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

HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2018 September 14.

Published in final edited form as:

Science. 2017 February 10; 355(6325): . doi:10.1126/science.aag1927.

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,*]

[1]Division of Cellular and Developmental Biology, MCB department, University of California at Berkeley, Berkeley, CA 94705, USA

[2]Division of Biostatistics, School of Public Health, University of California at Berkeley, Berkeley, CA 94720, USA

[3]Department of Genetics, Stanford University, Stanford, CA 94305, USA

[4]Thermo Fisher Scientific, 180 Oyster Point Blvd., South San Francisco, CA 94080

[5]Integrated DNA Technologies, 200 Chesapeake Drive, Redwood City, CA 94063, USA

[6]Department of Molecular and Cell Biology, University of Texas at Dallas, 800 West Campbell Road, Richardson, Texas 75080

[7]The Eunice Kennedy Shriver National Institute of Child Health and Human Development, the National Institutes of Health, Bethesda, MD, USA

[8]Department of Physics, the George Washington University, Washington, DC 20052, USA

[9]Department of Pathology, NYU School of Medicine, 540 First Ave, New York, NY 10016, USA

[10]Department of Statistics, University of California, Berkeley, CA 94720 USA; Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010 Australia and Bioinformatics Division. The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052 Australia

Abstract

Embryonic stem cells and induced pluripotent stem cells have pluripotent developmental potential, efficiently giving rise to all embryonic cell types, but rarely extraembryonic lineages (1). Here, we identify a microRNA *miR-34a*, whose deficiency in mouse pluripotent stem cells expands their developmental potential to generate both embryonic and extra-embryonic lineages *in vitro* and *in vivo*. *miR-34a*$^{-/-}$ pluripotent stem cells with this bidirectional cell fate potential resemble

*Correspondence to: lhe@berkeley.edu (LH) and ewcplin@gmail.com (CPL).
†These authors contributed equally.

Supplementary Materials:
Materials and Methods
Supplementary Text
Fig. S1–S6
Tables S1–S6
Supplementary Information

totipotent 2-cell (2C) blastomeres not only in their cell fate potential, but also in the key molecular signature, namely a strong induction of the MuERV-L (MERVL) family of murine endogenous retroviruses (ERVs). *miR-34a* represses MERVL expression through transcriptional regulation, at least in part, by repressing the transcription factor GATA-binding protein 2 (Gata2). Consistently, the *miR-34a*/Gata2 pathway restricts the acquisition of bidirectional cell fate potential in pluripotent stem cells. Altogether, our findings provide vital insights into the complex molecular network that defines and restrict the developmental potential of pluripotent stem cells.

Mouse embryonic stem cells (ESCs) derived from the inner cell mass (ICM) of blastocysts, as well as induced pluripotent stem cells (iPSCs) generated by somatic reprogramming, are classically defined as pluripotent stem cells (2–4). As a population, ESCs and iPSCs efficiently contribute to all embryonic cell types *in vitro* and *in vivo*, but rarely to extra-embryonic cell lineages in placenta and yolk sac (1). This restricted pluripotent potential contrasts with that of early blastomeres, which give rise to both embryonic and extra-embryonic cell lineages during normal development (5, 6). Interestingly, rare pluripotent stem cell populations with expanded cell fate potential have been identified in culture as a result of genetic alterations, specific culture and derivation conditions, or enrichment with specific molecular markers (7–10). Such pluripotent stem cells exhibit bidirectional cell fate potential, contributing to both embryonic and extra-embryonic lineages in a variety of *in vitro* and *in vivo* functional assays (7–10). In addition, these rare ESCs/iPSCs with bidirectional potential often share a key molecular signature, namely a strong induction of the MuERV-L (MERVL) family of murine endogenous retroviruses (ERVs), which only occurs in totipotent 2C blastomeres during normal mouse development (7, 9). While these studies suggest that a subset of cultured ESCs/iPSCs retain the cell fate plasticity to acquire features of early blastomeres, there clearly exists a strong molecular barrier restricting the ESC/iPSC developmental potential to a pluripotent cell state. In this study, we identified the *miR-34a* miRNA as the first non-coding regulator that restricts the pluripotent cell fate potential in cultured ESCs/iPSCs, the deficiency of which yields bidirectional cell fate potential and MERVL induction in pluripotent stem cells.

miR-34a$^{-/-}$ pluripotent stem cells exhibit expanded cell fate potential

microRNAs (miRNAs) are a class of small, regulatory non-coding RNAs that regulate gene expression post-transcriptionally through a combined mechanism of mRNA degradation and translational repression (11–13). These small non-coding RNAs are increasingly recognized as key regulators of cell fate specification in normal development and in pluripotent stem cells (14, 15).

Initially identified as *bona fide* p53 transcriptional targets in tumor suppression, the *miR-34* miRNAs (*miR-34a*, *miR-34b* and *miR-34c*), particularly *miR-34a*, have been previously characterized as a key barrier for somatic reprogramming (16). *miR-34a* deficiency significantly enhances the efficiency of iPSC generation (16), producing iPSCs with normal self-renewal and pluripotency (Fig. S1A, S1B and S1C; ref. 16). Surprisingly however, teratomas generated from *miR-34a*$^{-/-}$ iPSCs, but not wild-type iPSCs, contained cellular features reminiscent of trophoblast giant cells in the placenta, characterized by PL-1

(placental lactogen 1) expression, large cell volume, enlarged nuclei, and close proximity to internal hemorrhages (Fig. 1A). In ESCs, m*iR-34a* constitutes the majority of expressed *miR-34* miRNAs (Fig. S1D). Similarly, *miR-34a*$^{-/-}$ ESC derived teratomas, but not the wild-type controls, also contained areas reminiscent of extraembryonic placental cell lineages (Fig.1A) and exhibited an induction of trophectoderm (TE) markers (Fig. S1E), including *cdx2*, *elf5, psx1, fgfr2, egfr* and *mdfi* (17, 18). While we did not identify any areas morphologically resembling the visceral endoderm of the yolk sac, we detected a strong induction of primitive endoderm (PE) markers (*gata4, gata6 and sox17*) in *miR-34a*$^{-/-}$ teratomas, but not in wild-type controls (Fig. S1E). These findings suggest that *miR-34a*$^{-/-}$ pluripotent stem cells likely differentiate towards both embryonic and extra-embryonic cell lineages during teratoma formation.

The expanded potential of *miR-34a*$^{-/-}$ ESCs/iPSCs is also evident upon embryoid body (EB) differentiation (Fig. 1B and 1C). While markers from all three germ layers were similarly induced in wild-type and *miR-34a*$^{-/-}$ EBs, significant upregulation of TE markers (*cdx2*, *elf5*, *esx1*, *tfap2c* and *gata3*) (17, 19, 20) was observed only in *miR-34a*$^{-/-}$ EBs (Fig. 1B, 1C and S1F). Immunofluorescence (IF) staining confirmed that a significant percentage of *miR-34a*$^{-/-}$ EBs was Cdx2 positive (Fig. 1B), and these Cdx2-positive cells preferably localized to the periphery (Fig. 1B, S1G and S1H). Additionally, the extra-embryonic endoderm marker *gata4* and *pdgfra*, as well as the trophoblast lineage marker *mash2 (ascl2)* and *pl1 (prl3d1)*, were also induced in *miR-34a*$^{-/-}$ EBs (Fig. S1F). Thus, upon EB differentiation, *miR-34a*$^{-/-}$ ESCs exhibited expanded cell fate potential, generating cells with molecular features characteristic of both embryonic and extra-embryonic lineages.

To define the cell fate potential of *miR-34a*$^{-/-}$ pluripotent stem cells in normal development, we traced their lineage in chimeric blastocysts following microinjection or aggregation with recipient morulae. Initially, four GFP-labeled wild-type or *miR-34a*$^{-/-}$ ESCs were injected into each C57BL/6N recipient morula to generate chimeric blastocysts (Fig. 1D). While wild-type ESCs exclusively gave rise to cells localized to the ICM (Fig. 1D; Table S1), *miR-34a*$^{-/-}$ ESC progenies localized to both ICM and TE in ~60% of chimeric blastocysts (Fig. 1D; Table S1). This expanded cell fate potential is unlikely due to extra-embryonic contamination during *miR-34a*$^{-/-}$ ESC derivation, as *miR-34a*$^{-/-}$ iPSCs derived from mouse embryonic fibroblasts (MEFs) phenocopied *miR-34a*$^{-/-}$ ESCs in their developmental potential. When aggregated with recipient C57BL/6J morulae, *miR-34a*$^{-/-}$ ESCs and *miR-34a*$^{-/-}$ iPSCs colonize both ICM and TE of chimeric blastocysts, while passage- and littermate-controlled wild-type ESCs and iPSCs exclusively colonized the ICM (Fig. S1I).

The expanded cell fate potential of *miR-34a*$^{-/-}$ ESCs in chimeric blastocysts could be due to the presence of cells with bidirectional potential; alternatively, *miR-34a*$^{-/-}$ ESCs could contain a heterogeneous population of cells that preferentially differentiate into embryonic or extra-embryonic cell lineages. To distinguish between these two possibilities, we injected single, GFP-labeled *miR-34a*$^{-/-}$ ESCs into each recipient morula to generate chimeric blastocysts (Fig. 1E). In two independent *miR-34a*$^{-/-}$ ESC lines tested, single *miR-34a*$^{-/-}$ ESCs colonized both ICM and TE in 33% and 38% of chimeric blastocysts (n=13/40 and 8/21) (Fig. 1E; Table S1) respectively, suggesting that a significant portion of *miR-34a*$^{-/-}$ ESCs exhibit a bidirectional developmental potential at the single-cell level.

We then generated chimeric embryos by microinjecting 10–15 GFP-labeled wild-type or *miR-34a*$^{-/-}$ ESCs into C57BL/6N recipient blastocysts. While wild-type ESCs contributed exclusively to lineages of the three embryonic germ layers, *miR-34a*$^{-/-}$ ESCs contributed to both embryonic and extra-embryonic cell lineages in E9.5, E12.5 and E14.5 chimeric embryos (Fig. 1F, 1G and S1J; Table S1). In particular, we observed clusters of GFP-positive, *miR-34a*$^{-/-}$ ESC progenies in the visceral endoderm of the yolk sac, as well as in multiple extra-embryonic trophoblast lineages of the placenta (trophoblast giant cells, spongiotrophoblasts, syncytiotrophoblasts (STBs) and sinusoidal trophoblast giant cells (s-TGCs), Fig. 1F, 1G; Table S1). In these chimeric embryos, the number of GFP-positive cells in extra-embryonic cell lineages greatly surpasses the number of *miR-34a*$^{-/-}$ ESCs injected (Fig. 1F and 1G), suggesting that injected *miR-34a*$^{-/-}$ ESCs had undergone substantial proliferation before committing to multiple terminally differentiated extra-embryonic lineages.

miR-34a$^{-/-}$ pluripotent stem cells exhibit an induction of MERVL ERVs

To investigate the molecular basis for the bidirectional potential of *miR-34a*$^{-/-}$ pluripotent stem cells, we compared the transcriptomes of wild-type and *miR-34a*$^{-/-}$ iPSCs using RNA-sequencing (RNA-seq). We compared the abundance of all annotated transcripts between wild-type and *miR-34a*$^{-/-}$ iPSCs, including protein-coding genes, long non-coding RNAs (ncRNAs), pseudogenes, antisense transcripts, and retrotransposons using 100 bp paired end RNA-seq data (Fig. 2A). Given the repetitive nature of retrotransposons, we quantified retrotransposon expression at the family level using both uniquely and non-uniquely mapped reads (Supplemental Information S1). Surprisingly, the most highly expressed and differentially regulated transcript in *miR-34a*$^{-/-}$ iPSCs was transcribed from the MERVL family of ERVs (Fig. 2A and S2A), which were also highly induced in totipotent 2C blastomeres and reported ESCs with expanded potential (7, 9, 21, 22). ERV induction in *miR-34a*$^{-/-}$ ESCs/iPSCs was largely specific to the MERVL family (Fig. 2A, 2B and S2A; Table S2). The majority of differentially expressed retrotransposons in *miR-34a*$^{-/-}$ iPSCs belonged to the canonical MERVL family of ERVs (a class-III ERV) (Fig. S2A; Table S2); a small fraction of differentially expressed loci belonged to the MT2A, MT2B, MT2B1, and MT2B2 ERV families that are highly related to the canonical MERVL solo LTR, MT2_Mm (Fig. S2A; Table S3).

Consistent with our RNA-seq results, we invariably detected a significant increase of MERVL expression in *miR-34a*$^{-/-}$ iPSCs and ESCs, using real-time PCR primer pairs designed from multiple highly conserved MERVL regions (Fig. 2B and 2C; data not shown). Interestingly, while MERVL induction in *miR-34a*$^{-/-}$ iPSCs persisted for more than 27 passages (Fig. S2B), MERVL was only induced in early passages of *miR-34a*$^{-/-}$ ESCs and became completely silenced around passage 12 (Fig. S2B). It is conceivable that MERVL expression in *miR-34a*$^{-/-}$ ESCs triggers additional mechanisms to re-establish their silencing. The expanded cell fate potential of *miR-34a*$^{-/-}$ ESCs was highly correlated with the strong MERVL induction, as the late passage (passage 17) *miR-34a*$^{-/-}$ ESCs lost both MERVL induction and the bidirectional potential (Fig. S2B and S2C).

The MERVL ERVs have been retained throughout mammalian evolution, with independent expansion in the murine and primate genomes (23). There are 2502 loci in the C57B6/J mouse genome, ~26% of which encode elements with an intact retroviral structure, comprising 5'- and 3'-LTRs flanking the coding sequences for *gag*, *pol*, and *dUTPase*, but lacking *env*-like open reading frames (ORFs) (23) (Fig. 2C). Another 32% of MERVL loci exhibit truncated retroviral structure, missing one or both LTRs (Fig. 2C). The remaining 41% of MERVL loci have undergone homologous recombination, yielding solo LTRs (MT2_Mm) with varying degrees of sequence degeneration (Fig. 2C). We obtained bioinformatic estimates of locus-specific MERVL expression in wild-type and *miR-34a*$^{-/-}$ iPSCs using our RNA-seq data (Table S3; Supplemental Information). Notably, definitive evidence for MERVL reactivation in *miR-34a*$^{-/-}$ iPSCs was observed predominantly for loci harboring MERVLs with a complete retroviral structure, but not for those with truncated structure (Fig. 2D and S2D; Table S3). A fraction of MT2_Mm solo LTRs, along with a few elements from the highly related MERVL solo LTRs (MT2B, MT2B1, MT2B2 and MT2A), also exhibited a similar induction (Fig. S2A; Table S3).

Approximately 300 MERVL loci still encode intact Gag viral protein (7). We observed a significant increase in MERVL-Gag expression and in the percentage of MERVL-Gag-positive cells in *miR-34a*$^{-/-}$ pluripotent stem cells (Fig. 2E and 2F). Interestingly, *miR-34a*$^{-/-}$ ESCs and iPSCs were heterogeneous populations, containing ~12% and ~20% MERVL-Gag-positive cells, respectively, in otherwise Oct4-positive colonies (Fig. 2E, 2F, S2E and S2G). Consistent with this observation, a fraction of individual *miR-34a*$^{-/-}$ iPSC colonies exhibited a significantly greater MERVL induction than the bulk population (Fig. S2F), suggesting that the extent of MERVL induction in individual cells was largely underestimated using the bulk population. In *miR-34a*$^{-/-}$ ESCs and iPSCs, the expression of MERVL-Gag and Oct4 were mutually exclusive (Fig. 2E, S2E and S2G), suggesting that the MERVL-positive, Oct4 negative *miR-34a*$^{-/-}$ cells possess a unique state of developmental potency, distinct from that of classic pluripotent stem cells characterized by Oct4 expression (7, 9, 24).

The global protein-coding gene expression profiles of *miR-34a*$^{-/-}$ iPSCs resemble those of reported bi-potential ESCs in hierarchical clustering (Fig. S3A). Intriguingly, among the most differentially expressed protein-coding genes in *miR-34a*$^{-/-}$ iPSCs were those proximal to MERVL loci (Fig. 2G). Indeed, differential expression analysis between reported ESCs with bidirectional cell fate potential (*lsd1*$^{-/-}$, 2C+, *p60* knockdown and *p150* knockdown ESCs) and their pluripotent controls revealed the induction of MERVL proximal genes as a key feature (Fig. S3B) (7, 9, 25). The MERVL derepression in *miR-34a*$^{-/-}$ iPSCs and ESCs also correlates with the induction of many protein-coding genes that harbor either a proximal upstream MERVL or an intronic MERVL on the same strand (Fig. 2G; Fig. S3C). In many cases, MERVL or related loci, particularly solo LTRs, act as alternative promoters, generating chimeric transcripts of proximal genes that differ in 5'-UTRs (Fig. 2H, S3D and S3E; Table S4 and S5). These MERVL-gene chimeric transcripts can be unambiguously identified by the corresponding splice junctions from the RNA-seq data (Table S5). Using real-time PCR analyses, we validated the induction of multiple MERVL proximal genes in *miR-34a*$^{-/-}$ ESCs and iPSCs, including t*cstv1, tcstv3, zfp352, cml2 and p4ha2* (harboring a proximal upstream MERVL element), as well as *abcb5*, *tmem132c* and

chit1 (harboring an intronic MERVL element) (Fig. 2H; Fig. S3D and S3E). The induction level of MERVL-gene isoforms varied among individual *miR-34a*$^{-/-}$ iPSC colonies, and largely correlated with the extent of MERVL induction (Fig. S3F). Consistently, *miR-34a* overexpression in *miR-34a*$^{-/-}$ iPSCs not only decreased MERVL expression, but also reduced the level of the MERVL-driven chimeric transcript (Fig. 2I).

We repeated our analysis using 150 bp pair-end RNA-seq data to gain confidence in the accuracy of our retrotransposon mapping (Fig. S4A). The analysis using longer sequence reads confirmed all our observations (Fig. S4A and S4B).

MERVL induction in miR-34a$^{-/-}$ pluripotent stem cells is regulated transcriptionally

Given the correlation between MERVL induction and bi-potential pluripotent stem cells (Fig. S3A and S3B; ref. 7, 9), the molecular pathway that mediates *miR-34a*-dependent MERVL repression could also regulate *miR-34a*-dependent restriction on pluripotent potential. To determine the key sequences required for MERVL induction, we transfected wild-type and *miR-34a*$^{-/-}$ ESCs with a MERVL$_{1–1000}$-Luc (luciferase) reporter containing the full-length LTR (MT2_Mm) and a portion of the *gag* sequence as the promoter (25). This reporter showed an elevated luciferase activity in *miR-34a*$^{-/-}$ ESCs compared to wild-type ESCs, faithfully recapitulating the endogenous MERVL induction (Fig. 3A). The LTR sequence was both necessary and sufficient for the MERVL$_{1–1000}$-Luc reporter activity in *miR-34a*$^{-/-}$ ESCs (Fig. 3A); furthermore, a minimal fragment, MERVL$_{125–375}$, containing a direct repeat and a TATA box, was sufficient to drive strong luciferase activity specifically in *miR-34a*$^{-/-}$ ESCs (Fig. 3A). The MERVL$_{125–375}$ fragment is highly conserved among all highly induced MERVL loci in *miR-34a*$^{-/-}$ iPSCs (Fig. S5A), suggesting a sequence-dependent transcriptional mechanism for MERVL induction. Consistently, *miR-34a*$^{-/-}$ ESCs and iPSCs also exhibited H3K4Me3 enrichment near the LTR of MERVL retrotransposons and the MERVL LTR proximal to *tcstv1*, but not near other ERVs such as IAP or MMERK10C (Fig. 3B). Thus, MERVL loci are specially enriched for active transcription machinery in *miR-34a*$^{-/-}$ pluripotent stem cells.

Gata2 mediates elevated MERVL expression in *miR-34a*$^{-/-}$ pluripotent stem cells

The MERVL$_{125–375}$ fragment likely contains *cis*-regulatory elements necessary and sufficient to enable MERVL induction in *miR-34a*$^{-/-}$ ESCs/iPSCs. We failed to detect any significant sequence complementarity between *miR-34a* (pri-, pre-, or mature miRNA sequences) and the MERVL$_{125–375}$ sequence, thus precluding a direct, RNA base-pairing-dependent repression mechanism. We predicted 70 candidate transcription factors that bind within MERVL$_{125–375}$, among which, only GATA-binding protein 2 (Gata2) exhibits an expression pattern similar to that of MERVL during early pre-implantation development (Fig. S5A; ref. 52, 53). Gata2 is also reported to play an important role in cell fate potency of ESCs (26).

We aligned the LTR sequences from 18 MERVL loci that were strongly induced in *miR-34a*$^{-/-}$ iPSCs, most of which contain three predicted Gata2 binding sites (Fig. S5A). Mutating the two most conserved sites within the MERVL$_{125–375}$-Luc reporter significantly reduced its activity in *miR-34a*$^{-/-}$ pluripotent stem cells (Fig. 3C). Similarly, *gata2* knockdown in *miR-34a*$^{-/-}$ pluripotent stem cells effectively abolished the induction of MERVL and MERVL proximal genes (*zfp352, tmem132c* and *chit1*) in cell culture (Fig. 3D; Fig. S5C), and significantly decreased MERVL and *cdx2* induction during teratoma formation (Fig. S5D). *gata2* knockdown also reduced H3K4Me3 deposition on MERVL elements and on the MERVL proximal gene *tcstv1*, suggesting a specific decrease in active transcriptional machinery on MERVL loci (Fig. 3E). Finally, using chromatin immunoprecipitation (ChIP), we demonstrated specific binding of Gata2 to the LTR region of MERVL, and not to the MERVL internal region or to other ERVs such as IAP or MMERK10C (Fig. 3F). Taken together, Gata2 plays an essential role in directly promoting the induction of MERVL ERVs and MERVL proximal genes in *miR-34a*$^{-/-}$ pluripotent stem cells.

Epigenetic modifications constitute another possible mechanism for MERVL induction in *miR-34a*$^{-/-}$ pluripotent stem cells. We investigated the role of DNA methylation on MERVL induction, but no difference was detected between wild-type and *miR-34a*$^{-/-}$ iPSCs (Fig. S5E), consistent with intact MERVL silencing in *dnmt3a*$^{-/-}$*; dnmet3b*$^{-/-}$ ESCs/iPSCs (Fig. S5F). MERVL was previously shown to be induced by a global decrease in H3K9Me2 or a global increase in H3K4Me1 in *g9a/glp*$^{-/-}$ ESCs or *lsd1*$^{-/-}$ ESCs, respectively (25, 27). Using ChIP, we found that H3K27Ac and H3K9Me2 deposition on MERVL was unaltered in *miR-34a*$^{-/-}$ pluripotent stem cells (Fig. S5G and S5H). Additionally, while *miR-34a*$^{-/-}$ ESCs and iPSCs exhibited a ~2–3 fold H3K4Me1 enrichment on MERVL loci (Fig. S5I), *miR-34a* overexpression silenced MERVL expression in wildtype and *lsd1*$^{-/-}$ ESCs with a similar efficiency (Fig. S5J), suggesting that the direct mechanism through which *miR-34a* silenced MERVL was likely independent of a global alteration of H3K4Me1. Hence, none of the tested epigenetic mechanisms appeared to be essential mechanisms for *miR-34a*-mediated MERVL repression in ESCs/iPSCs.

miR-34a restricts cell fate potential of pluripotent stem cells by directly repressing Gata2

Gata2 not only plays an essential role in mediating MERVL activation in *miR-34a*$^{-/-}$ ESCs/iPSCs (Fig. 3D), it also emerges as a strong candidate as a direct *miR-34a* target. *gata2* harbors three predicted *miR-34a* binding sites (28, 29) (Fig. 4A), and exhibited *miR-34a*-dependent regulation in pluripotent stem cells—*gata2* mRNA and protein are increased in *miR-34a*$^{-/-}$ pluripotent stem cells and reduced upon ectopic *miR-34a* overexpression (Fig. 4B and 4C). Similarly, a luciferase reporter containing a fragment of the *gata2* gene with all three predicted *miR-34a* binding sites exhibited *miR-34a*-dependent repression in luciferase assays (Fig. 4D). Mutating all three predicted *miR-34a* binding sites in this reporter abolished this *miR-34a*-dependent regulation (Fig. 4D). These findings suggest that *gata2* is a direct *miR-34a* target in ESCs/iPSCs that mediates MERVL regulation. Consistent with Gata2 derepression in *miR-34a*$^{-/-}$ ESCs/iPSCs, a number of previously characterized Gata2 targets, including *dab2, cd34* and *gata1* (26, 30, 31), exhibited a Gata2-dependent upregulation in *miR-34a*$^{-/-}$ iPSCs (Fig. S6A and S6B).

Knockdown of *gata2* in *miR-34a*$^{-/-}$ iPSCs phenocopies *miR-34a* overexpression, not only downregulating the expression of MERVL and MERVL-proximal genes (Fig. 3D and S5C), but also abolishing their bidirectional developmental potential to differentiate into both embryonic and extra-embryonic lineages (Fig. 4E and 4F). In EB differentiation assay, *gata2* knockdown or *miR-34a* overexpression in *miR-34a*$^{-/-}$ pluripotent cells impaired the induction of the TE marker *cdx2*, without affecting the induction of the markers for all three germ layers (Fig. 4E and 4F). More importantly, while the control infected *miR-34a*$^{-/-}$ iPSCs contributed to both ICM and TE in 53% of the chimeric blastocysts (n=8/15, Fig. 4G; Table S1), *miR-34a*$^{-/-}$ iPSCs with *gata2* knockdown lost this expanded cell fate potential (n=0/13, Fig. 4G; Table S1). To our knowledge, *miR-34a* is the first non-coding RNA known to restrict pluripotent cell fate potential in ESCs/iPSCs. While multiple *miR-34a* targets could act collectively to restrict cell fate potential and repress MERVL expression in pluripotent stem cells, the *miR-34a*/Gata2 axis clearly plays an essential role in this process (Fig. 4H).

Conclusion

Mouse zygotes and early blastomeres possess a totipotent cell fate potential, generating both embryonic and extra-embryonic cell types during normal development. This totipotent cell fate potential is gradually restricted during preimplantation development. By the late blastocyst stage, the separation of TE (which develops into the placenta to support embryonic development), epiblast (which forms the embryo proper) and PE (which develops into the yolk sac) signals the completion of the first cell fate specification event, which commits the developmental potential of cells to either embryonic or extra-embryonic lineages. ESCs and iPSCs in culture faithfully recapitulate the developmentally restricted, pluripotent cell fate potential of the epiblast, efficiently contributing to all embryonic cell lineages *in vivo*, but rarely to extra-embryonic lineages (1). Experimentally, ESCs can be induced into cells with bidirectional developmental potential, using somatic nuclear transfer, genetic modifications, and specific ESC/iPSC enrichment/culture/derivation procedures (7–10, 17, 32). The low efficiency of such experimental manipulations reflects the existence of multiple cellular and molecular impediments for the acquisition of a bidirectional cell fate potential.

miR-34a plays an important role in maintaining cell fate identity in multiple contexts, as it restricts pluripotent cell fate potential of ESCs/iPSCs and acts as a barrier for somatic reprogramming (16). To our knowledge, *miR-34a* is the first non-coding RNA whose deficiency in ESCs or iPSCs expands their developmental potential, generating a significant fraction of cells with bidirectional developmental potential. It is intriguing that *miR-34a* deficient pluripotent stem cells also induce the expression of MERVL ERVs; and this MERVL induction in ESCs/iPSCs appears correlated with their expanded cell fate potential in a number of experimental systems (Fig. 4H). MERVL induction could simply be an indicator of an unique 2C-like transcriptional and epigenetic state; alternatively, MERVL could functionally contribute to the establishment and/or maintenance of the 2C-like cell fate potential by rewiring gene regulatory networks to induce many 2C-specific, MERVL-driven genes (7, 9). An interesting parallel can be drawn to human ESCs, wherein fluctuating levels HERV-H family ERVs marks a dynamic population of naive-state pluripotency (33). Taken

together, these findings suggest the possibility that ERVs, while traditionally viewed as evolutionary remnants of invading foreign DNA sequences, could have important yet unrecognized developmental functions.

While *miR-34a* deficiency expands cell fate potential in pluripotent stem cells, *miR-34a*$^{-/-}$ mice undergo normal preimplantation development in laboratory conditions (16, 34). It is conceivable that the *miR-34* family miRNAs act redundantly with other mechanisms to repress MERVL expression in preimplantation development. It is also possible that an impaired totipotency to pluripotency transition in *miR-34a*$^{-/-}$ embryos is tolerated in mouse preimplantation development due to its considerable cell fate plasticity (6, 35). Nevertheless, *miR-34a*$^{-/-}$ pluripotent stem cells constitute a powerful experimental system to investigate the molecular basis underlying the developmental potential to both embryonic and extra-embryonic linages. Our studies provides vital insights into an intricate network of protein-coding genes, ncRNAs, and retrotransposons that act cooperatively to define cell fate plasticity and cell fate potential in pluripotent stem cells.

Materials and Methods

Derivation of ESCs and generation of iPSCs

For ESC derivation, uteri containing E3.5 wild-type or *miR-34a*$^{-/-}$ embryos were isolated from timed pregnancies, and transferred individually to a 12-well plate with irradiated MEF (mouse embryonic fibroblasts) feeders. After 5 days of incubation, embryo outgrowth was separated from TE, individually picked, and expanded in mouse ES medium. For iPSC generation, wildtype and *miR-34a*$^{-/-}$ primary MEFs were isolated from littermate-controlled E13.5 wild-type and *miR-34a*$^{-/-}$ embryos, infected with retroviruses generated by pMX retroviral vectors that encode mouse Oct4, Sox2 and Klf4, and cultured on irradiated MEF feeder in ES medium. Subsequently, single iPSC-like colonies were individually picked and expanded on irradiated MEF feeders to establish stable lines. Cell lines used for *in vivo* experiments were validated free from microplasma contamination by PCR.

RNA-seq data analysis

RNA-seq reads were mapped to the GRCm38 (mm10) reference genome with TopHat to quantify gene and retrotransposon expression levels (36). 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) (Table S4). We used EdgeR to test for differential expression between *miR-34a*$^{-/-}$ and wild-type iPSCs (37). We defined a gene as differentially expressed (DE) if it had an absolute log2-fold-change greater or equal than 2 and a False Discover Rate of 0.05 (Supplementary Text; Table S2–S5).

We performed the analysis using three datasets: HiSeq2000 100bp paired-end data (100PE); NextSeq500 150bp paired-end data (150PE); and a combined dataset obtained by pooling the reads of the two (combined). This allowed us to quantify the effect of read length and

sequencing depth on our analyses: the results are highly reproducible across the three datasets (Fig. S4; Table S2–S5).

Real-time PCR analysis for gene expression

RNA was isolated by Trizol extraction following manufacturer's instruction (Life Technologies, Cat. # 15596). cDNA was reverse-transcribed using iScript Advanced Reverse-Transcriptase (Bio-Rad, Cat. # 1725037). For single colony analysis, cDNA was prepared using a Single Cell-to-Ct qRT-PCR kit (Life Technologies, Cat. # 4458236). All real-time qPCR analyses were performed using SYBR FAST qPCR Master Mix (Kapa Biosystems, Cat. # KK4604). All primers used are listed in Table S6. To detect MERVL expression, four pairs of primers were designed to amplify specific regions of MERVL (Fig. 3C) and yielded similar results (data now shown). One pair of primers detecting the MERVL *pol* region was used for all other MERVL real-time PCR analyses.

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. Samples were collected at day 0, 3, 6 and 9 post differentiation for real-time PCR analyses and for immunofluorescence staining.

Generation of chimeric blastocysts or chimeric mice from ESCs/iPSCs

To generate chimeric blastocysts by microinjection, four ESCs or one ESC of the desired genotype were injected into each E2.5 C57Bl/6N wild-type, recipient morulae, which was then cultured overnight to obtain chimeric blastocysts. To generate chimeric blastocysts by morula aggregation, one-cell embryos were cultured for 48h into morulae. After the removal of zona pellucida, morulae were aggregated with ESCs or iPSCs and then cultured overnight. To generate chimeric mice by microinjection, 10–15 ESCs of the desired genotype were injected into E3.5 recipient blastocyst embryos before implanted into the uterus of pseudopregnant mothers. Chimeric embryos were collected at E9.5, E12.5 and E14.5 for immunofluorescence (IF) and immunohistochemistry analyses.

Immunofluorescence (IF) and Immunohistochemistry (IHC)

For IF staining, cells, EBs or embryos paraffin sections were fixed and blocked before incubated with primary antibodies (1:100, Cdx2, Abcam, Cat. # ab76541; 1:100, Gata4, Santa Cruz Biotechnology, Cat. # sc-9053), followed by secondary antibody incubation and DAPI staining. For IHC of teratomas and chimeric mouse embryos, tissue paraffin sections were incubated with primary antibodies against PL-1 (1:75, Santa Cruz Biotechnology, Cat. # sc-34713) or GFP (1:100, Abcam, Cat. # ab38689).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

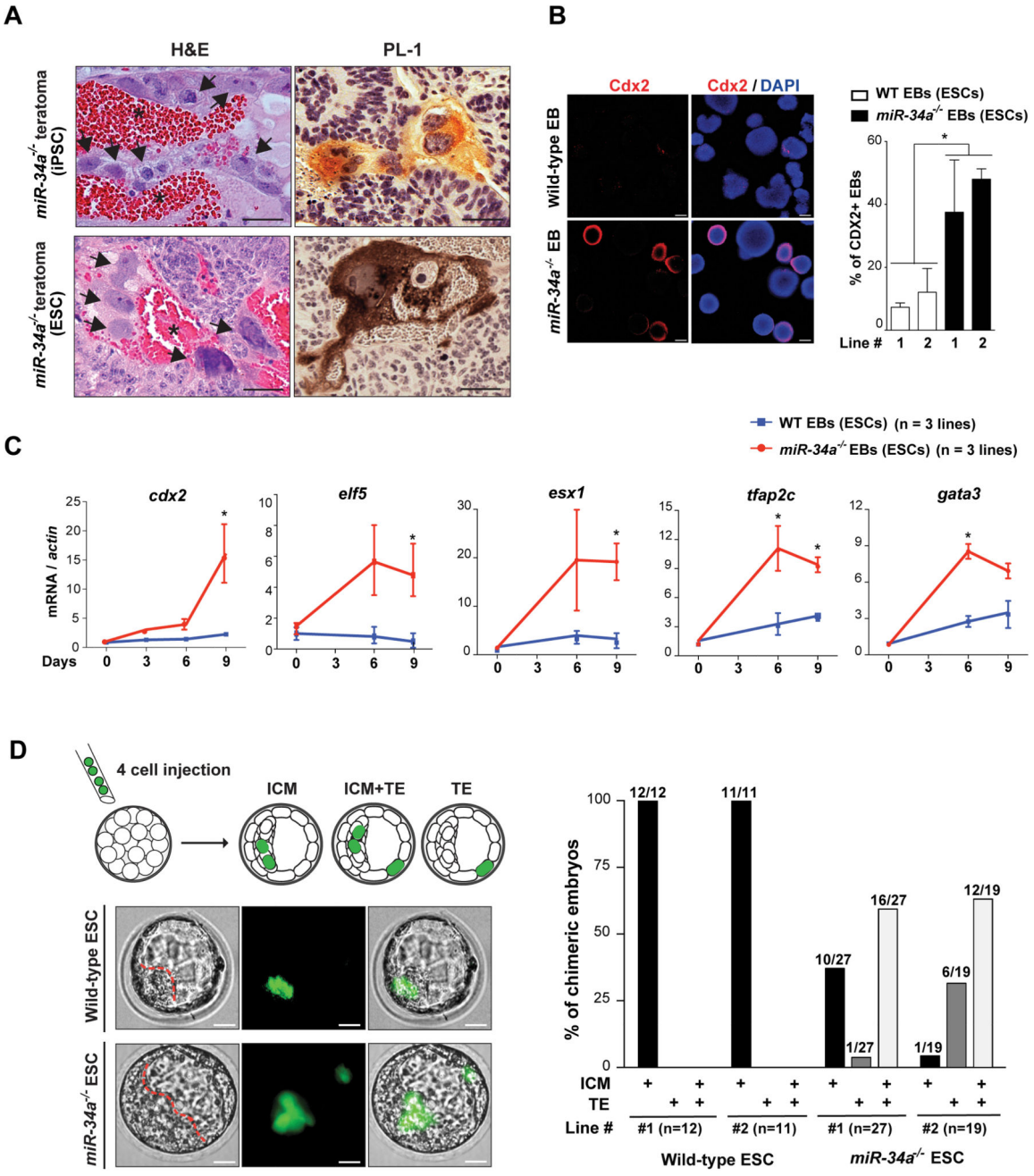
We thank V. Prideaux, W. Wang, T. Kim, R. Huang, K. N. Li, H. Aaron, J-Y. Lee, W. Xu, J. Ong, P. Cheung, B. Zaghi, M. Chung, J. Choi, A. Li, A. Perez, W. Bao, S. Tindall, K. Zhao, K. Cui, B. Xue, O. Tam, K. Heydari, A. Valeros, MJ Bennett, and H. Noller for technical assistance. We thank L. Xie, V. A. Modzelewski and R. Song for discussion and input. We also thank T. Heidmann, J. Rossant, A. Li, V. Krizhanovsky, M. Stadtfeld, M.C. Lorincz, Y. Shinkai, D. Trono, T. Chen and R. Jaenisch for sharing valuable reagents. P. Margolis for carefully reading the manuscript and M. Rape and N. Patel for sharing Olympus Revolution XD spinning disk confocal microscope and Zeiss LSM 700 confocal microscope with us. L.H. is supported by a new faculty award by California Institute for Regenerative Medicine (RN2-00923-1) and an R01 (R01 CA139067) from the National Cancer Institute (NCI). S.C. is supported by a CIRM predoctoral fellowship. C-P.L. is supported by a Siebel postdoctoral fellowship and a CIRM postdoctoral fellowship.

References and Notes

1. Beddington RS, Robertson EJ. An assessment of the developmental potential of embryonic stem cells in the midgestation mouse embryo. Development. 1989; 105:733–737. [PubMed: 2598811]
2. Evans MJ, Kaufman MH. Establishment in culture of pluripotential cells from mouse embryos. Nature. 1981; 292:154–156. [PubMed: 7242681]
3. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006; 126:663–676. [PubMed: 16904174]
4. Martin GR. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proceedings of the National Academy of Sciences. 1981; 78:7634–7638.
5. Papaioannou VE, Mkandawire J, Biggers JD. Development and phenotypic variability of genetically identical half mouse embryos. Development. 1989; 106:817–827. [PubMed: 2562672]
6. Tarkowski AK. Experiments on the development of isolated blastomeres of mouse eggs. Nature. 1959; 184:1286–1287. [PubMed: 13836947]
7. Macfarlan TS, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 2012; 487:57–63. [PubMed: 22722858]
8. Morgani SM, et al. Totipotent Embryonic Stem Cells Arise in Ground-State Culture Conditions. Cell Reports. 2013; 3:1945–1957. [PubMed: 23746443]
9. Ishiuchi T, et al. Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. Nat Struct Mol Biol. 2015; 22:662–671. [PubMed: 26237512]
10. Abad M, et al. Reprogramming in vivo produces teratomas and iPS cells with totipotency features. Nature. 2013; 502:340–345. [PubMed: 24025773]
11. Ambros V. The functions of animal microRNAs. Nature. 2004; 431:350–355. [PubMed: 15372042]
12. Bartel DP, Lee R, Feinbaum R. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function Genomics. Cell. 2004; 116:281–297. [PubMed: 14744438]
13. Zamore PD, Haley B. Ribo-gnome: the big world of small RNAs. Science. 2005; 309:1519–1524. [PubMed: 16141061]
14. Kanellopoulou C, et al. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes & development. 2005; 19:489–501. [PubMed: 15713842]
15. Judson RL, Babiarz JE, Venere M, Blelloch R. Embryonic stem cell-specific microRNAs promote induced pluripotency. Nature biotechnology. 2009; 27:459–461.
16. Choi YJ, et al. miR-34 miRNAs provide a barrier for somatic cell reprogramming. Nature Cell Biology. 2011; 13:1353–1360. [PubMed: 22020437]
17. Niwa H, et al. Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. Cell. 2005; 123:917–929. [PubMed: 16325584]
18. Giakoumopoulos M, Golos TG. Embryonic stem cell-derived trophoblast differentiation: a comparative review of the biology, function, and signaling mechanisms. Journal of Endocrinology. 2013; 216:R33–R45. [PubMed: 23291503]
19. Kubaczka C, et al. Direct Induction of Trophoblast Stem Cells from Murine Fibroblasts. Stem Cell. 2015; 17:557–568.

20. Benchetrit H, et al. Extensive Nuclear Reprogramming Underlies Lineage Conversion into Functional Trophoblast Stem-like Cells. Stem Cell. 2015; 17:543–556.
21. Kigami D, Minami N, Takayama H, Imai H. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. Biology of reproduction. 2003; 68:651–654. [PubMed: 12533431]
22. Peaston AE, et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Developmental Cell. 2004; 7:597–606. [PubMed: 15469847]
23. Bénit L, Lallemand JB, Casella JF, Philippe H, Heidmann T. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. Journal of virology. 1999; 73:3301–3308. [PubMed: 10074184]
24. Nichols J, et al. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. Cell. 1998; 95:379–391. [PubMed: 9814708]
25. Macfarlan TS, et al. Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes & development. 2011; 25:594–607. [PubMed: 21357675]
26. Zhang C, Ye X, Zhang H, Ding M, Deng H. GATA factors induce mouse embryonic stem cell differentiation toward extraembryonic endoderm. Stem cells and development. 2007; 16:605–613. [PubMed: 17784834]
27. Maksakova IA, et al. Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. Epigenetics & chromatin. 2013; 6:1–1. [PubMed: 23289424]
28. Lewis BP, Shih I-H, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 2003; 115:787–798. [PubMed: 14697198]
29. Miranda KC, et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell. 2006; 126:1203–1217. [PubMed: 16990141]
30. Orkin SH. GATA-binding transcription factors in hematopoietic cells. Blood. 1992; 80:575–581. [PubMed: 1638017]
31. Suzuki M, et al. GATA factor switching from GATA2 to GATA1 contributes to erythroid differentiation. Genes to Cells. 2013; 18:921–933. [PubMed: 23911012]
32. Wilmut I, Schnieke AE, McWhir J, Kind AJ, Campbell KH. Viable offspring derived from fetal and adult mammalian cells. Nature. 1997; 385:810–813. [PubMed: 9039911]
33. Wang J, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014; 516:405–409. [PubMed: 25317556]
34. Concepcion CP, et al. Intact p53-dependent responses in miR-34-deficient mice. PLoS Genet. 2012; 8:e1002797. [PubMed: 22844244]
35. Rossant J. Postimplantation development of blastomeres isolated from 4- and 8-cell mouse eggs. Journal of embryology and experimental morphology. 1976; 36:283–290. [PubMed: 1033982]
36. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
37. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139. [PubMed: 19910308]
38. Tan MH, et al. RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. Genome Research. 2013; 23:201–216. [PubMed: 22960373]
39. Westerman BA, et al. A genome-wide RNAi screen in mouse embryonic stem cells identifies Mp1 as a key mediator of differentiation. The Journal of experimental medicine. 2011; 208:2675–2689. [PubMed: 22143885]
40. He L, et al. A microRNA component of the p53 tumour suppressor network. Nature. 2007; 447:1130–1134. [PubMed: 17554337]
41. Li MA, Pettitt SJ, Yusa K, Bradley A. Genome-wide forward genetic screens in mouse ES cells. Methods in Enzymology. 2010; 477:217–242. [PubMed: 20699144]
42. Willett RT, Greene LA. Gata2 is required for migration and differentiation of retinorecipient neurons in the superior colliculus. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2011; 31:4444–4455. [PubMed: 21430145]

43. Cunningham F, et al. Ensembl 2015. Nucleic Acids Research. 2014; 43:D662–D669. [PubMed: 25352552]
44. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30:923–930. [PubMed: 24227677]
45. Gentleman RC, Carey VJ, Bates DM. Others, Bioconductor: Open software development for computational biology and bioinformatics. Genome biology. 2004; 5:R80. [PubMed: 15461798]
46. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011; 12:323. [PubMed: 21816040]
47. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC bioinformatics. 2010; 11:94. [PubMed: 20167110]
48. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8:118–127. [PubMed: 16632515]
49. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; 28:882–883. [PubMed: 22257669]
50. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. 1979
51. Pereira V. Automated paleontology of repetitive DNA with REANNOTATE. BMC genomics. 2008; 9:614. [PubMed: 19094224]
52. Xue Z, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013; 500:593–597. [PubMed: 23892778]
53. Liu W-M, et al. Sperm-borne microRNA-34c is required for the first cleavage division in mouse. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:490–494. [PubMed: 22203953]



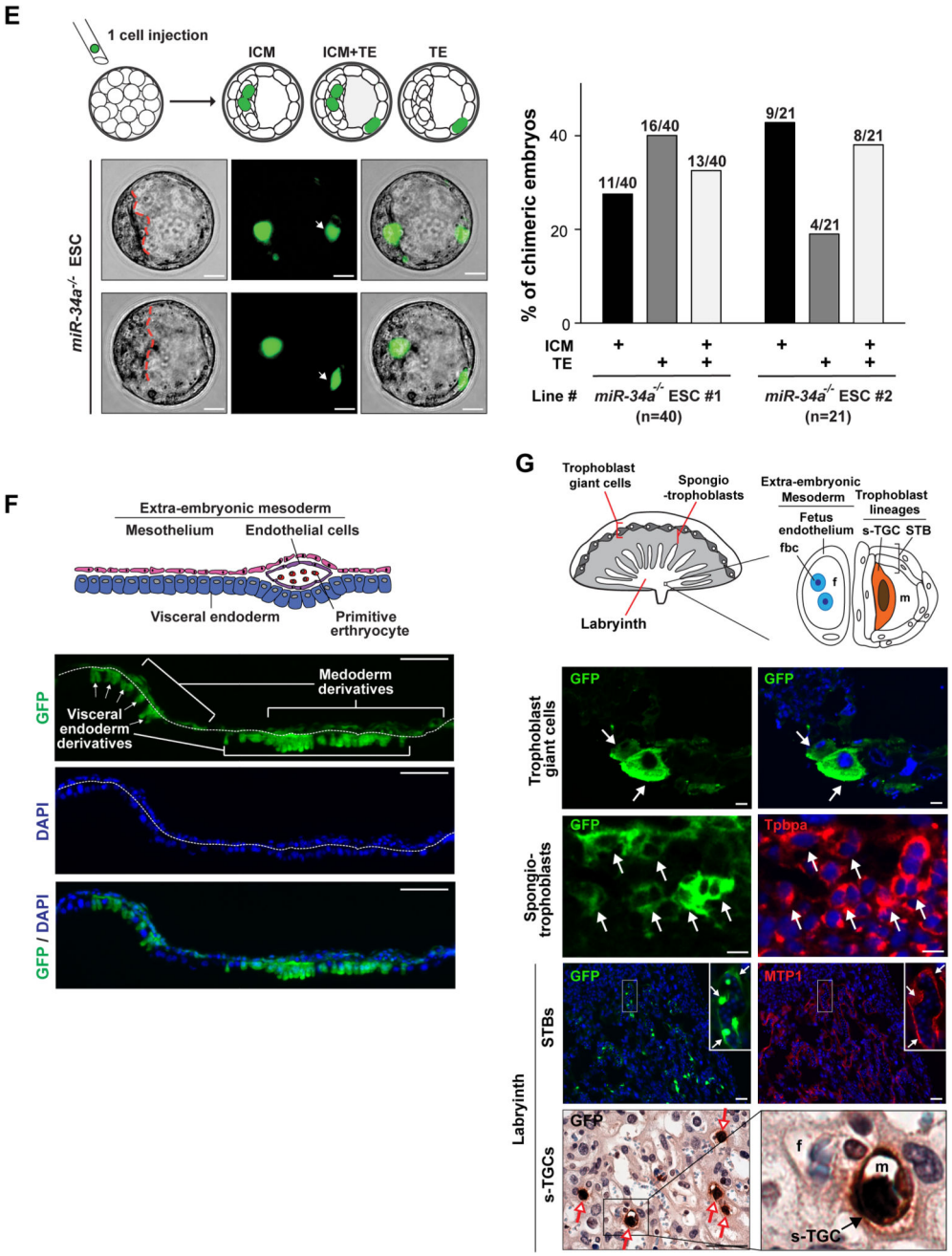
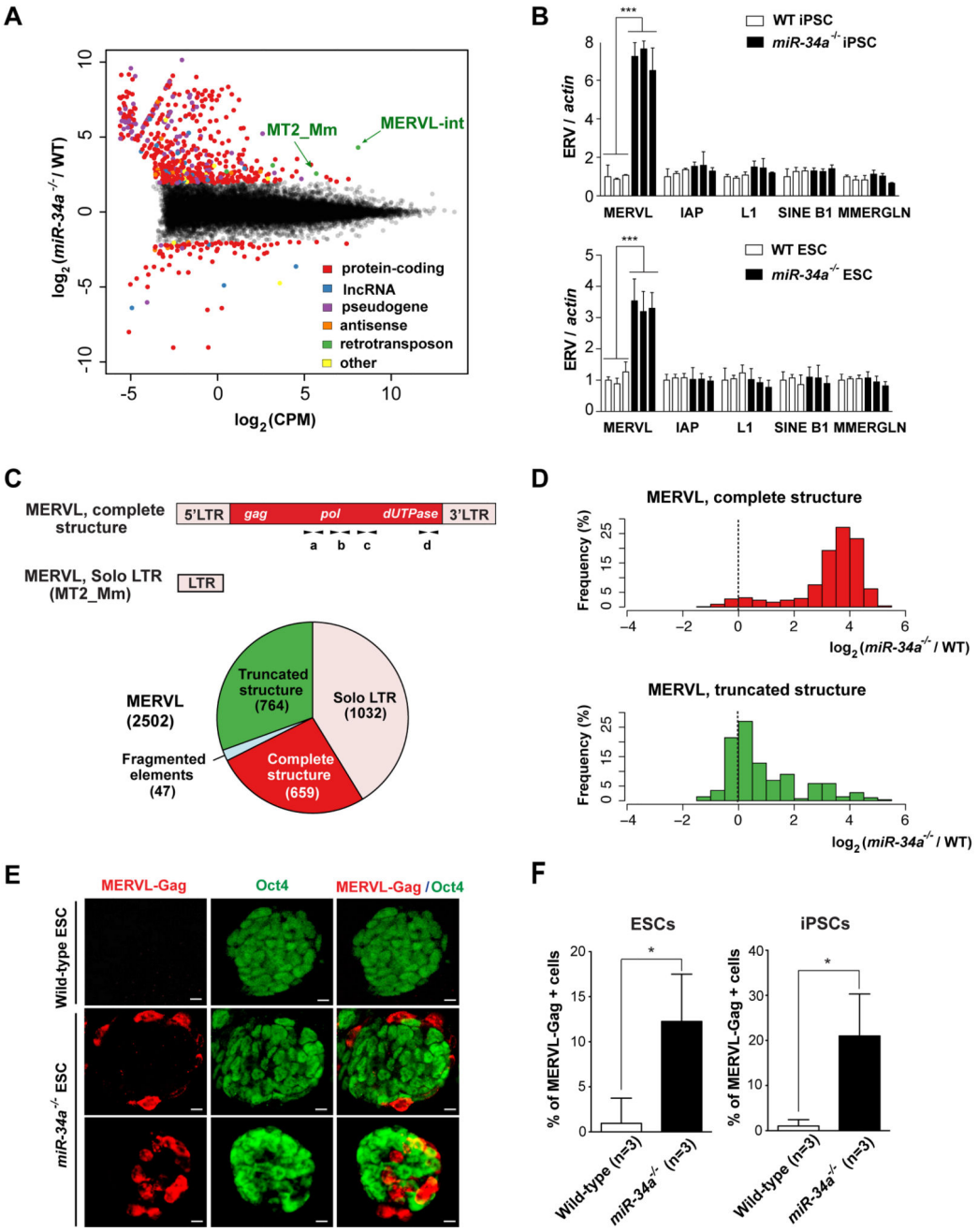


Fig 1. *miR-34a*$^{-/-}$ pluripotent stem cells exhibit expanded cell fate potential

A. *miR-34a*$^{-/-}$ teratomas contain extra-embryonic cell lineages and extra-embryonic cell markers. Teratomas generated from *miR-34a*$^{-/-}$ iPSCs and *miR-34a*$^{-/-}$ ESCs contain cells with the typical placental trophoblast giant cell morphology (black arrows) and placental lactogen 1 (PL-1) expression. Asterisks: the blood-filled lacunae associated with placenta giant cell-like cells. Scale bars, 50 μm. **B**, **C**. *miR-34a*$^{-/-}$ embryoid bodies (EBs) exhibit an induction of both embryonic and extra-embryonic cell markers in immunofluorescence (IF) staining (**B**) and real-time PCR analyses (**C**). **B**. *miR-34a*$^{-/-}$ EBs yield a greater percentage of Cdx2-positive EBs in IF staining and exhibit an induction of the TE marker Cdx2

primarily in cells at the periphery. Scale bars, 100 μm. Error bars: *s.d.*, n=5–7 (randomly selected 10× fields). **C**. *miR-34a*$^{-/-}$ EBs showed an increase in TE markers *cdx2*, *elf5*, *esx1*, *tfap2c* and *gata3*. Three independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ ESC lines were compared. Error bars, *s.d.*, n=3. **D–G.** *miR-34a*$^{-/-}$ ESCs contribute to both embryonic and extra-embryonic cell lineages in chimeric assays *in vivo*. **D**. Four GFP-labeled wild-type or *miR-34a*$^{-/-}$ ESCs were microinjected into each C57BL/6N recipient morula, and the contribution of their progenies to the inner cell mass (ICM) and the trophectoderm (TE) were determined by the localization of GFP-positive cells (left). Scale bar, 20 μm. The percentage of chimeric blastocyst embryos with ESC contribution to the ICM, the TE, and ICM+TE were measured for both wild-type and *miR-34a*$^{-/-}$ ESCs (right). Two independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ ESCs were compared. n, the number of chimeric embryos obtained for each ESC line from three independent injections (Table S1). Two independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ ESC lines were compared. **E**. Single GFP-labeled *miR-34a*$^{-/-}$ ESCs are able to contribute to both ICM and TE (white arrows) of chimeric blastocysts. Representative images were shown for two chimeric blastocysts (top). Scale bar, 20 μm. The percentage of chimeric embryos with ESC contribution to the ICM, the TE, and ICM+TE were quantified (bottom). Two independent *miR-34a*$^{-/-}$ ESC lines were examined. n, the number of chimeric blastocyst embryos for each ESC line from three independent injections for each line. All *P*-values were calculated on a basis of a two-tailed student's *t*-test. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. Two independent *miR-34a*$^{-/-}$ ESC lines were examined. **F, G**. *miR-34a*$^{-/-}$ ESCs contribute to multiple differentiated cell lineages in embryo, yolk sac and placenta in chimeric analyses *in vivo*. 10–15 GFP-labeled wild-type or *miR-34a*$^{-/-}$ ESCs were microinjected into C57BL/6N blastocysts or aggregated with CD1 recipient morulae to generate chimeric embryos. GFP-labeled *miR-34a*$^{-/-}$ ESCs contributed to the visceral endoderm derivatives of the yolk sac (**F**, scale bar, 50 μm), and multiple trophoblast lineages of the placenta (trophoblast giant cells, spongiotrophoblasts, syncytiotrophoblasts (STBs) and sinusoidal trophoblast giant cells (s-TGCs) (**G**, scale bar, 50 μm) of the chimeric embryos at E12.5 and E14.5. **F**. A diagram illustrates the yolk sac tissue architecture and major cell types (top). GFP-positive visceral endoderm cells are identified based on the bilaminar structure of the yolk sac and their characteristic columnar epithelial morphology (bottom). **G**. A diagram illustrates the placenta tissue architecture and major cell types (top). (Bottom) GFP-positive trophoblast giant cells (white arrows) and s-TGCs (red arrows) are identified based on their specific distribution in placenta and their unique cell morphology; GFP-positive spongiotrophoblasts or STBs (white arrows) are identified based on the IF co-staining with trophoblast specific protein alpha (Tpbpa) or ferroportin (MTP1). fbc, nucleated fetal blood cells; f, fetal blood; m, maternal blood. The statistic summary of multiple passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ ESC lines were summarized in Table S1.



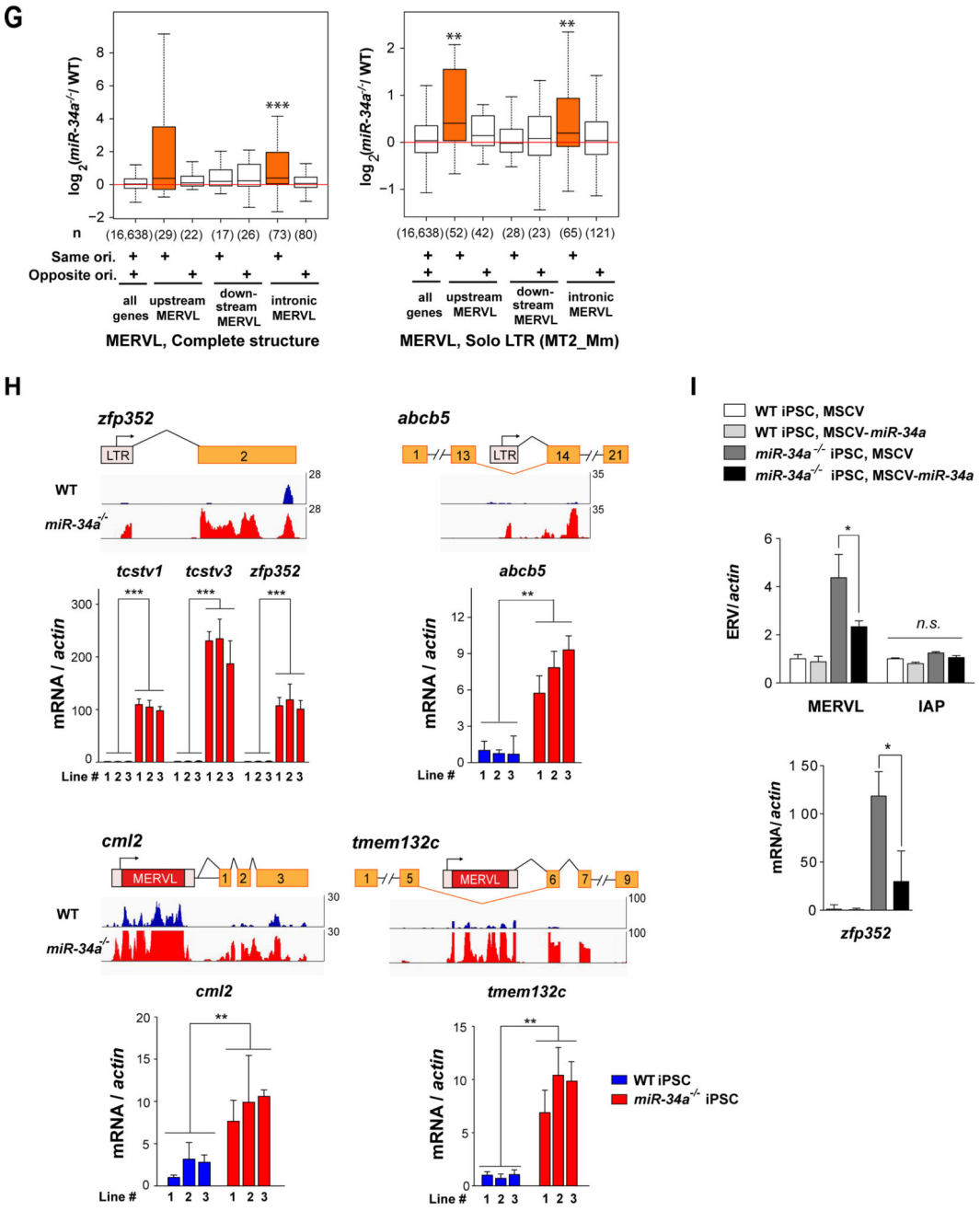
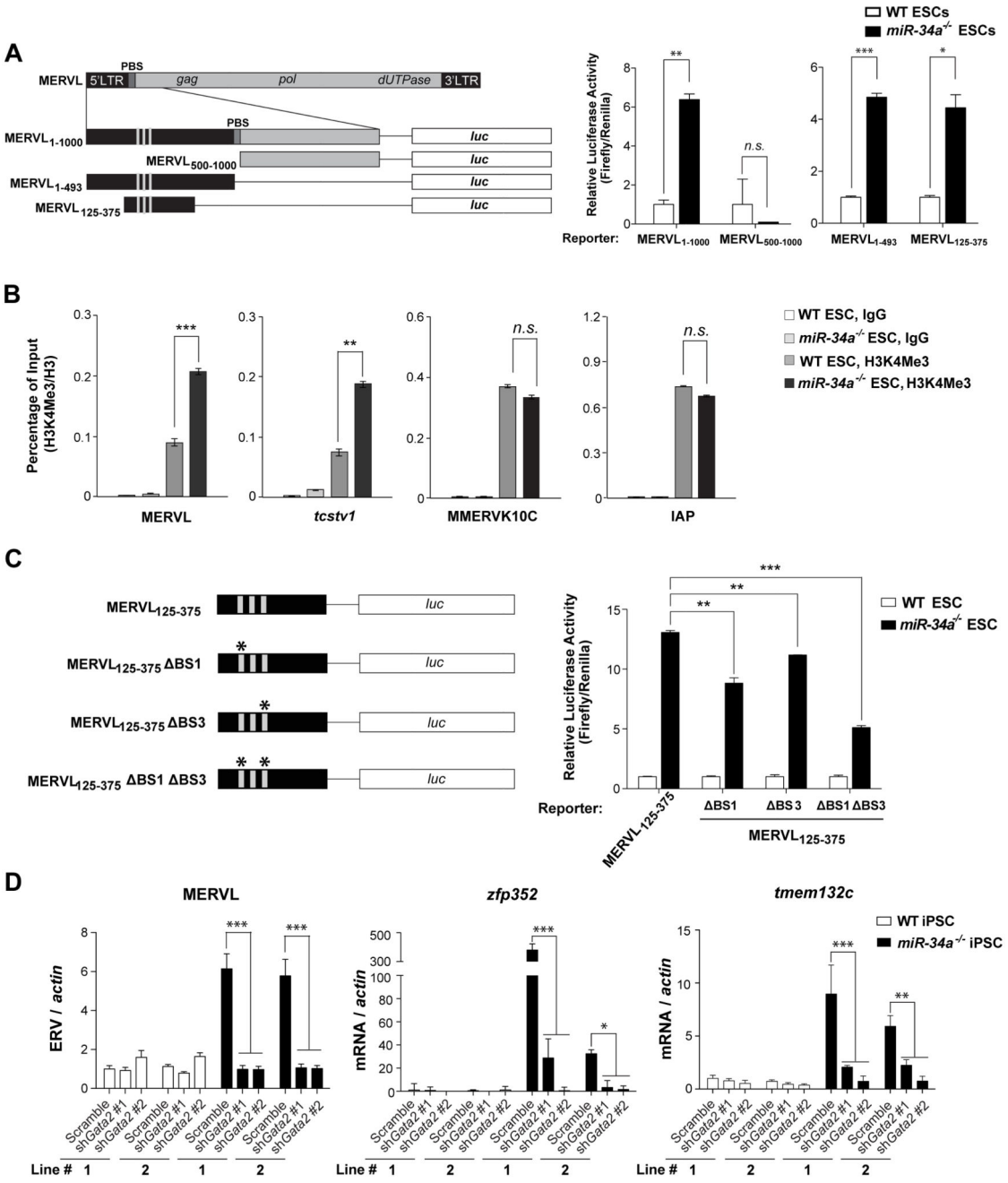


Fig. 2. *miR-34a*$^{-/-}$ pluripotent stem cells exhibit specific induction of the MERVL ERVs

A. The MERVL ERVs are highly induced and differentially expressed (DE) transcriptional unit in *miR-34a*$^{-/-}$ iPSCs compared to wild-type iPSCs. An MA-plot compares the transcription profiles of *miR-34a*$^{-/-}$ and wild-type (WT) iPSCs using RNA-seq data. DE transcriptional units, including protein-coding genes, long non-coding RNAs (lncRNAs), pseudogenes, antisense transcripts, and retrotransposons, are color-coded by class. The 441 DE transcription units between *miR-34a−/−* and wild-type iPSCs (False Discovery Rate 5%, absolute log2 fold-change ≥ 2) include 352 protein-coding genes, 54 pseudogenes, 13 lncRNAs, 6 antisense transcripts, 4 retrotransposon families, and 12 others. Among these

DE genes, MERVL-int (the MERVL internal sequence that encodes *gag, pol* and *dUTPase*) and MT2_Mm (the solo-LTR of the canonical MERVL) are strongly induced and highly expressed in *miR-34a*$^{-/-}$ iPSCs. RNA-seq data were generated from three pairs of independently derived, passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ iPSCs. CPM: counts per million. **B**. MERVL ERVs are specifically induced in *miR-34a*$^{-/-}$ iPSCs and ESCs, as measured by real-time PCR analyses. All other ERVs tested, including intracisternal A-particle (IAP), LINE L1, SINE B1, and MMERGLN, exhibited no significant difference between wild-type and *miR-34a*$^{-/-}$ pluripotent stem cells. Three independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ iPSC and ESC lines were compared. Error bars: *s.d.*, n=3. **C**. A schematic diagram illustrates the transcript structure of the MERVL family of ERVs, with the position of four pairs of validated real-time PCR primers indicated (a, b, c and d) (top). (Bottom) A pie chart shows the relative abundance for the different MERVL subclasses, categorized according to their structural features. MERVL loci with a complete structure (red) carry both 5' and 3' LTRs, along with an internal sequence (annotated by Repeat Masker as MERVL-int); truncated MERVL elements (green) lack one or both LTRs; and solo- LTRs of canonical MERVL (pink), also designated as MT2_Mm, are generated through homologous recombination during evolution. **D**. The MERVL induction in *miR-34a*$^{-/-}$ iPSCs occurs primarily in loci with a complete structure. Histograms are shown for the $\log_2$ fold-change of the expressed MERVL loci between *miR-34a*$^{-/-}$ and wild-type iPSCs, either with a complete structure (top) or with a truncated structure (bottom). **E**. The MERVL-Gag and Oct4 expression is mutually exclusive in a subset of *miR-34a*$^{-/-}$ ESCs, while MERVL-Gag staining is absent in wild-type ESCs. Scale bars, 20 μm. **F**. The percentage of MERVL-Gag positive cells were quantified in early passage of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ ESCs (left) and iPSCs (right). Error bars: *s.d.*, n=5–6 (randomly selected 10× fields). **E–F**. Three independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ iPSC and ESC lines were compared. **G–I**. The MERVL activation in *miR-34a*$^{-/-}$ pluripotent stem cells induces many MERVL proximal gene isoforms. **G**. MERVL proximal genes with an upstream or intronic MERVL element on the same strand are preferentially up-regulated in *miR-34a*$^{-/-}$ iPSCs. Boxplots of log2 fold-change between *miR-34a*$^{-/-}$ and wild-type iPSCs are shown for genes proximal to annotated, complete MERVL loci (left) or MERVL solo LTR (MT2_Mm) loci (right). The numbers in parentheses indicate the number of protein-coding genes in each category; the box plot of the log2 fold-change of all protein-coding genes is included for reference. Same ori. (same orientation): MERVL and its proximal gene are on the same strand; opposite ori. (opposite orientation): MERVL and its proximal gene are on opposite strands. Genes harboring a complete MERVL copy in their introns on the same strand, as well as genes with MT2_Mm on the same strand either upstream or in their introns, have significantly larger fold-changes than the rest of the genes. **, $P < 0.01$; *** $P < 0.001$; Wilcoxon-Mann-Whitney test. **H**. Examples are shown for induced expression and altered transcript structure of MERVL proximal genes in *miR-34a*$^{-/-}$ iPSCs. The MT2B1 solo LTR elements upstream of z*fp352, tcstv1 or tcstv3* act as promoters to strongly induce their expression in *miR-34a*$^{-/-}$ iPSCs. Similarly, a complete MERVL element upstream of *cml2* acts as an alternative promoter to induce the expression of an MERVL-*cml2* isoform in *miR-34a*$^{-/-}$ iPSCs. Intronicly localized MERVLs, either a solo LTR (as that in intron 13 of *abcb5*) or a complete ERV element (as that in intron 5 of

tmem132c), also act as alternative promoters to drive the expression of truncated gene isoforms that contain only the downstream exons. Error bars: *s.d.*, n=3, *** $P < 0.001$. **I**. *miR-34a* overexpression in *miR-34a*$^{-/-}$ iPSCs using a MSCV retroviral vector significantly suppresses the level of MERVL and the MERVL-*zfp352* chimeric transcript in real-time PCR analysis, but causes no alteration in the level of IAP. One pair of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ iPSCs were examined. Error bars: *s.d.*, n=3. All *P*-values were calculated on a basis of a two-tailed Student's *t*-test unless stated otherwise. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.



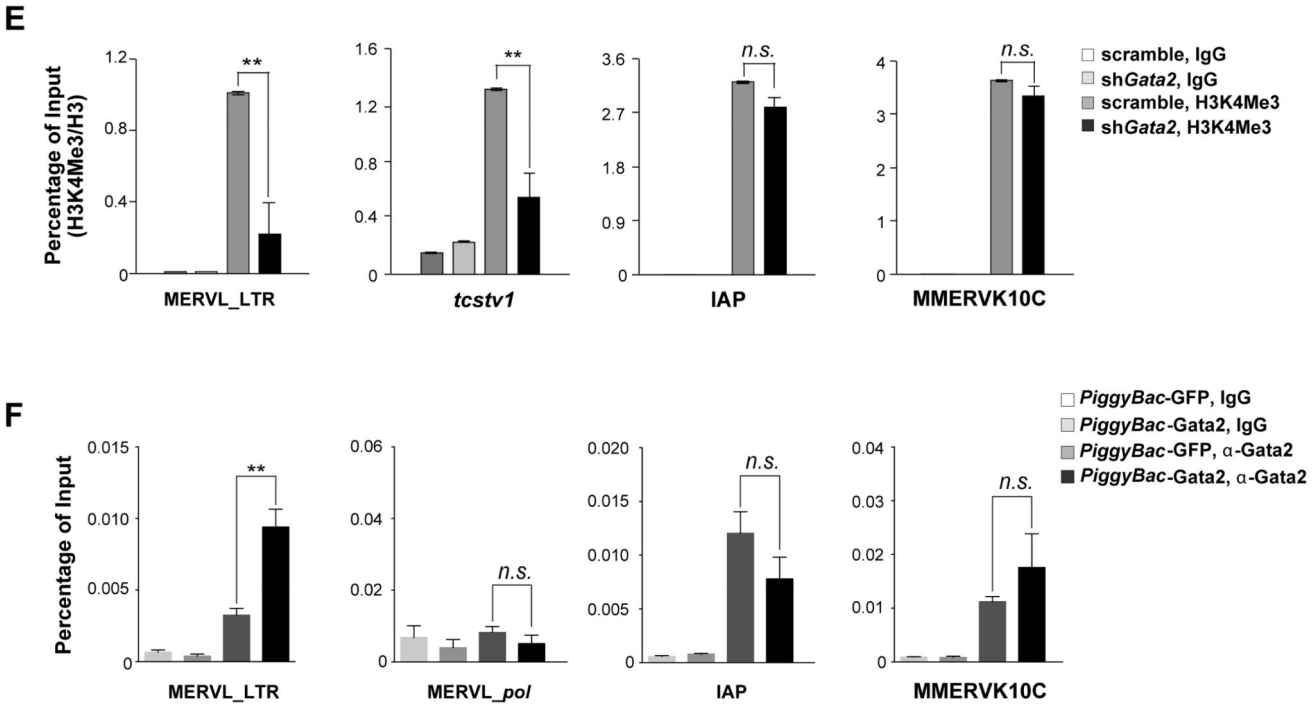
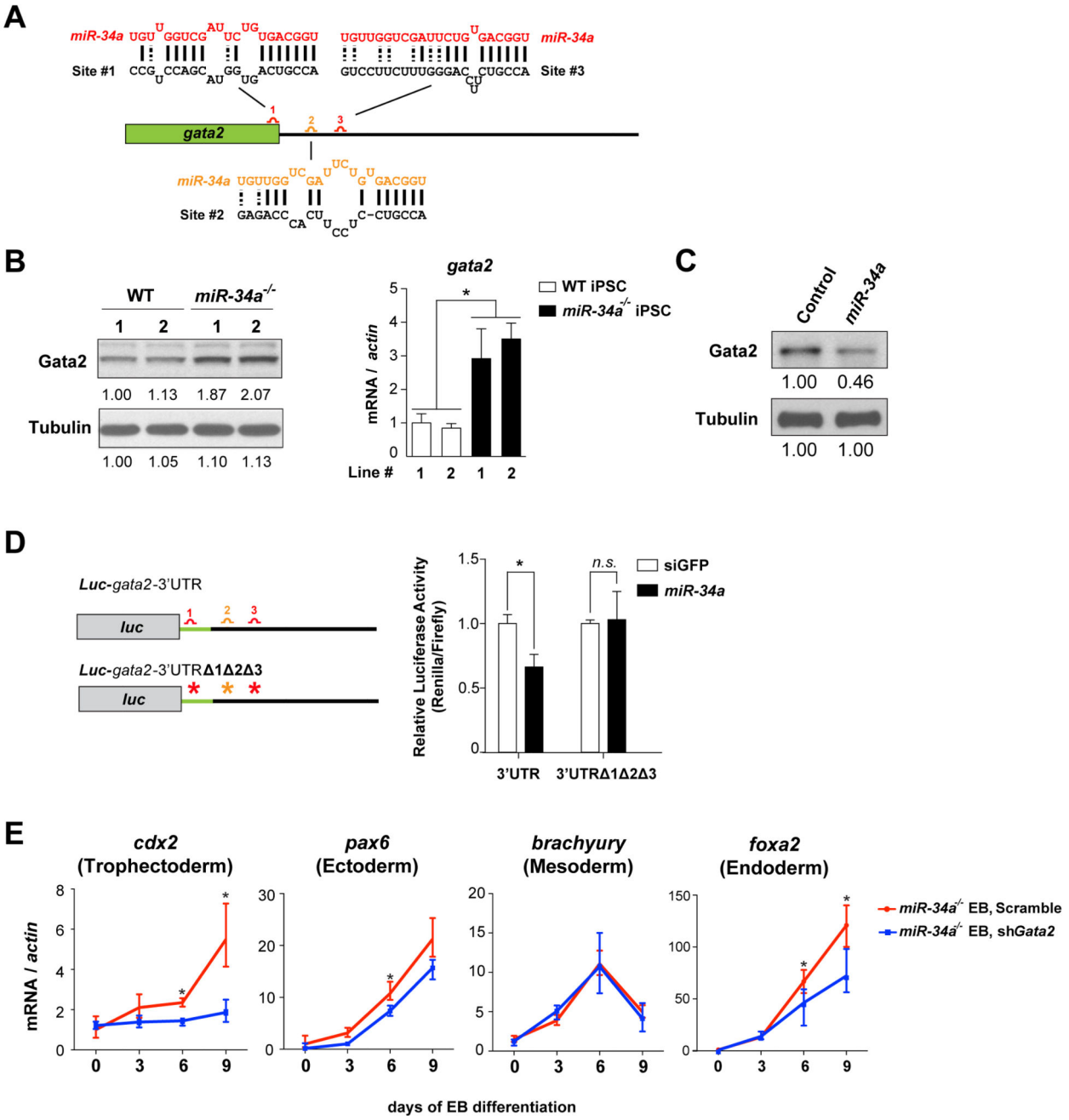


Fig.3. Gata2 is essential for the MERVL induction in *miR-34a*$^{-/-}$ pluripotent stem cells

A. A full length or a fragment of MERVL LTR can be activated specifically in *miR-34a*$^{-/-}$ESCs. (Left) A schematic diagram shows the MERVL LTR and the truncated fragments that were tested for promoter activity using luciferase (Luc) assays. (Right) The Luc reporters driven by MERVL fragments containing the full length LTR (MERVL$_{1–1000}$-Luc and MERVL$_{1–493}$-Luc) exhibit a strong activation in *miR-34a*$^{-/-}$ ESCs, but not in wild-type ESCs. A Luc reporter containing the truncated MERVL$_{125–375}$ fragment completely recapitulates this differential reporter activity in wild-type and *miR-34a*$^{-/-}$ ESCs. Error bars: *s.d.*, n=2. PBS: primer binding site. **B**. Chromatin ImmunoPrecipitation (ChIP) reveals an increased H3K4Me3 modification on the MERVL LTR and the MERVL-*tcstv1* chimeric gene in *miR-34a*$^{-/-}$ ESCs. In comparison, the H3K4Me3 level is unaltered on IAP LTR and MMERVK10C LTR. **C**. Mutations of two predicted Gata2 binding sites (BS1 and BS3) in the MERVL$_{125–375}$-Luc reporter synergistically impair the reporter activity in *miR-34a*$^{-/-}$ ESCs. (Left) A schematic diagram shows the MERVL$_{125–375}$-Luc reporter and the mutated derivatives. (Right) While the mutation of BS1 or BS3 alone in the MERVL$_{125–375}$-Luc reporter modestly impairs its activity in *miR-34a*$^{-/-}$ ESCs, mutations of both BS1 and BS3 significantly reduces the reporter activity. Error bars: *s.d.*, n=2. **D**. *gata2* knockdown significantly decreases the expression of MERVL and MERVL proximal genes. Using two independent shRNAs targeting *gata2*, we effectively knocked down *gata2* in *miR-34a*$^{-/-}$ iPSCs using RNA interference (RNAi) (also see Fig. S5C), and observed a decreased expression of MERVL and MERVL proximal genes (*zfp352* and *tmem132c*). **E**. *gata2* knockdown in *miR-34a−/−* iPSCs significantly decreases the H3K4Me3 deposition on MERVL elements and on the specific MERVL element upstream of *tcstv1*, but has no effects on H3K4Me3 deposition on IAP or MMERVK10C. **F**. ChIP reveals an increase of Gata2 binding to the MERVL LTR region upon Gata2 overexpression in *miR-34a*$^{-/-}$ iPSCs. Gata2

binding to MERVL *pol*, IAP LTR, or MMERVK10C LTR is unaltered upon Gata2 overexpression. **C**, **E–F**. Two independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ iPSC lines were compared. 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.



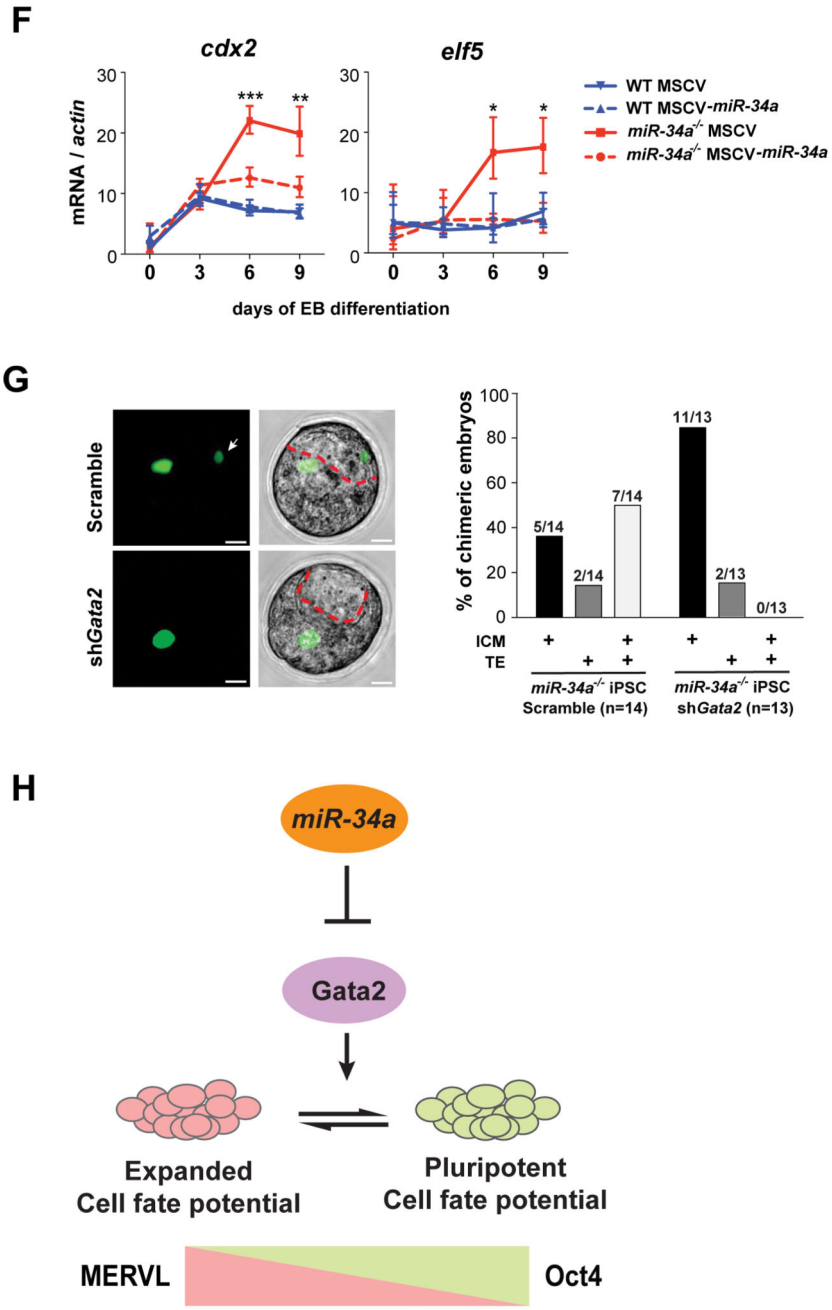


Fig. 4. *miR-34a* restricts cell fate potential of pluripotent stem cells by targeting *gata2*.

A. A schematic representation of three predicted *miR-34a* binding sites in the *gata2* mRNA, with one site (1) located at the 3' end of the open reading frame (ORF) and two sites (2 and 3) located within the 3'UTR. Site 1 (red) is predicted as a strong *miR-34a* binding site by both duplex folding energy and the 7mer-A1 seed-match rule (28, 29). While site 3 (red) does not have a perfect seed sequence, it contains a compensatory 3' base-pairing and exhibits a strong folding energy. In contrast, site 2 (orange) represents a weaker prediction (28). **B**. Gata2 exhibits *miR-34a*-dependent repression in *miR-34a*$^{-/-}$ pluripotent stem cells. Gata2 protein (left) and *gata2* mRNA (right) are elevated in *miR-34a*$^{-/-}$ iPSCs as compared

with wild-type iPSCs. Two independent pairs of passage- and littermate-controlled wild-type and *miR-34a*$^{-/-}$ iPSC lines were measured by Western blotting and real-time PCR analyses. Error bars: *s.d.*, n=3, * $P < 0.05$. **C**. Overexpression of *miR-34a* in *miR-34a*$^{-/-}$ iPSCs represses the Gata2 protein level. The quantification of western blot analyses was performed by ImageJ. **D**. Mutating all three predicted *miR-34a* binding sites within the Luc-gata2-3UTR luciferase reporter completely abolishes its *miR-34a*-dependent repression. Error bars: *s.d.*, n=2. **E**. *gata2* knockdown in *miR-34a*$^{-/-}$ iPSCs by RNAi abolishes the induction of the TE marker *cdx2* during EB differentiation, but has no effects on the induction of ectoderm (*pax6*), mesoderm (*brachyury*), or endoderm (*foxa2*) markers. Error bars: *s.d.*, n=3. **F**. *miR-34a* overexpression in *miR-34a*$^{-/-}$ iPSCs significantly suppresses the induction of TE markers *cdx2* and *elf5* upon EB differentiation. Error bars: *s.d.*, n=3. **E–F**. Results shown are representative of two independent experiments using the same *miR-34a*$^{-/-}$ iPSC line. **G**. The expanded cell fate potential of *miR-34a*$^{-/-}$ iPSCs is restricted by *gata2* knockdown in chimeric analyses. We injected four GFP-labeled cells into each recipient morula. While 50% of chimera blastocysts generated from control infected *miR-34a*$^{-/-}$ iPSCs contain iPSC contribution to both ICM and TE (n=7/14), *miR-34a*$^{-/-}$ iPSCs with *gata2* knockdown lost this expanded cell fate potential (n=0/13), and primarily contributed to the ICM (n=11/13). All *P*-values were calculated on a basis of a two-tailed Student's *t*-test. * $P < 0.05$, ** $P < 0.01$. **H**. A diagram shows our proposed model in which the *miR-34a*/gata2 pathway restricts the pluripotent cell fate potential of pluripotent stem cells.