs (**D**) and *miR-34a*$^{-/-}$ ESCs (**E**). **D.** In *miR-34a*$^{-/-}$ iPSCs*,* the *p4ha2* transcription is induced by an upstream MT2B element, a solo LTR related to MERVL LTR; the *chit1* gene contains an intronic MERVL element in intron 1, which acts as an alternative promoter to drive a *chit1* isoform containing all downstream exons. The expression level of *chit1* and *p4ha2* was measured using three independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ iPSC lines. Error bars: *s.d.*, n=3. **E.** Real-time PCR analyses validate the induction of MERVL-gene isoforms in *miR-34a*$^{-/-}$ ESCs, including *cml2* and *p4ha2* (with an upstream MERVL element) as well as *abcb5*, *tmem132c* and *chit1* (with an intronic MERVL element). Intronicly localized MERVLs, either a solo LTR (as that in intron 13 of *abcb5*) or a complete ERV (as that in intron 5 of *tmem132c* or intron 1 of *chit1*), act as alternative promoters to drive the expression of truncated gene isoforms that only contain the downstream exons. Three independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ ESC lines were compared. Error bars: *s.d.*, n=3. **F.** Induction of MERVL proximal genes is heterogeneous among different *miR-34a*$^{-/-}$ iPSC colonies. Using single cell real-time PCR analyses, we measured the expression of MERVL proximal genes (*abcb5*, *zfp352*, *tcstv1*, *tcstv3*, and *chit1*) in individual colonies of wild-type and *miR-34a*$^{-/-}$ iPSCs, demonstrating the colonial heterogeneity in their expression level. Interestingly, one of a previously reported totipotency marker *hex1* (*8*) is expressed at a similar level between wild-type and *miR-34a*$^{-/-}$ ESCs, suggesting that the Hex1-positive ESCs are likely a different cell population from the totipotent-like *miR-34a*$^{-/-}$ ESC population. Error bars: *s.d.*, n=3. All *P*-values were calculated on a basis of a two-tailed Student's *t*-test. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, *n.s.*, not significant.

Fig. S4

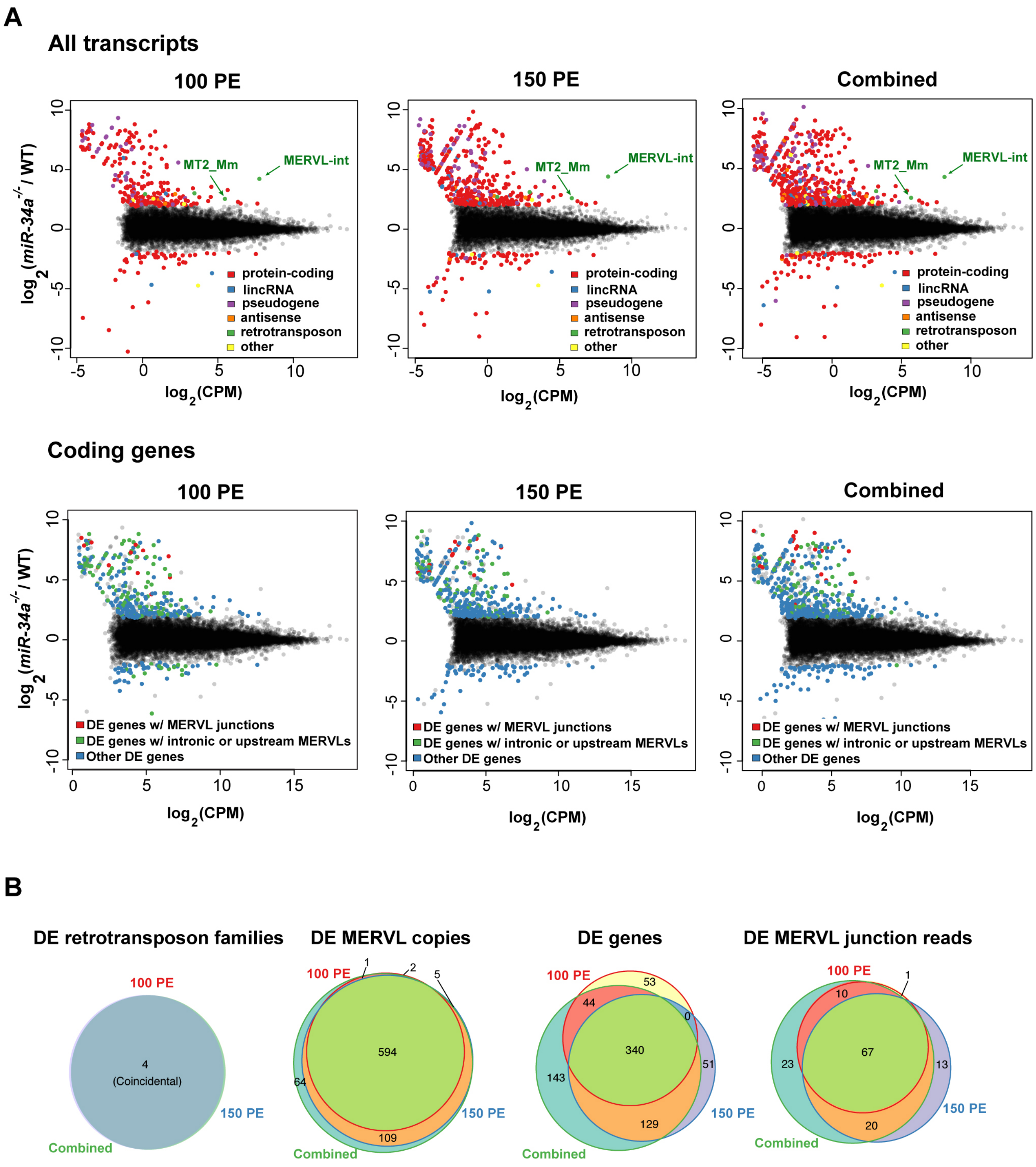
A
All transcripts
100 PE
150 PE
Combined
MT2_Mm
MERVL-int
protein-coding
lincRNA
pseudogene
antisense
retrotransposon
other
$\log_2(miR\text{-}34a^{-/-}$ / WT)
-10
-5
0
5
10
log$_2$(CPM)
Coding genes
100 PE
150 PE
Combined
DE genes w/ MERVL junctions
DE genes w/ intronic or upstream MERVLs
Other DE genes
15
B
DE retrotransposon families
100 PE
4
(Coincidental)
150 PE
Combined
DE MERVL copies
100 PE
1
2
5
594
64
109
150 PE
Combined
DE genes
100 PE
53
44
0
340
51
143
129
150 PE
Combined
DE MERVL junction reads
100 PE
1
10
67
23
13
20
150 PE
Combined

Fig. S4. The effect of the length and depth of RNA-seq data on the transcriptional profile characterization of *miR-34a$^{-/-}$* iPSCs.

We re-sequenced the wild-type and *miR-34a$^{-/-}$* iPSC RNA-seq libraries to a greater length (150 bp paired-end (PE)) and sequencing depth (additional 192 million reads), in order to explore to what extent our previous RNA-seq analysis is limited by the technical difficulty to precisely map the repetitive RNA-seq reads. **A**. (Top) The MERVL ERVs are the most highly induced and differentially expressed (DE) transcriptional unit in *miR-34a$^{-/-}$* iPSCs. MA-plots compare the transcription profiles of *miR-34a$^{-/-}$* and wild-type (WT) iPSCs. DE transcriptional units, including protein-coding genes, long non-coding RNAs (lncRNAs), pseudogenes, antisense transcripts, and retrotransposons, are color-coded by class, as in Fig. 2A. The results of 100 bp PE RNA-seq (left, same as Fig. 2A), the results of 150 bp PE RNA-seq (middle), and the results of the combined dataset by pooling together the two sequencing datasets (right) are shown as MA plots. (Bottom) The upregulated protein-coding genes in *miR-34a$^{-/-}$* iPSCs are enriched for genes with an observed MERVL-gene junction or genes proximal to MERVL. MA-plots compare the transcription profiles of protein-coding genes between *miR-34a$^{-/-}$* and WT iPSC using RNA-seq data. The DE protein-coding genes are color-coded by their relation to MERVL, as in Fig. S3C. The results of 100 bp PE RNA-seq (left, same as Fig. S3C), the results of 150 bp PE RNA-seq (middle), and the results of the combined datasets (right) are shown as MA plots. **B**. Venn diagrams showing the concordance between the 100 bp PE and 150 bp PE RNA-seq datasets. For each of the datasets (100 bp PE, 150 bp PE, and combined), we identified DE retrotransposons families, MERVL loci, DE genes, and DE MERVL-gene junctions (see Supplementary Text). The datasets are largely in agreement, suggesting that 100 bp PE reads are sufficient to detect most DE MERVL-gene junctions and DE MERVL loci. As expected, the combined dataset has more power to detect differential expression, since we are effectively doubling the sequencing depth.

Fig. S5

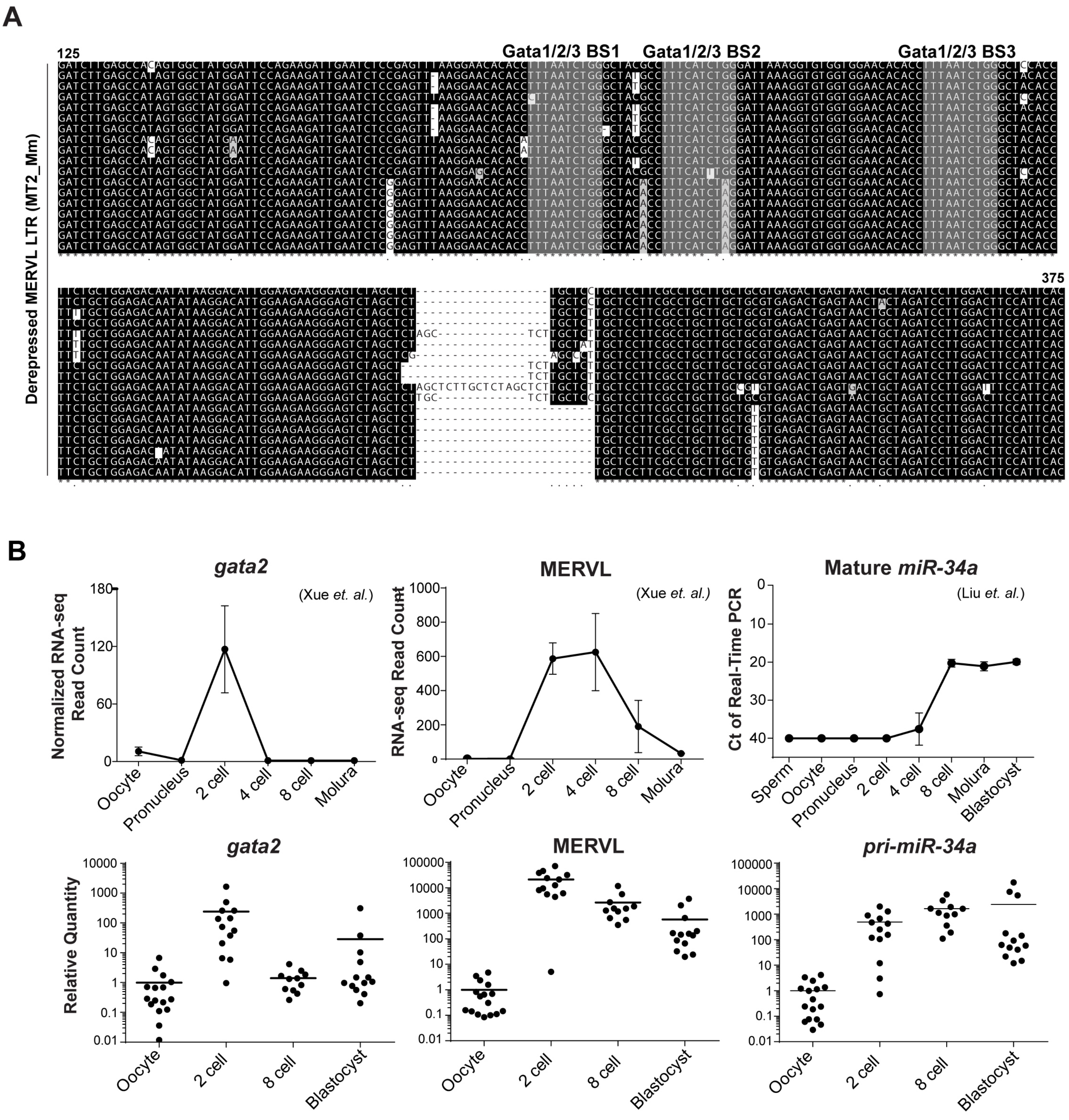
A
125
Gata1/2/3 BS1
Gata1/2/3 BS2
Gata1/2/3 BS3
Derepressed MERVL LTR (MT2_Mm)
375
B
gata2
(Xue et. al.)
Normalized RNA-seq Read Count
180
120
60
0
Oocyte
Pronucleus
2 cell
4 cell
8 cell
Molura
MERVL
(Xue et. al.)
RNA-seq Read Count
1000
800
600
400
200
0
Oocyte
Pronucleus
2 cell
4 cell
8 cell
Molura
Mature miR-34a
(Liu et. al.)
Ct of Real-Time PCR
0
10
20
30
40
Sperm
Oocyte
Pronucleus
2 cell
4 cell
8 cell
Molura
Blastocyst
gata2
Relative Quantity
10000
1000
100
10
1
0.1
0.01
Oocyte
2 cell
8 cell
Blastocyst
MERVL
100000
10000
1000
100
10
1
0.1
0.01
Oocyte
2 cell
8 cell
Blastocyst
pri-miR-34a
100000
10000
1000
100
10
1
0.1
0.01
Oocyte
2 cell
8 cell
Blastocyst

Fig. S5 (Cont'd)

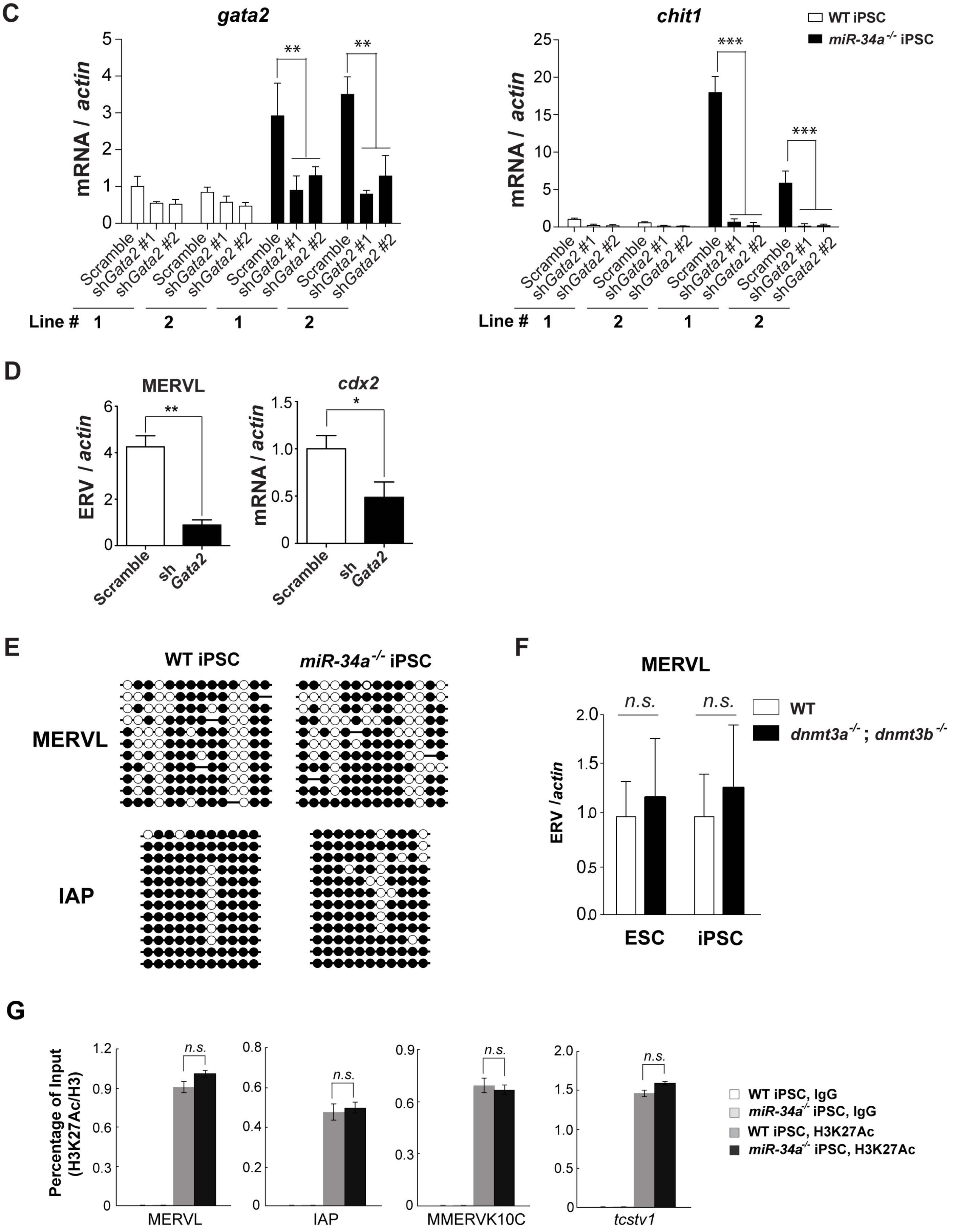
C
gata2
chit1
WT iPSC
miR-34a-/- iPSC
mRNA / actin
0
1
2
3
4
5
10
15
20
25
**
***
Scramble
shGata2 #1
shGata2 #2
Line #
1
2
D
MERVL
cdx2
ERV / actin
mRNA / actin
0.5
1.0
1.5
6
*
sh Gata2
E
WT iPSC
miR-34a-/- iPSC
MERVL
IAP
F
n.s.
WT
dnmt3a-/-; dnmt3b-/-
ERV /actin
0.0
2.0
ESC
iPSC
G
Percentage of Input (H3K27Ac/H3)
0.3
0.6
0.9
1.2
0.2
0.4
0.8
MERVL
IAP
MMERVK10C
tcstv1
WT iPSC, IgG
miR-34a-/- iPSC, IgG
WT iPSC, H3K27Ac
miR-34a-/- iPSC, H3K27Ac

Fig. S5 (Cont'd)

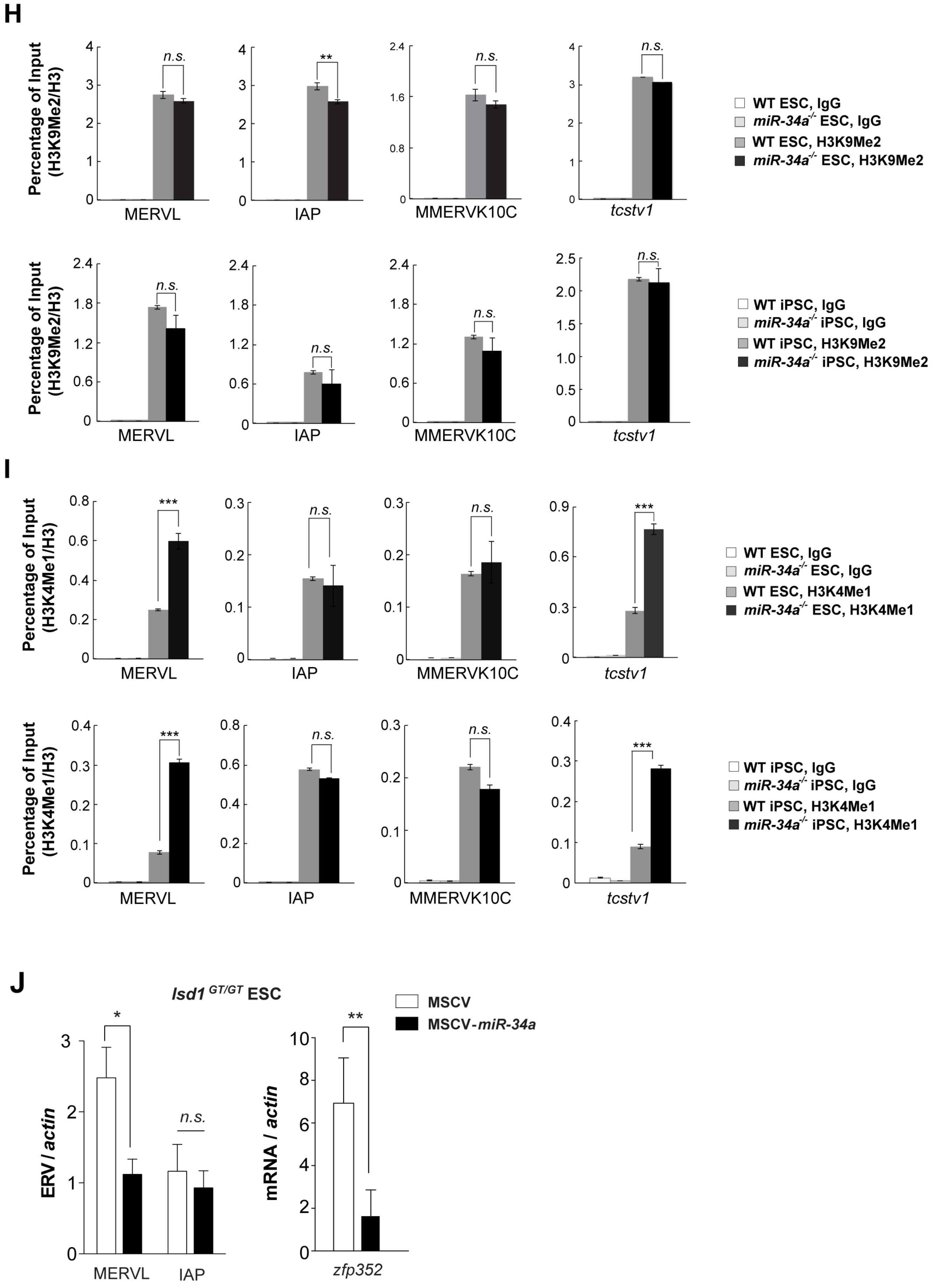
H
Percentage of Input (H3K9Me2/H3)
n.s.
**
n.s.
n.s.
MERVL
IAP
MMERVK10C
tcstv1
□ WT ESC, IgG
□ miR-34a-/- ESC, IgG
■ WT ESC, H3K9Me2
■ miR-34a-/- ESC, H3K9Me2
Percentage of Input (H3K9Me2/H3)
n.s.
n.s.
n.s.
n.s.
MERVL
IAP
MMERVK10C
tcstv1
□ WT iPSC, IgG
□ miR-34a-/- iPSC, IgG
■ WT iPSC, H3K9Me2
■ miR-34a-/- iPSC, H3K9Me2
I
Percentage of Input (H3K4Me1/H3)
***
n.s.
n.s.
***
MERVL
IAP
MMERVK10C
tcstv1
□ WT ESC, IgG
□ miR-34a-/- ESC, IgG
■ WT ESC, H3K4Me1
■ miR-34a-/- ESC, H3K4Me1
Percentage of Input (H3K4Me1/H3)
***
n.s.
n.s.
***
MERVL
IAP
MMERVK10C
tcstv1
□ WT iPSC, IgG
□ miR-34a-/- iPSC, IgG
■ WT iPSC, H3K4Me1
■ miR-34a-/- iPSC, H3K4Me1
J
lsd1 GT/GT ESC
□ MSCV
■ MSCV-miR-34a
ERV / actin
*
n.s.
MERVL
IAP
mRNA / actin
**
zfp352

Fig. S5. Gata2 mediates the MERVL induction in *miR-34a$^{-/-}$* pluripotent stem cells.

A. Clustal-W LTR sequence alignment of 18 differentially expressed MERVL loci in *miR-34a$^{-/-}$* iPSCs reveals three conserved predicted Gata1/2/3 binding sites within the minimal region of MERVL LTR, MERVL$_{125-375.}$ Among the three predicted GATA binding sites (designated as BS1, BS2, and BS3 and highlighted in yellow), BS1 and BS3 are fully conserved across all 18 MERVL elements, while BS2 is partially conserved. **B**. Expression patterns of MERVL, *gata2* and *miR-34a* during mouse preimplantation development. In published datasets (*52*, *53*), the levels of MERVL and *gata2* both peak in 2C embryos; and the mature *miR-34a* is highly expressed from the 8C to the blastocyst stage (top). Using single-embryo real-time PCR analyses, we validated the expression patterns of MERVL, *gata2,* and *pri-miR-34a* in mouse preimplantation embryos (bottom). **C**. The expression level of the MERVL proximal genes is dependent on *gata2*. Two independent shRNAs against *gata2* are able to effectively knock down *gata2* and the MERVL proximal gene *chit1* in *miR-34a$^{-/-}$* iPSCs. Two independent pairs of passage- and littermate-controlled wild-type and *miR-34a$^{-/-}$* iPSC lines were compared. Error bars: *s.d.*, n=3. **D**. *gata2* is necessary for MERVL and *cdx2* induction during teratoma formation. In teratomas generated from *miR-34a$^{-/-}$* iPSCs, knockdown of *gata2* in *miR-34a$^{-/-}$* iPSCs reduces the MERVL and *cdx2* levels during teratoma formation. **E**, **F**. DNA methylation is not essential for the MERVL induction in *miR-34a$^{-/-}$* pluripotent stem cells. **E**. Wild-type and *miR-34a$^{-/-}$* iPSCs have similar level of modest DNA methylation on MERVL elements, as determined by bisulfite sequencing. In contrast, iPSCs of both genotypes exhibit a high level of DNA methylation on the IAP elements. Black circle, methylated CpG; open circle, unmethylated CpG. **F**. No MERVL induction is detected in *dnmt3a$^{-/-}$*; *dnmt3b$^{-/-}$* ESCs and iPSCs that are deficient for *de novo* DNA methylation. Error bars: *s.d.*, n=3. *n.s.*, not significant. **G-I** Characterization of epigenetic modifications on MERVL in wild-type and *miR-34a$^{-/-}$* pluripotent stem cells. Wild-type and *miR-34a$^{-/-}$* pluripotent stem cells have similar deposition of H3K27Ac (**G**) and H3K9Me2 (**H**) on the MERVL LTR and the MERVL-*tcstv1* chimeric gene, yet the deposition of H3K4Me1 (**I**) on MERVL is increased in *miR-34a$^{-/-}$* pluripotent stem cells. As a control, H3K27Ac (**G**), H3K9Me2 (**H**) and H3K4Me1 (**I**) deposition on IAP LTR or MMERVK10C LTR is similar between wild-type and *miR-34a$^{-/-}$* pluripotent stem cells. Error bars: *s.d.*, n=3. Two independent pairs of passage- and littermate-controlled wild-type and *miR-34a$^{-/-}$* ESCs and iPSCs are compared. **J**. *miR-34a* overexpression in *lsd1* deficient (*lsd1$^{GT/GT}$*) ESCs using a murine stem cell virus (MSCV) retroviral vector effectively suppresses the level of MERVL and the MERVL proximal gene *zfp352*, but causes no alteration in IAP. Error bars: *s.d.*, n=3. All *P*-values were calculated on a basis of a two-tailed Student's *t*-test. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, *n.s.*, not significant.

Fig. S6

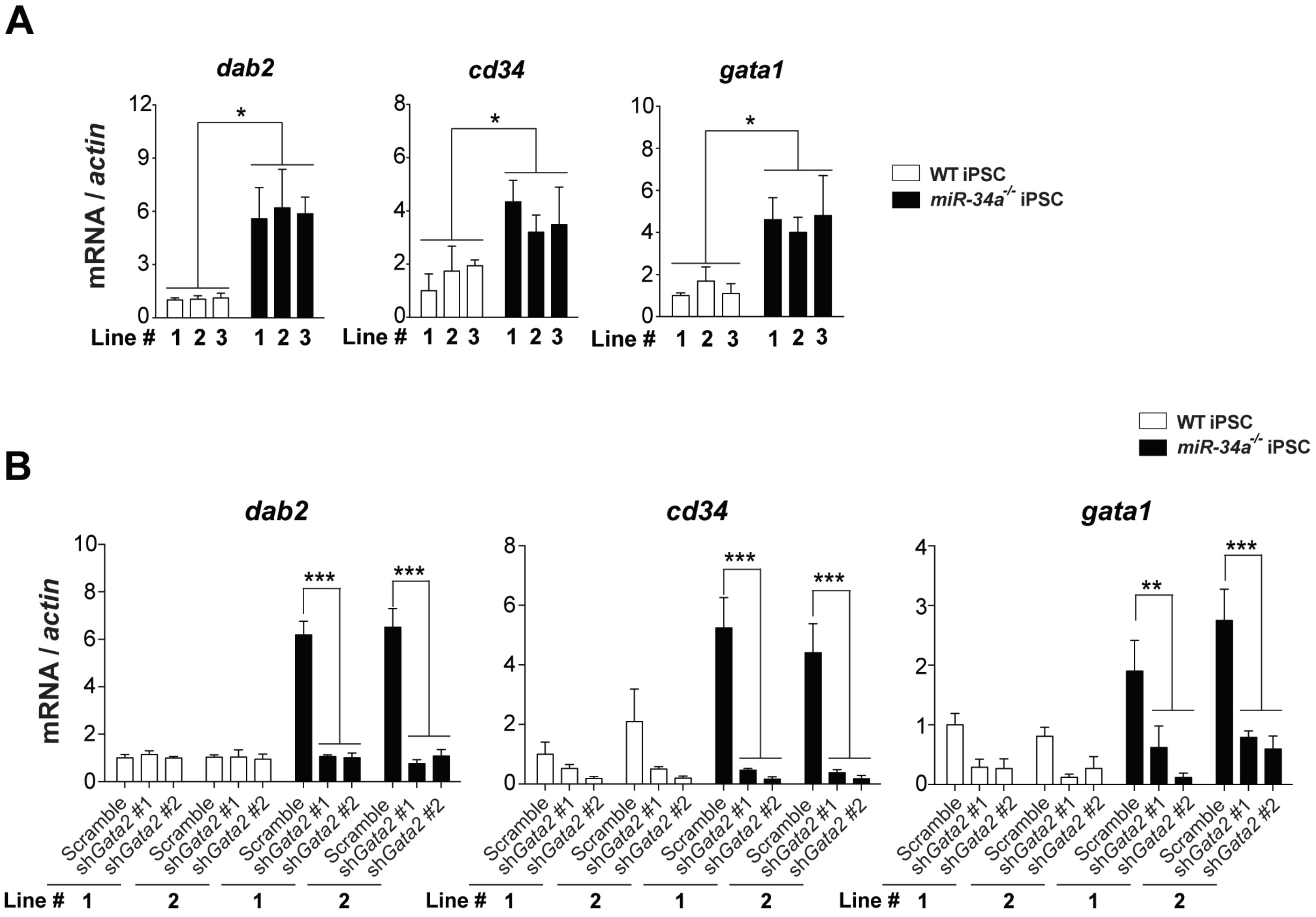
A
dab2
cd34
gata1
mRNA / actin
12
9
6
3
0
8
6
4
2
0
10
8
6
4
2
0
*
*
*
Line # 1 2 3 1 2 3
Line # 1 2 3 1 2 3
Line # 1 2 3 1 2 3
WT iPSC
miR-34a-/- iPSC
B
WT iPSC
miR-34a-/- iPSC
dab2
cd34
gata1
mRNA / actin
10
8
6
4
2
0
8
6
4
2
0
4
3
2
1
0
***
***
***
***
**
***
Scramble
shGata2 #1
shGata2 #2
Line # 1 2 1 2
Line # 1 2 1 2
Line # 1 2 1 2

Fig. S6. Gata2 is a key target of *miR-34a* in pluripotent stem cells.

A. Consistent with Gata2 being a key target for *miR-34a*, *miR-34a*$^{-/-}$ iPSCs exhibit an increase of multiple well-characterized *gata2* targets (*dab2, cd34* and *gata1*) in our real-time PCR analyses. Three independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ iPSC lines were compared. Error bars, *s.d.*, n=3. **B**. The expression level of characterized *gata2* targets is dependent on *gata2* in *miR-34a*$^{-/-}$ iPSCs. Two independent *gata2* shRNAs are able to effectively suppress *dab2*, *cd34* and *gata1* in *miR-34a*$^{-/-}$ iPSCs. Two independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ iPSC lines were compared by real-time PCR analyses. Error bars: *s.d.,* n=3. All *P*-values were calculated on a basis of a two-tailed Student's *t*-test. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Table S1. *miR-34a*$^{-/-}$ ESCs contribute to both embryonic and extra-embryonic cell lineages in chimeric analyses *in vivo*.

Table S2. Expression quantification of all retrotransposon families in wild-type and *miR-34a*$^{-/-}$ iPSCs using RNA-seq data.

Table S3. Expression quantitation of individual MERVL loci and MERVL-related ERV loci in wild-type and *miR-34a*$^{-/-}$ iPSCs using RNA-seq data.

Table S4. A summary of genes differentially expressed between wild-type and *miR-34a*$^{-/-}$ iPSCs using RNA-seq data.

Table S5. Quantitation of chimeric junction reads between MERVL or MERVL-related loci and proximal protein-coding genes in wild-type and *miR-34a*$^{-/-}$ iPSCs using RNA-seq data.

Table S6. The quantitative PCR primers used in this study.

Supplemental Information. The reannotation of retrotransposons in GFF format using the *REAnnotate* program to merge fragmented Repeat Masker annotations that belong to a single retrotransposons element. (see “Retrotransposon annotation” in Supplementary Text for the detail). This reannotation allows us to distinguish between ERVs with a complete or truncated gene structure, as well as ERVs that only contain a solo LTR. For each element we record a unique id, the retrotransposon family and class, percentage of divergence, insertions and deletions (from Repeat Masker).