Published in final edited form as:
Nat Genet. 2018 January ; 50(1): 106–119. doi:10.1038/s41588-017-0016-5.

A molecular roadmap for the emergence of early-embryonic-like cells in culture

Diego Rodriguez-Terrones[#1,2], **Xavier Gaume**[#1,2], **Takashi Ishiuchi**[1,3], **Amélie Weiss**[2], **Arnaud Kopp**[2], **Kai Kruse**[4], **Audrey Penning**[1], **Juan M. Vaquerizas**[4], **Laurent Brino**[2], and **Maria-Elena Torres-Padilla**[1,5]

[1]Institute of Epigenetics and Stem Cells (IES), Helmholtz Zentrum München D-81377 München, Germany

[2]Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM U964, U de S, F-67404 Illkirch, CU de Strasbourg, France

[4]Max Planck Institute for Molecular Biomedicine, Münster, Germany

[5]Faculty of Biology, Ludwig Maximiliams Universität, München, Germany

[#] These authors contributed equally to this work.

Abstract

Unlike pluripotent cells, which generate only embryonic tissues, totipotent cells can generate a full organism, including extraembryonic annexes. A rare population of cells resembling 2-cell stage embryos arises in pluripotent embryonic stem (ES) cell cultures. These 2-cell-like-cells display molecular features of totipotency and broader developmental plasticity. However, their specific nature and the process through which they arise remain outstanding questions. Here, we identify intermediate cellular states and molecular determinants during the emergence of 2-cell-like-cells. By deploying a quantitative single cell expression approach, we identified an intermediate population characterised by the expression of the transcription factor ZSCAN4 as precursor of 2-cell-like-cells. Using an siRNA screening, we uncovered novel epigenetic regulators of 2-cell-like-cell emergence, including the non-canonical PRC1 complex PRC1.6 and Ep400/Tip60. Our data shed light on the mechanisms underlying the exit from the ES cell state towards the formation of early-embryonic-like cells in culture and identify key epigenetic pathways that promote this transition.

[*]Corresponding author: torres-padilla@helmholtz-muenchen.de, Tel. + 49(0) 89-3187-3317, Fax. + 49(0) 89-3187-3389.
[3]Current address: Division of Epigenetics and Development, Medical Institute of Bioregulation, Kyushu University, 812-8582 Fukuoka, Japan

Author Contributions

D.R.-T., X. G. and T. I. designed, performed and analysed experiments. D.R.-T. performed most computational analyses. A.W. performed the screening together with X.G. under the supervision of L.B. A.K. implemented the screening analysis pipeline with L.B. K.K performed bioinformatics analysis under the supervision of J.M.V. A. P. performed experiments for screening validation. M.E.T-P. designed and supervised the study. All authors contributed to the manuscript preparation, read, commented and approved the manuscript.

The authors declare no competing financial interest.

Introduction

Cellular plasticity, the ability to give rise to different cellular fates, is essential for multicellularity. Multicellular organisms derive ultimately from one single cell, the one cell embryo, which forms at fertilization and has the capacity to generate a full organism. This capacity is referred to as totipotency[1–3]. Pluripotency, on the other hand, emerges later in development and relates to the ability to form all germ layers of the embryo proper, but not the extra-embryonic annexes[4]. In the mouse, only the zygote and individual 2-cell stage blastomeres are, strictly speaking, totipotent, as they can generate a full organism on their own[5–7].

Pluripotent embryonic stem cells (ESCs) derived from the inner cell mass of the blastocyst[8,9] recapitulate some molecular features of the pre-implantation epiblast, including expression of transcription factors such as NANOG and OCT4[10–12]. ES cell cultures are heterogeneous, with subpopulations of cells differing in gene expression in a dynamic equilibrium[13–18]. Much of this heterogeneity results from changes in expression of pluripotency-associated transcription factors (TFs), which are part of the core regulatory network of ESCs[19]. In addition, the developmental potential of ESCs grown under different conditions is not equivalent[20–23].

Unlike pluripotency, the molecular features of totipotency remain largely unknown. Cells resembling 2-cell stage embryos arise spontaneously in ES cell cultures, constituting less than 1% of the culture [24]. These 2-cell-like-cells display a transcriptome highly similar to 2-cell stage embryos, including 2-cell-stage specific genes like *Zscan4* and the MERVL retrotransposon[24,25], and have distinctive molecular features compared to ESCs, including downregulation of OCT4 protein[24], higher histone mobility[26] and dispersed chromocentres[27]. Two-cell-like cells seem to have higher developmental plasticity and higher nuclear reprogrammability than ESCs[24,27].

The molecular regulatory networks underpinning the molecular identity and emergence of 2-cell-like-cells have not been established. Also, it is unclear whether they can self-renew, but given their higher plasticity, knowledge of the underlying biology of the mouse 2-cell-like-cells could potentially be applied to expand the potency of existing human pluripotent cells.

Results

Single cell analysis reconstructs the transition to the 2-cell-like state

To shed light into the molecular mechanisms underlying the emergence of embryonic-like features, we set out to identify intermediate steps in the transition from ES to 2-cell-like-cells. We used transcriptional profiling at the single cell level, which can reveal dynamic cellular states, thereby identifying cell fate determinants[28,29]. We focused mainly on chromatin modifiers and transcription factors, which generally display low expression levels. Because of the crucial need for highly precise and sensitive gene expression measurements, we used a qPCR-based microfluidics approach based on the Biomark Fluidigm platform as opposed to a poly-A-based RNA-seq method[30,31].

We performed quantitative gene expression analysis in single cells using an ES cell line containing the *2C::tbGFP* reporter, driving turboGFP expression under the MERVL promoter, which recapitulates the 2-cell-like state27. We identified a representative set of genes to profile based on bulk RNA-seq data comparing ESCs versus endogenous and CAF-1 KD-induced 2-cell-like-cells27. We selected 93 genes based on their differential expression between ES and 2-cell-like-cells and on their functional significance (Fig. 1a, Supplementary Fig. 1a, Supplementary Table 1). Through an experimental setup designed to enrich for cells at an early stage of the transition towards the 2-cell-like state (Fig. 1b and Supplementary Fig. 1b and c) we analysed gene expression in 136 individual tbGFP- and tbGFP+ cells (Supplementary Table 2). Two spike-in controls allowed us to assess technical noise (Supplementary Fig. 1d). *tbGFP* mRNA expression occurred in those cells selected as tbGFP-positive based on their fluorescence, but not in cells lacking tbGFP (Fig. 1c). Two-cell-like cells (tbGFP+) maintained mRNA levels of *Oct4/Pou5f1* and displayed high expression of *Zscan4* and of the *MT2_mm* and *MT2B/C* LTR-driven chimeric transcripts *Spz1*, *Naalad2* and *Sp110*, thereby validating our dataset (Fig. 1c).

To identify transitional cellular states, we performed Principal Component Analysis (PCA), which revealed three main sources of variability in the data. Principal component 1 (PC1) separates tbGFP- from tbGFP+ cells, and therefore defines the ES to 2-cell-like-cell transition (Fig. 1d-e and Supplementary Fig. 1e). Expression of *Zscan4* and of chimeric transcripts, in addition to that of *tbGFP*, provides the strongest identity to PC1 (Fig. 1f-g). PC2 revealed heterogeneity within the 2-cell-like-cell population (Fig. 1d-e). For example, 2-cell-like-cells display graded expression of *Id1*, *Id2* and *Id3* (Fig. 1f). The PC3 contrasts naïve versus primed pluripotency (Fig. 1g)(Supplementary Fig. 1e)32–34. Importantly, computing the same PCA but excluding the *tbGFP* expression produced the same results (not shown), suggesting that the dispersion of the data is inherent to the global expression profile of individual cells, and is not determined by the expression of the *2C::tbGFP* reporter. Thus, our single cell analysis allows us to model the acquisition of the 2-cell like state.

Zscan4 positive cells exhibit an intermediate expression profile between ES and 2-cell-like-cells

We next interrogated the single cell dataset for transcripts showing graded expression between the ES and 2-cell-like-cells. The PCA identified a set of cells with an intermediate expression profile located between the two clusters of cells, along PC1 (Fig. 1e, black arrows). These cells lacked expression of *2C::tbGFP* (Fig. 1d-e), indicating that they had not yet entered the 2-cell-like state. However, they express the transcription factor *Zscan4c/d/f* (Fig. 1h-i). The intermediate clustering of these cells was maintained even when omitting expression data of *Zscan4c/d/f* and *tbGFP* (Supplementary Fig. 1f, g). Thus, while *Zscan4c/d/f* expression delineates an intermediate cellular state, this state is not solely determined by changes in *Zscan4c/d/f* expression, suggesting that cells undergo larger transcriptional reprogramming concomitantly with *Zscan4c/d/f* induction, but prior to the acquisition of a 2-cell-like identity.

Two-cell-like cells arise primarily from Zscan4 positive cells

We next addressed whether *Zscan4c/d/f*-positive (*Zscan4+* hereafter) cells constitute an intermediate state during the transition from ES to 2-cell like cells. While knock-down of ZSCAN4 in ESCs results in genome instability35, its ectopic overexpression can induce a limited part of the 2-cell-specific transcriptome36,37. We reasoned that if 2-cell-like-cells arise from *Zscan4+* cells, we could formulate four predictions. Firstly, profiling of single cells expressing *Zscan4* but not *2C::tbGFP* should reveal a cluster of cells occupying an intermediate position along the original PC1. Indeed, additional single cell expression profiling of *Zscan4+* cells sorted based on *Zscan4*::tdTomato fluorescence (Fig. 2a, Supplementary Fig. 2a and Supplementary Note) and projection of the data into the previous dataset indicated that *Zscan4+* cells (but tbGFP negative) localized in the middle along the PC1 (n=189 cells; Fig. 2b, red dots). This intermediate character persisted when computing both datasets together and when data for *tbGFP* and *Zscan4* was excluded from the analysis (Supplementary Fig. 2b-c). Thus, *Zscan4* is a marker of an intermediate cell population between the ES and 2-cell-like state, and *Zscan4*+ cells are not necessarily 2-cell-like-cells. Secondly, we reasoned that the majority – if not all– 2-cell-like-cells should also express *Zscan4*. We found that all 2-cell-like-cells co-express endogenous ZSCAN4 protein (Fig. 2c, d and Supplementary Note) and mRNA (Fig. 2d), in line with a recent report38. Thus, the 2-cell-like-cell population is contained within the *Zscan4+* population. Thirdly, we asked whether *Zscan4+* cells convert more frequently than *Zscan4-* cells to 2-cell-like-cells. Using a destabilized *Zscan4::mCherry* reporter (Fig. 2e, Supplementary Fig. 2d-e), we found that *Zscan4+* cells give rise to significantly more (~9 times) 2-cell-like-cells, compared to *Zscan4-* cells (4% ± 2.6 versus 0.5% ± 0.13, p=0.04)(Supplementary Fig. 2f-g). Fourthly, we performed time-lapse microscopy (Fig. 2e, Supplementary Fig. 2d-e and Supplementary Note). We imaged 383 emerging 2-cell-like-cells and tracked them individually, which demonstrated that the majority (81%; 312/383) emerged from *Zscan4*+ cells (Fig. 2f and Supplementary Videos 1-3). The remaining ~19% (71/383) appeared directly from *Zscan4*- cells (Fig. 2f), suggesting that 2-cell-like-cells may arise through different pathways. While we cannot formally rule out that these cells also derive from *Zscan4+* cells because of incomplete *Zscan4::mCherry* reporter penetrance (Supplementary Fig. 2e), our results suggest that 2-cell-like-cells arise primarily from *Zscan4+* precursors, thereby identifying *Zscan4* transcription as the first molecular marker in the transition from ES toward 2-cell like cells.

ATAC-seq analysis38 further demonstrated that *Zscan4*+ cells display an intermediate pattern of chromatin accessibility between ES and 2-cell-like-cells (Supplementary Fig. 2h). Importantly, ATAC-seq38 also revealed that while 2-cell-like-cells display an open chromatin at MERVL sites, *Zscan4*+ cells do not (Fig. 2g-h). This suggests that the chromatin landscape of *Zscan4*+ cells differs from that of 2-cell-like-cells and that *Zscan4* activation precedes chromatin opening at MERVL.

Next, we investigated the dynamics of *Zscan4* expression. ZSCAN4 is encoded by 6 genes (*a* through *f*)25. Our single cell dataset revealed that three distinct cell clusters are defined according to *Zscan4* transcript abundance: low, medium and high (Supplementary Fig. 3a). Their position along the PC1, from the lowest to the highest mRNA content (Supplementary

Fig. 3b-c), suggests a gradual activation of *Zscan4* upon exit towards the 2-cell-like state (Supplementary Fig. 3a and Supplementary Note). We classified cells into 5 groups based on their combined expression of *2C::tbGFP* and *Zscan4*: i) ESCs, ii) Zscan4-low (tbGFP-negative), iii) Zscan4-medium (tbGFP-negative), iv) Zscan4-high (tbGFP-negative); and v) 2-cell-like-cells (tbGFP-positive). Transcript levels of *Oct4* did not vary across the five groups (Fig. 2i; note Ct value; and Supplementary Table 3). Zscan4-low cells showed no difference in expression of chimeric transcripts compared to ESCs, but chimeric transcripts showed a bimodal distribution in the Zscan4-mid and high intermediates (Fig. 2i). This suggests that LTRs are activated in different subpopulations of cellular intermediates, pointing towards a sequential model of transcriptional changes whereby *Zscan4* activation precedes activation of chimeric LTRs and MERVL itself, in line with the ATAC-seq analysis (Fig. 2g-h).

To address whether our conclusions above reflect sequential transcriptional changes genome-wide, we interrogated available single cell RNA-seq datasets38. We modelled single cell RNA-seq data according to their abundance of *Zscan4* transcripts, which confirmed the presence of subpopulations of cells expressing various levels of *Zscan4*, which we classified into ES, Zscan4-low, -mid and –high, and 2-cell-like-cells (Supplementary Fig. 3d-e). Differential expression analysis revealed unique transcriptional changes between each transitional state and confirmed a step-wise transition from ESCs towards the 2-cell-like state globally, with the Zscan4+ intermediates showing an intermediate expression profile (Supplementary Fig. 3f, g).

Zscan4+ cells downregulate protein levels of pluripotency factors

ES cell identity relies primarily on a gene regulatory network in which OCT4, NANOG and SOX2 are essential components39. Two-cell-like cells downregulate protein levels of OCT424,27. To dissect the mechanism through which ESCs transition towards the 2-cell-like state, we determined at which stage of the transition the down-regulation of pluripotency-associated TFs occurs. We addressed whether other pluripotency-associated factors are also affected in 2-cell-like-cells. OCT4 was downregulated in 2-cell-like-cells and also in *Zscan4*+ intermediates (Fig. 3a, f and Supplementary Fig. 4a and Supplementary Note). SOX2, PRDM14 and AP2γ were also down-regulated in *Zscan4*+ cells (Fig. 3b-d, f). REX1 localization changed in 2-cell-like-cells, suggesting an alteration of REX1 function during the emergence of 2-cell-like-cells (Fig. 3e). In addition to changes in protein levels, the single cell expression dataset revealed a gradient in the expression of *Sox2, Prdm14*, *AP2γ* and *Rex1*, but not of *Oct4*, along the PC1 (Fig. 3g). Transcript levels of *Sox2, AP2γ* and *Rex1* were down-regulated in Zscan4-high intermediates, to a similar level to that of 2-cell-like-cells (Fig. 3h and Supplementary Table 3). We conclude that the pluripotency regulatory network is down-regulated in 2-cell-like-cells, in line with an exit from the ES state, but that changes in the levels of pluripotency TFs are already apparent in the Zscan4-positive, but *2C::tbGFP* negative state. Therefore, upregulation of *Zscan4* precedes changes in the pluripotency core regulatory network in the transition towards 2-cell-like-cells.

Exit of ESCs towards the 2-cell-like state is unrelated to differentiation

The results above prompted us to investigate whether 2-cell-like-cells emerge through a similar mechanism to differentiation, since both fates entail the eventual down-regulation of pluripotency TFs. To address whether 2-cell-like-cells exit through an early differentiation intermediate or rather exit independently of differentiation cues, we used a *Rex1* knock-in reporter40, as *Rex1* expression faithfully reflects the earliest exit towards differentiation17,22. Rex1-negative cells, in contrast to Rex1-positive cells, lose clonogenic self-renewal capacity when plated in serum21. In addition, Rex1-low and Rex1-negative cells differentiate more efficiently than Rex1-positive cells17,22. We reasoned that if 2-cell-like-cells emerge from differentiating ESCs, Rex1-low cells should generate Zscan4+ cells with higher propensity. Using a double reporter cell line (Fig. 4a and Supplementary Fig. 4b-c), we removed *Zscan4*+ cells from serum/LIF grown ESCs and replated Rex1-low and Rex1-high cells separately (Supplementary Fig. 4d). Rex1-low cells showed the expected flattened morphology after 24h of culture (Fig. 4b)17. FACS revealed that Rex1-low cells generate fewer Zscan4+ cells than Rex1-high cells (Fig. 4c, d), in line with the single cell data showing an enrichment of Zscan4+ cells in Rex-1 high cells (Supplementary Fig. 4e, f). To rule out the possibility that Zscan4+ cells within the Rex1-low fraction derive from a population that interconverted to the Rex1-high state 17,21,40 (Fig. 4e), we performed the same experiment as above, but included Rex1-negative cells, which cannot convert to Rex1-positive state21(Fig. 4f, g). This time, cells were plated without LIF because after 24h of LIF withdrawal, Rex1-low cells irreversibly lose ES cell identity and are closer to lineage specification21. FACS confirmed that Rex1-low cells have a decreased ability to generate Zscan4+ cells (Fig. 4h, i). In addition, the number of Zscan4+ cells in the Rex1-negative population was significantly lower than in the Rex1-positive fraction. Thus, Zscan4+ cells – and therefore 2-cell-like-cells – do not emerge from a differentiating precursor. In agreement, siRNA for *Oct4, Nanog, Sox2* or *Rex1* did not affect the percentage of Zscan4+ and *2C::*tbGFP+ cells (Supplementary Fig. 4g-i). We conclude that 2-cell like cells exit the ES cell state through a different path than differentiation.

Defining a molecular roadmap during the emergence of the 2-cell like state

We next sought to define a molecular roadmap for 2-cell-like-cells emergence based on the sequential changes in gene expression across the 5 cellular states described above (Supplementary Fig. 5a-b and Supplementary Table 3). Specific TFs showed significant changes of expression upon 2-cell-like-cell emergence, mostly among the Zscan4+ intermediates (Fig. 5a). For example, the pioneer factor *FoxA1* was sharply upregulated at the Zscan4-low to Zscan4-mid transition. Amongst the 22 chromatin modifiers profiled, a few displayed significant changes accompanying each of the transitions between ES and 2-cell-like-cells (Fig. 5a). The transition with the most changes in the chromatin factors analysed was between the Zscan4-mid and the Zscan4-high states. These observations prompted us to undertake a broad functional analysis to identify epigenetic factors that promote 2-cell-like fate.

Identification of epigenetic pathways regulating 2-cell-like emergence

We performed a screening using an siRNA library targeting 1167 genes41 (Fig. 5b, Supplementary Table 4). Based on our results above, we employed three readouts to identify *bona fide* 2-cell-like-cells: 2C::eGFP expression, ZSCAN4 expression, and loss of OCT4 protein (Fig. 5c). The screening achieved single cell resolution through nuclear segmentation (Supplementary Fig. 6a). We included siRNA against the CAF-1 subunit p150 as positive control27(Fig. 5c and Supplementary Fig. 6b) and applied stringent thresholds based on individual and combined Z-scores for hit selection (Fig. 5d-e and Supplementary Fig. 6c). Some hits led to reduced cell numbers (Supplementary Fig. 6d), suggesting toxicity and/or cell proliferation effects. Chromatin modifiers previously shown to affect 2-cell-like emergence, such as KAP1 and LSD1 were also in our screening; however, their down-regulation had a mild effect on 2-cell-like-cell induction, and a lower Z-score. We extracted the top 50 hits (Fig. 5f) and validated them using a secondary screening (Supplementary Fig. 7a-b). Because top hits of the primary screen include polycomb proteins and histone chaperones, we included additional siRNAs to investigate all known components of these pathways. We validated all top 50 hits, with the exception of *Dnmt3b* (Supplementary Fig. 7c). In addition, siRNA for polycomb-related proteins MGA, MAX and RYBP 42,43 and the histone chaperone DAXX, induced 2-cell-like-cells from 3- to 15-fold (Supplementary Fig. 7d-e). The combined top 50 hits from both screenings varied in their relative impact on 2C::eGFP, ZSCAN4 and OCT4 expression (Fig. 5f). Quantification of 2-cell-like-cells in the top 50 hits revealed an induction that ranged between 0.28 and 2.52%, which represents a fold induction of 3.5 to 32 times compared to the controls (Fig. 5g and Supplementary Table 4).

Analysis of the protein networks of the validated top 49 hits revealed 5 major complexes that regulate 2-cell-like-cell emergence, including 23 components of the spliceosome, 4 major members of the Ep400/Tip60 (KAT5) complex, 7 members of the Polycomb (PcG) Repressor Complex 1 (PRC1), and proteins involved in DNA replication (Fig. 6a). The members of the PRC identified belong to a non-canonical PRC1 complex44,45 characterised by the presence of RYBP and PCGF646,47, suggesting specificity among the PcG proteins in their ability to regulate 2-cell-like-cell emergence. Importantly, our screening identified the second CAF-1 subunit, p60 (*Chaf1b*), in agreement with our previous work27. The mRNA expression profile of the identified hits in endogenous and CAF-1-KD-induced 2-cell-like-cells suggests that many, but not all, of the hits analysed are down-regulated in 2-cell-like-cells (Supplementary Fig. 8a). In the embryo, the expression level of all the hits obtained varied across developmental stages, with most hits displaying a sharp down-regulation or up-regulation at the 4-cell stage (Supplementary Fig. 8b).

We selected 11 hits from the major chromatin pathways identified: PRC1, Ep400/Tip60 and the replication factor Rif1, which we validated individually with single siRNAs, firstly by assessing the efficiency of knock-down (Fig. 6b). Because RNAi for spliceosome proteins resulted in increased cell death (Supplementary Fig. 6d), we did not focus on its subunits for further analysis. Instead, our subsequent analysis concentrated on *Ep400*, *Dmap1, Ring1b, Pcgf6, Rif1, L3mbtl2, Usp7, Tip60, Mga, Max and Rybp*, all of which were effectively and specifically down-regulated upon siRNA transfection (Fig. 6b, and data not shown). We included the p150 and p60 CAF-1 subunits in all our subsequent experiments as positive controls. All 2C::eGFP positive cells induced displayed robust expression of ZSCAN4 protein, similarly to endogenous and CAF-1KD-induced 2-cell-like-cells (Fig. 6c and Supplementary Fig. 9). In addition, these cells expressed the GAG protein, reflecting expression of endogenous MERVL loci (Fig. 6d) and faithful recapitulation of MERVL transcription by the 2C::eGFP (Supplementary Fig. 9). RT-qPCR demonstrated that siRNA for all 11 hits provoked a strong increase of MERVL, major satellites, *Zscan4* and transcripts from chimeric genes (Fig. 6e and Supplementary Fig. 10a). OCT4 protein was undetectable in 2C:eGFP positive cells that emerged after knock-down of our 11 hits analysed (Supplementary Fig. 9c). All these results together indicate that the 2-cell-like-cells induced upon down-regulation of our novel 11 hits are *bona fide*, as they display the known molecular characteristics of endogenous 2-cell-like-cells24,27. FACS quantification of 2C::eGFP fluorescence revealed an induction of 2-cell-like-cells between 5- and 30-fold compared to the controls (Fig. 6f). *Ep400*, *Dmap1* and *Mga* were the most effective hits: their siRNA resulted in an increase in the 2-cell-like-cell population of up to 5 to 6% of the culture, compared with only ~0.2% in the controls (Supplementary Table 4). Thus, our screening effectively led to the identification of novel regulators of 2-cell-like emergence.

We next addressed whether selected subunits of the PRC1.6 and Ep400/Tip60 complexes function in the maintenance and/or induction of 2-cell-like-cells by sorting out Zscan4+ cells and 2-cell-like-cells followed by siRNA for *Pcgf6, Ring1b, Ep400* and *Dmap1*. The percentage of 2-cell-like-cells after siRNA for all 4 subunits was the same in cultures where Zscan4+ and 2-cell-like-cells had been removed, compared to cultures with pre-existing 2-cell-like-cells (Fig. 7a), suggesting that the primary function of these proteins is in the induction, rather than in the maintenance of the 2-cell-like state. In addition, while most of the hits identified affected the percentage of both Zscan4+ and of 2-cell-like-cells similarly, some of the proteins identified had a bigger effect on 2-cell-like-cell induction (Fig. 7b and Supplementary Fig. 10b). We conclude that members of the Ep400/Tip60 complex, as well as PRC1 and Rif1, act as inhibitors of 2-cell-like emergence in ES cultures.

To understand the molecular nature of PRC1 activity in 2-cell-like-cell formation we determined which polycomb complex inhibits 2-cell-like emergence by performing siRNA for all known PRC1 subunits and associated polypeptides (Supplementary Fig. 11a). Nine of the 28 siRNAs tested led to significant increase in the 2-cell-like-cell population (Supplementary Fig. 11b; protein names highlighted in red in Supplementary Fig. 11a), with subunits specific of the non-canonical PRC1 (ncPRC1) complex PRC1.6, such as PCGF6 and L3MBTL2, acting as main gatekeepers of 2-cell-like emergence (Supplementary Fig. 11b). In line with this, downregulation of MGA and MAX, known interactors of Myc also known to assemble into PRC1.6 46,48,49, robustly induced 2-cell-like-cells (Supplementary Fig. 11b). In addition to revealing PRC1.6 as a regulator for 2-cell-like-cells, our results suggest specificity in the ncPRC1 subunits involved. Down-regulation of RYBP, but not of YAF-2 which is mutually exclusive with RYBP and defines different complexes46, induced 2-cell-like-cells (Supplementary Fig. 11c-f). Likewise, down-regulation of *Ring1b*, but not of *Ring1a*, induced 2-cell-like-cells (Supplementary Fig. 11c-f). In agreement, down-regulation of *Usp7* and *Skp1*, known interactors of RING1B, but for which a tight specific allocation to a given ncPRC1 subcomplex is unknown50, efficiently induced 2-cell-like-cells. Notably, siRNA for PRC2 components EED and EZH2, had no effect on 2-cell-like-cells emergence or MERVL transcription (Supplementary Fig. 11g-j).

Because PRC1 catalyses H2AK119 ubiquitylation, we addressed H2AK119ub levels in 2-cell-like-cells. Immunostaining for H2AK119ub revealed that endogenous and induced 2-cell-like-cells display significantly lower levels of H2AK119ub, compared to their neighbouring ESCs (Supplementary Fig. 12a and b). This is relevant considering that reduced H2AK119ub levels occurred not only upon *Ring1b/Pcgf6* down-regulation – for which this is expected – but also upon down-regulation of *Ep400,* and, albeit to a lesser extent, of *Dmap1*, suggesting that emergence of 2-cell-like-cells, regardless of the molecular pathway involved, entails decrease in H2AK119ub.

We next asked whether the epigenetic pathways identified act in parallel in regulating 2-cell-like-cell fate. siRNA for any combination of Ep400/Tip60 and PRC1.6 subunits had a clear additive effect (Fig. 7c and Supplementary Note), indicating that Ep400/Tip60 and PRC1.6 act through different pathways to induce 2-cell-like-cells. In agreement, *Rex1* and *Nanog* siRNA did not affect the fold change of 2-cell-like-cells upon *Ep400* and *Dmap1* siRNA, but did affect slightly the extent to which downregulation of *Pcgf6* and *Ring1b* induced 2-cell-like-cells (Supplementary Fig. 12c). This suggests that part of the effect of PRC1.6 – but not of Ep400 or Tip60 – in inducing 2-cell-like-cells, is dependent on REX1 and NANOG function. In addition, 2i treatment decreased the number of 2-cell-like-cells emerging after siRNA for *Pcgf6, Ring1b, Ep400* and *Dmap1* (Fig. 7d), and overexpression of *Nanog* had no effect (Supplementary Fig. 12d).

Lastly, to investigate the potential mechanism of action of PRC1.6 and Ep400/Tip60 in 2-cell-like emergence, we asked whether they occupy genes differentially expressed in 2-cell-like-cells, in ESCs. We analysed ChIP-seq data for PRC1.6 and Ep400/Tip60 subunits in ESCs. We classified differentially expressed genes as up- or down-regulated27 and analysed the enrichment of the PRC1.6 subunits RING1B, RYBP, MAX, of H2AK119ub and of Ep400 and H3K4me3 over their TSSs. Because most transcriptional changes observed in 2-cell-like-cells are in up-regulation27,51, we focused on up-regulated TSSs. K-means clustering of ChIP-seq profiles revealed 5 main binding profiles (Fig. 8a). From these, 3 clusters were strongly co-occupied by RING1B, RYBP and H2AK119ub (odds ratio for enrichment 2.84, p-value=1.24 e-06), suggesting that these genes are normally repressed by PRC1 in ESCs. The TSSs and associated transcripts from these 3 clusters were mostly silent in ESCs, and their expression was induced in 2-cell-like-cells, representing around ~17% of the upregulated TSSs (Fig. 8a). Two of the 3 PRC1-bound clusters were comprised mainly of ‘bivalent’ promoters marked by H3K4me3 and PRC1 occupancy (in addition to H3K27me3, not shown). Approximately half of the up-regulated TSSs (~57%) did not show binding of any of the chromatin modifiers or modifications analysed, presumably due to the low mappability of some of these regions (Fig. 8a). Interestingly, we found that many of them contained a MERVL (MT2_Mm) within 50 kb upstream (Fig. 8a), contrary to down-regulated genes. The up-regulation of the genes within this cluster is therefore presumably explained by proximity to MERVL1,51,52. Notably, we did not observe significant binding at MERVL for Ep400 or for PRC1 subunits RING1B or RYBP, except for a minor fraction of complete MERVL that is bound by Max (Fig. 8b; see Online Methods and data not shown). Since a large proportion of the down-regulated genes in 2-cell-like-cells are bound by Ep400 (Fig. 8c)(odds ratio for enrichment 3.70, ~60% of down-regulated genes, p-value <2.2 e-16), down-regulation of the latter may help repressing part of the ES transcriptome in 2-cell-like-cells. This also supports the observations that simultaneous down-regulation of both complexes synergistically induces 2-cell-like-cells. Together, these data suggest that the transition towards the 2-cell-like-cell transcriptome is combinatorial and can be achieved at least at two different levels: i) regulation through the binding and action of the identified chromatin modifiers and ii) activation of a proximal MERVL, presumably through the binding of the transcription factor Dux53,54.

Discussion

We have identified an intermediate cellular state during the transition to the 2-cell-like state characterized by a transcriptional profile distinctive from ES and 2-cell-like-cells. We note that, in contrast to previous findings where the Zscan4+ and the MERVL+ populations have been considered interchangeable38, our approach of classifying individual cells based on their single-cell transcriptome rather than by their fluorescence or population-wide transcriptome allowed us to uncover differences between the populations in a more robust and quantitative manner. Zscan4 is a *bona fide* marker for this intermediate population. While the activation of *Zscan4* expression demarcates the first molecular change that we detect, the extent of the transcriptional changes in Zscan4+ cells is not limited to *Zscan4* expression. This anticipates that ZSCAN4 itself may not necessarily have a directive, essential role in 2-cell-like emergence. Indeed, ZSCAN4 overexpression in ESCs seems to elicit cell death and can trigger part, but not all, repertoire of 2-cell-like-cell characteristics36.

Early differentiating cells generate fewer Zscan4+ cells compared to naïve ESCs, indicating that exit towards the 2-cell stage demarcates a different process compared to loss of pluripotency upon differentiation. Our data also indicate that 2-cell-like-cells themselves are heterogeneous, but the nature of their heterogeneity differs from ES cell heterogeneity. While 2-cell-like-cells have been proposed to constitute a metastable state24, it remains to be seen whether such a state is also in an internal dynamic equilibrium or whether the heterogeneity that we observed represents additional transitional states.

A more open chromatin characterises the 2-cell state *in vivo* and *in vitro*26,54,55. Ep400/Tip60 regulates nucleosome stability56,57, potentially providing a molecular basis for chromatin opening. Our data suggest that several activities are in play to regulate the transition from ES to 2-cell-like-cells. This is supported by the synergistic effects of PRC1.6 and Ep400/Tip60 down-regulation in 2-cell-like emergence. In addition, PRC1.6 is the only PRC1 complex that possesses histone deacetylation activity46,47, which may also impact on the global histone hyperaceylation observed in 2-cell-like-cells24,27. Notably, specific components of PRC1.6, such as L3MBTL2 and RYBP, are low or absent from embryonic chromatin in 2-cell-stage embryos58. While work on induced pluripotency indicates that reprogramming does not necessarily recapitulate developmental progression in reverse order, investigating whether the factors that we identified are responsible for regulating the 2-cell transition *in vivo* will be an important task for the future.

Our data supports a role for several biochemical activities such as chromatin assembly, nucleosome remodelling and histone acetylation and ubiquitylation in reshaping the chromatin landscape from ES to 2-cell-like cells (Fig. 8d). Altogether, our work identifies transitional states during the transition from ES to 2-cell-like-cells and chromatin pathways involved. Investigating the molecular mechanisms behind the emergence of the 2-cell state in culture, is an essential prerequisite to any statement about a potential similarity with the molecular transitions occurring in development *in vivo*.

Online Methods

Cell culture

All cell lines in this study, unless otherwise stated, were grown in media containing DMEM-Glutamax-I, 15% fetal calf serum, 2x LIF, 0.1 mM 2-betamercaptoethanol, non-essential amino acids, penicillin and streptomycin over a bed of feeder cells. For the LIF withdrawal experiment, LIF and feeders were omitted from the culture conditions. Medium supplemented with 2i (3 μM CHIR99021 and 1 μM PD0325901, Miltenyi Biotec) was used for the establishment of stable cell lines and for their expansion and maintenance. After removal of 2i, cells were always cultured for at least 7 days in Serum/LIF conditions over a bed of feeder cells before being used for experiments (except for siRNA transfection).

siRNA transfection

Two days before transfection, cells were plated in gelatin-coated dishes. The 2i inhibitors were removed from the medium 1 day before transfection. Lipofectamine RNAi MAX (Life Technologies) was used for siRNA transfection according to the manufacturer's instructions. 75,000 cells were reversed transfected in 24-well-gelatin-coated plates using 30 nM siRNA final concentration (the siRNAs are listed in Supplementary Table 5). We used Silenced Negative Control No.1 siRNA (Life Technologies) as a negative control for siRNA treatment. The effect of RNAi was examined 2 days after transfection. Pcgf6, Ring1b, Ep400, Dmap1, Mga, Max and Rybp siRNA effects on 2-cell-like cell induction were validated by FACS, IF and RT-qPCR (qPCR primers are listed Supplementary Table 6) with an additional second siRNA sequence available upon request (data not shown). The effects of Snrpd1 and Lsm6 siRNA on 2-cell-like cell induction observed in the screening were validated by FACS, IF and RT-qPCR (data not shown). However, because of the high cell toxicity observed upon siRNA to spliceosome (Supplementary Fig. 5d), we did not focus on these hits for the remainder of the work. Lipofectamine 2000 (Life Technologies) was used for co-transfection of siRNA and Nanog expression vector 59 according to the manufacturer's instructions. Cells were analysed by RT-qPCR 2 days after transfection.

RNAi Screening

Screening was performed in the high throughput screening facility of the IGBMC, using a custom siRNA library (Dharmacon) of chromatin factor related siRNA siGenome smartpool (4 different siRNAs/pool). Controls were performed with smartpool siRNAs from Dharmacon (Thermo Scientific). Transfection efficiency was optimized using a cell death siRNA that trigger cell death when transfected in the cells (quantification of transfection efficiency by assessing the toxicity) and the p150 siRNA (quantification of transfection efficiency by measuring the induction of 2-cell-like cells). For each target, 20 nM final concentration of siRNA is reverse transfected in triplicate in 5,000 mESCs by using INTERFERin®-HTS (Polyplus-Transfection). For the primary screening 1167 genes were targeted (see Supplementary Table 4). The screening was performed in 96-well microplates coated with gelatin. 2 days after siRNA transfection, cells were fixed with 1.5% paraformaldehyde, permeabilized with 0.1% Triton X-100, blocked with 2% BSA, and incubated with eGFP, ZSCAN4 and OCT4 antibodies followed by Alexa Fluor 647, 555 and 488 conjugated second antibodies (Invitrogen). Cell nuclei were counterstained with 1 μg/mL DAPI. The screening was achieved owing to a Tecan Freedom EVO 150 (for cell transfection, staining, and immunocytochemistry) and to an Orbitor™ RS Microplate Mover robotic arm coupling microplate stacks to a Cellomics CellInsight™ NXT High-Content Screening Platform (Thermo Scientific™). Images were acquired with the CellInsight™ NXT (Thermo Scientific™) and analyzed with HCS Studio™ Cell Analysis Software (nuclei segmentation and eGFP, ZSCAN4 and OCT4 intensities). Quantification of positive cells for each of these factors was done automatically, based on nuclear segmentation and across >5 fields (for actual number of cell counts see Supplementary Tables S7 and S8. The percentage of cell positive for eGFP, ZSCAN4 or OCT4 staining were quantified for each well. 2-cell-like cells are defined as cells that are positive for eGFP and ZSCAN4 but negative for OCT4 staining.

To validate selected targets from the siRNA screen, a secondary screen was performed with individual siRNAs. In the secondary screen, 81 genes of the primary screening were assessed by transfecting 4 different individual siRNAs, in triplicate for each target. The secondary screening showed high reproducibility with the primary screening (Supplementary Fig. 7a-b). In addition, 32 new genes were also investigated in the secondary screening by transfecting a pool of 4 siRNA per target as well as 9 genes already present in the primary screening as internal control (Supplementary Table 4). The methods used for the secondary screen were as described for the primary screening. For each condition, z-scores were calculated as follow: $z{=}mean\left(\frac{xi-\overline{x}}{s}\right)$; where x_i are the values for the triplicates, $\bar{x}$ and s are the mean value and the standard deviation for the negative control conditions, respectively.

Reporter cell lines

EGFP and turboGFP 2C-reporter cell lines were previously described27. To generate the Zscan4 reporter cell line, we replaced the emerald cassette of the reporter plasmid kindly provided by M. Ko 35 with a destabilized NLS-tagged tdTomato or mCherry cassette. To generate the Zscan4 reporter cell lines, the turboGFP 2C-reporter cell line was transfected with the respective plasmids using Lipofectamine 2000 and afterwards a single clone was selected. In the case of the Rex1 and Zscan4 reporter, a stable cell line carrying a knocked-in EGFP cassette into the ORF of Rex1 was kindly provided by the laboratory of Austin Smith 40 and afterwards transfected with the Zscan4c::tdTomato reporter construct. The reporter cell lines used in this study are summarized below:

Transgene 1	Transgene 2	Transgene 3	Experiments
2C::3XturboGFP-NLS-PEST	n.a.	n.a.	-RNA-seq data in Fig 1 -ZSCAN4 immunofluorescence in Fig 2 -SOX2 and REX1 immunofluorescence in Fig 4
2C::3XturboGFP-NLS-PEST	CAG::NLS-tdTomato	n.a,	-Single cell expression profiling in Fig 1, S1 and S4
2C::3XturboGFP-NLS-PEST	Zscan4c::tdTomato-PEST	n.a.	-Single cell expression profiling in Fig 2 -PRDM14 and TFAP2C immunofluorescence in Fig 3
2C::3XturboGFP-NLS-PEST	Zscan4c::mCherry-NLS-PEST	CAG::H2B-tdiRFP	- Time-lapse experiments in Fig 2 - FACS analysis experiments in Fig. 7a, 7c and S7.
2C::EGFP	n.a.	n.a.	-OCT4 immunofluorescence in Fig 3 -Screening, Figures 5 to 7
Rex1::EGFP-PEST (knock in, ref. 40)	Zscan4c::tdTomato-PEST	n.a.	-All experiments in Fig 4

Analysis of RNA-seq data and selection of genes for the single cell expression profiling experiments

RNA-seq data for 2-cell-like cells was reported previously in ref. 27. Genes used in the Biomark single cell analysis were selected on the basis of their functional significance and differential expression from our previous bulk RNA-seq analysis. Chimeric genes were defined on the basis of transcription starting from an LTR from the MT2 families by genome browser analysis of the RNA-seq in Ishiuchi et al27. Heatmaps of this transcriptomics data were generated using the ggplot2 R package.

Validation of Taqman assays and custom designs

Taqman assays (ThermoFisher Scientific) were first tested on 3 different 5-fold serial dilutions of cDNA from turboGFP$^+$ cells on a LightCycler qPCR instrument. Taqman assays that failed to amplify or that did not exhibit linearity in their measurements when compared between the different dilutions were omitted from the single cell analysis. All Taqman assays used are described in Supplementary Table 1. The Taqman used for Zscan4 amplifies Zscan4c, d and f. Custom primers were designed to target turboGFP reporter and were mixed to a final concentration equivalent to that of the Taqman assays (18 μM for the primer and 4 μM for the probe).

Single cell expression profiling

After thawing, reporter cell lines were cultured for 6 days in Serum/LIF conditions over a bed of feeders and passaged every single day except for the second day of culture. On the 6th day of culture, cells were sorted with the help of a FACS machine and only 2C::turboGFP$^-$

cells were replated; 2-cell-like cells were discarded. 24 hours later, cells were sorted once again but this time either the ESC, the 2-cell-like or the Zscan4$^+$ fraction was preserved and prepared according to the manufacturer's protocol (Fluidigm, PN1006117) for use in Fluidigm's C1 microfluidics-based single cell sample preparation platform. Of note, 1 μL of a 1:286 dilution of ERCC spike-in mix was added to the lysis solution of each C1 run instead of water. Single cell expression data was generated in Fluidigm's Biomark qPCR platform in technical duplicates using Taqman probes (ThermoFisher Scientific) according to the manufacturer's protocol (Fluidigm, PN68000130). Only cells which passed the quality control check were included in the analyses performed throughout the manuscript. In total, we profiled 136 cells across three biological and two technical replicates. A constitutive *pCAG* promoter driving *NLS-tdTomato* expression allowed us to sort feeder cells out (Fig. 1b and Supplementary Fig. 1c). In addition, we used two spike-in controls, which allowed us to assess the quality of the normalization to endogenous control genes and therefore to constrain technical noise (Supplementary Fig. 1d)

Normalization, quality control and modeling of single cell data

Ct values obtained from the Biomark platform were processed as described previously60. Briefly, Ct values higher than 28 or with quality scores lower than the threshold of 0.65 were substituted with Ct values of 28. Subsequently, Ct values were subtracted from a baseline value of 28 so that 0 implies no expression and 28 a high level of expression. For normalization purposes, the average of *Actb* and *Gapdh* Ct values was subtracted from the values of all other genes for the same cell in order to obtain positive values in all assays and samples. The Ct values of both technical replicates of the same cell were averaged. Please note that, unlike RNA-seq approaches, higher levels of expression of specific genes, relative to other is unlikely to bias the gene expression data because the usage of a pool of primers specific for each of the assayed genes provides additional robustness to a single highly-expressed gene being preferentially amplified and biasing a cell's transcriptional profile. For the principal component analysis, the principal components of the dataset were computed using the svd method in R, and no scaling was performed. For the projection of the Zscan4 dataset into the principal components of the ESC/2-cell-like dataset, principal components were calculated using the svd method for the ESC/2-cell-like dataset and afterwards its loadings matrix and the Zscan4 dataset's matrix were multiplied. All plotting was done in R with the ggplot2 package. In order to classify cells into the ESCs, Zscan4-low, Zscan4-mid, Zscan4-high and 2-cell-like categories we first classified them into two groups based on whether they expressed *Zscan4c/d/f* or not. Cells that did not express *Zscan4* were termed ESCs. The remaining cells were subsequently classified based on whether the *turboGFP* transcript level exceeded the established cutoff, which was determined with the help of the density function of *turboGFP* expression. Cells with expression values above the cutoff were classified as 2-cell-like cells. Note that this threshold had to be defined because the *tbGFP* reporter is intron-less and it therefore has to be distinguished from the genomic background (1C, threshold selection not shown). The remaining cells were classified into Zscan4-low, -mid or -high cells according to the thresholds that were set over *Zscan4's* density function in Supplementary Fig. 2c.

Gapdh and *Actb* were used as internal controls for normalisation since they reflect independent molecular pathways and are unrelated to pluripotency. Indeed, their expression was stable across samples and showed a high correlation. The Ct values for ES (GFP-) and 2-cell-like cells (GFP+) were consistently similar for the spike-in RNAs after normalization (Supplementary Fig. 1c), supporting the robustness of the normalization. Once data processing was performed, all cells expressing outlier values in *Actb*, *Gapdh* or the spike-in RNAs were removed from the analysis.

For the principal component analysis, the principal components of the dataset were computed using the pcaMethods package in R (ref. 61), and no scaling was performed. For the projection of the Zscan4 dataset into the principal components of the ESC/2-cell-like dataset, principal components were calculated for the ESC/2-cell-like dataset and afterwards its loadings matrix and the Zscan4 dataset's matrix were multiplied.

All plots were done in R with the ggplot2 package.

Flow cytometry

Cells were washed with room temperature sterile PBS, trypsinized and re-suspended in ice-cold sterile 0.5% BSA PBS solution. Sorting was performed using a BD BioSciences FACS Aria II or III. During sorting, cells were collected in culture medium and kept at 4 °C during the sort. Analysis of FACS data was performed using the FlowJo software and the same gatings were used for all replicates of a same experiment. Cells were not index sorted and the purity of the sort was estimated at 96% for 2C-like and 97% for ESCs. A control flow profile for wt ESCs is shown in Supplementary Figure 1c. For the Rex1 experiments, the Rex1$^-$ gate was defined based on the fluorescence of WT ESCs and the Rex1^{high} gates were defined based on the fluorescence of Rex1-GFP cells cultured in 2i. FACS Calibur (BD Biosciences) was used to quantify the population of eGFP-positive cells. Cells that had been frozen in 2i were thawed and the Rex1 sorting was performed 4 days after 2i withdrawal. For data presented in Figure 7a, the 2C::turboGFP and Zscan4c::mCherry cell line was FACS sorted just before transfection and the 2-cell-like cells or Zscan4::mCherry positive cells were removed respectively from the population.

Immunofluorescence, image processing and quantification

Cells were cultured over feeder-coated coverslips, fixed in PFA for 10 minutes at room temperature and permeabilized with 0.2% Triton X-100 for another 10 minutes. Primary antibodies were incubated overnight followed by 3 washes in PBS. Secondary antibodies were incubated for 1 hour. Mounting was done in Vectashield mounting medium (Vector Labs). Image acquisition was performed using a Leica SP8 confocal microscope. For immunofluorescence quantifications, manual segmentation of the cell nucleus was performed on the DAPI channel using ImageJ and the average fluorescence intensity was measured. ImageJ software and afterwards the average fluorescence intensity was measured. Only ESCs, ZSCAN4$^+$ cells and 2-cell-like cells from the same coverslip and imaging session were used for each comparison. Density plots for each of these groups were computed using the kernel density estimation function in R using the Sheather & Jones bandwidth selection method. For 2C::eGFP immunostaining, cells were fixed as described and blocked for 30 min at 37°C in 10% FCS, 3% BSA and 0.1% Triton X-100 in PBS (blocking buffer). Primary and secondary antibodies were each incubated for 30 min at 37°C in the blocking buffer solution.

Real-time PCR

Total RNA was extracted from ESCs with the RNeasy Plus mini kit (Qiagen) and treated with turbo DNase (Life Technologies) to remove genomic DNA. Reverse transcription was performed with SuperScript II (Life Technologies) with random hexamers. Real-time PCR was performed with Lightcycler 480 SYBR Green I Master Mix (Roche) on a LightCycler 96 Real-time PCR System (Roche). The relative expression level of each gene was normalized to *Gapdh* and *Actb*. The primers used in this study are listed in Supplementary Table 6.

Antibodies

Antibodies used in this work were the following: mouse turboGFP (OTI2H8, Origene), rabbit turboGFP (PA5-22688, ThermoFisher), rabbit Zscan4 (AB4340, EMD Millipore), rabbit MuERVL-Gag (R1501-2, Hangzhou HuaAn Biotechnology), chicken eGFP (ab13970, Abcam), rabbit H2AK119ub (8240, Cell Signaling), goat Rex1 (sc-50670, Santa Cruz), goat Oct4 (sc-8628, Santa Cruz), mouse Oct4 (611203, BDBiosciences), goat Sox2 (sc-17320, Santa Cruz), rabbit Prdm14 (gift from D. Reinberg, 62, rabbit Tfap2c (sc-8977, Santa Cruz).

Time-lapse experiments

Prior to time-lapse, cells were cultured overnight on glass-bottom laminin-coated (Sigma-Aldrich) Ibidi micro-Insert cell culture dishes to allow the cells to attach. Image-acquisition was carried out over the entire well with a 20x 0.75 NA Plan-Apochromat objective every 30 minutes using a Nikon Ti-E equipped with Bruker Opterra II multipoint confocal system for 48 hours. Images were recorded on an EMCCD camera using emission filters for turboGFP (BP520/40), mCherry (570LP) and iRFP (655LP) mounted on a FLI filter wheel. Spontaneously arising 2-cell-like cells were manually identified using ImageJ software and scored based on whether they arose from a Zscan4$^+$ cell based on the intensity of mCherry channel. This analysis was ran independently by two different people and cross-compared for accuracy.

ATAC-seq data analysis

Previously published ATAC-seq data in ESCs, Zscan4$^+$ and 2-cell-like cells was obtained from GEO accession GSE75751 38. Paired end reads were trimmed for adaptor sequences using trimmomatic 0.36 and mapped to the mm10 reference genome using bowtie2 with parameters --dovetail -X 2000 --no-discordant --no-mixed. The resulting bam files were then filtered for non-uniquely mapping reads using samtools with a MAPQ threshold of 10 and filtered for duplicates using Picard Tools' MarkDuplicates. Finally, mitochondrial reads were removed using samtools and signal tracks were generated using macs2.1.1 63 with parameters --SPMR --nomodel --nolambda --shift -100 --extsize 200 for the combined reads of all replicates of the same population. For the MERVL/MT2_Mm analysis in Figure 2g and 2h, MT2_Mm coordinates were obtained from RepeatMasker release 20140131 and plots were generated using deepTools. For clarity, only solo-LTRs are shown in Figure 2h but the same pattern of enrichment was also observed for full-length MERVLs. For the differential accessibility analysis in Supplementary Figure 2h, regions of differential accessibility between ESCs and 2-cell-like cells were called using the bdgdiff macs2 command, and plots were generated using deepTools.

Single cell RNA-seq data analysis

Single cell RNA-seq data for ESCs, Zscan4^{+} and 2-cell-like cells was obtained from ArrayExpress accession E-MTAB-5058 (ref. 38). Paired end reads were trimmed for adaptor sequences using trimmomatic 0.36 and mapped to the mm10 reference genome using STAR 2.5.3a (ref. 64). The resulting bam files were then filtered for unmapped reads and secondary alignments using samtools. Finally, reads intersecting repetitive elements were quantified by intersecting the RepeatMasker annotation to the aligned bam files using bedtools intersect with the –split parameter. To classify the individual cells in a comparable manner to the classification performed on the single cell qPCR data, uniquely-mapping read counts mapping to the full-length Zscan4 isoforms c, d and f were summed. Note that the Zscan4 Taqman assay used for the single cell qPCR experiments cannot discriminate bertween the c, d and f isoforms, and hence our decision to pool the read counts for these isoforms. The following thresholds were used to classify individual cells: cells with $\log_2$(MT2_Mm, CPM + 1) > 12.25 were considered 2-cell-like cells; cells with $\log_2$(Zscan4c/d/f, CPM + 1) > 24 were considered Zscan4-high; cells with $\log_2$(Zscan4c/d/f, CPM + 1) > 12 were considered Zscan4-mid; cells with $\log_2$(Zscan4c/d/f, CPM + 1) > 2 were considered Zscan4-low; all remaining cells were considered ESCs. The positioning of these thresholds in the context of the population-wide distribution of expression levels is shown in Supplementary Figure 3d. For the heatmap shown in Figure 2g, 2313 upregulated and 951 downregulated genes were selected based on the 2-cell-like and ESCs RNA-seq data reported in Ishiuchi et al27. The mean expression level for each gene within each respective group of single cell transcriptomes was taken, and the row-wise Z-score calculated. Plotting was performed using ggplot2. The differential expression analysis in Supplementary Figure 3f was performed using DESeq2 65 by considering each individual cell as a replicate. Changes were considered statistically significant when p(adj.)<0.05.

ChIP-seq data analysis

ChIP-seq data for H3K4me3 (GSE74112)66, Ep400 (GSE64825)67, H2AK119ub (GSE89949)68, Ring1b (GSE40860)69, Rybp (GSE42466)70 and Max (GSE48175)71 from ESCs grown under serum/LIF conditions were obtained from GEO. Note that none of these datasets were generated in the presence of 2i. Single end reads were trimmed for adaptor sequences and low quality bases using trimmomatic 0.36 and mapped to the mm10 reference genome using bowtie with parameters -S --best --strata -v 2 -M 1. The resulting bam files were then filtered for non-uniquely mapping reads using samtools with a MAPQ threshold of 10 and filtered for duplicates using Picard Tools' MarkDuplicates. Finally, fragment size was estimated using MaSC and sequencing-depth normalized signal tracks were generated using macs2.1.1 using the MaSC 72estimated fragment size. Biomart was used to obtain 10 Kb-wide genomic windows centered on all TSSs of the 2313 upregulated (6829 TSSs) and the 951 downregulated (4189 TSSs) genes between 2-cell-like cells and ESCs as reported 27. Mappability tracks were generated using the GEM suite using a 36bp window. Heatmaps shown on Figure 8a and c were plotted using deepTools73 and custom scripts. A TSS was considered MT2_Mm associated if an MT2_Mm copy was found within 50 Kb upstream. For the enrichment calculation of given chromatin modifier, peaks were called using macs2.1.1 using a q-value cutoff of 10^{-6} and peaks within 1 kb of each other were merged. Genes with at least one of their TSSs occurring within 5 Kb of a ChIP-seq peak were considered bound by the respective factor. Fisher's exact tests for these overlaps were performed using R.

ChIP-seq enrichment at MERVL and MT2_Mm

Input and treatment reads were mapped individually to the mm10 reference genome using Bowtie74 (1.1.1) with the following options: "-k 201 --best –strata –m 200", allowing for up to 200 alignments of the highest quality to be reported. MACS63 (2.1.1) "predictd –g mm" was used to estimate the fragment size on the uniquely mapping treatment reads. Because repetitive regions are the focus of this analysis, CSEM 75 (2.4) was used to assign multi-mapping reads to their most likely origin. It was run with the option "—upper-bound 500", and the fragment size estimated above. For each multi-mapping read, only the most likely alignment is selected. If multiple equally likely alignments exist, one is selected at random. MACS "bdgcmp" is used to generate fold-enrichment tracks from these input and treatment alignments. Mappability tracks were generated by splitting the mm10 genome into overlapping reads of 36 bp length, starting at each base, and aligning them back to the reference genome using Bowtie74 (1.1.1). A base is "mappable", if it can be aligned to the reference genome uniquely, "unmappable" otherwise. To avoid mappability issues, for transposable elements (TEs) MERVL and MT2_MM enrichment P-values were calculated using a permutation test. First, the enrichment profile was calculated as the mean of all rows in the enrichment heatmap. Then, the maximum value of the enrichment profile emax was recorded around windows 1kb up- and 1k downstream of the TE. A permuted set of regions was then generated by selecting random genomic locations that maintain number, size, and chromosome distribution of the original elements. As before, the maximum value of the enrichment profile in the random set rmax was recorded. The permutation P-value was then calculated as: p = emax / Σi rimax. , where i denotes the ith permutation. For every P-value, 10000 randomisations were performed. For the analysis, MERVLs were classified into complete MERVL according to the presence of contiguous MT2_Mm LTRs within 7kb distance with the same 5' to 3' orientation interspersed with an internal MERVL element as annotated by Repeat Masker (v4.0.3). MT2_Mm LTRs without a corresponding MERLV internal part or a nearby (<7kb) LTR pair in the same orientation were classified as solo LTRs. Occurrences of MERLV internal elements without matching MT2_Mm LTRs flanking these elements were classified as internal MERVLs. The *p* values for the enrichment for the following factors for the MERVL ChIPseq analyses are: for MAX: 0.0001, 1.0 and 0.9801 for complete MERVL, MT2_Mm and MERVL-int, respectively; for RYBP: 0.0537, 0.9610 and 0.5597 for complete MERVL, MT2_Mm and MERVL-int, respectively; for H2AK119ub (not shown): 0.6739, 1.0 and 1.0000 for complete MERVL, MT2_Mm and MERVL-int, respectively.

Statistical analyses

Statistical tests were performed minding data distribution and the number of data points available. Details on sample sizes, in addition to the statistical tests conducted, are presented on the corresponding figure legends. Asterisks indicate P-values below the 0.05 threshold.

Primers and Taqman assays used in this study

The list of Taqman assays is in Supplementary Table 1, and of primers used in Supplementary Table 6.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank A. Smith for providing the knock-in Rex1 reporter cell line, M. Ko for the Zscan4c promoter plasmid, R. Enriquez-Gasca for providing a classification of MERVLs prior to publication. A. Ettinger for timelapse analysis, C. Ebel, D. Pich, T. Hofer and W. Hammerschmidt for help and access to FACS, the INGESTEM infrastructure for access to the IGBMC high throughput high content screening workstation, C. Thibault, F. Recillas-Targa and M. Zurita-Ortega for helpful discussions and A. Burton for critical reading of the manuscript. M.E.T.-P. acknowledges funding from EpiGeneSys NoE, ERC-Stg 'NuclearPotency' (280840), the EMBO Young Investigator Programme, the Fondation Schlumberger pour l'Education et la Recherche (2016-Torres-Padilla) and the Helmholtz Association. J.M.V. acknowledges funding from the Max Planck Society and Epigenesys NoE. T.I. was a recipient of post-doctoral fellowships from Uehara Memorial Foundation and Human Frontier Science Programme (LT000015/2012-L) and D. R.-T was partially supported by a DGECI fellowship (2890/2014) from the National University of Mexico.

References

1. Ishiuchi T, Torres-Padilla ME. Towards an understanding of the regulatory mechanisms of totipotency. Current opinion in genetics & development. 2013; 23:512–8. [PubMed: 23942314]

2. Surani MA, Hayashi K, Hajkova P. Genetic and epigenetic regulators of pluripotency. Cell. 2007; 128:747–62. [PubMed: 17320511]

3. Wu G, Scholer HR. Lineage Segregation in the Totipotent Embryo. Current topics in developmental biology. 2016; 117:301–17. [PubMed: 26969985]

4. Nichols J, Smith A. The origin and identity of embryonic stem cells. Development. 2011; 138:3–8. [PubMed: 21138972]

5. Tarkowski AK. Experiments on the development of isolated blastomers of mouse eggs. Nature. 1959; 184:1286–7. [PubMed: 13836947]

6. Tarkowski AK, Wroblewska J. Development of blastomeres of mouse eggs isolated at the 4- and 8-cell stage. J Embryol Exp Morphol. 1967; 18:155–80. [PubMed: 6048976]

7. Tsunoda Y, McLaren A. Effect of various procedures on the viability of mouse embryos containing half the normal number of blastomeres. Journal of reproduction and fertility. 1983; 69:315–22. [PubMed: 6887141]

8. Evans MJ, Kaufman MH. Establishment in culture of pluripotential cells from mouse embryos. Nature. 1981; 292:154–6. [PubMed: 7242681]

9. Smith AG, et al. Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. Nature. 1988; 336:688–90. [PubMed: 3143917]

10. Mitsui K, et al. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ESCs. Cell. 2003; 113:631–42. [PubMed: 12787504]

11. Chambers I, et al. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell. 2003; 113:643–55. [PubMed: 12787505]

12. Scholer HR, Hatzopoulos AK, Balling R, Suzuki N, Gruss P. A family of octamer-specific proteins present during mouse embryogenesis: evidence for germline-specific expression of an Oct factor. Embo J. 1989; 8:2543–50. [PubMed: 2573523]

13. Canham MA, Sharov AA, Ko MS, Brickman JM. Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. PLoS biology. 2010; 8:e1000379. [PubMed: 20520791]

14. Chambers I, et al. Nanog safeguards pluripotency and mediates germline development. Nature. 2007; 450:1230–4. [PubMed: 18097409]

15. Hayashi K, Lopes SM, Tang F, Surani MA. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. Cell stem cell. 2008; 3:391–401. [PubMed: 18940731]

16. Kalmar T, et al. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS biology. 2009; 7:e1000149. [PubMed: 19582141]

17. Toyooka Y, Shimosato D, Murakami K, Takahashi K, Niwa H. Identification and characterization of subpopulations in undifferentiated ES cell culture. Development. 2008; 135:909–18. [PubMed: 18263842]

18. Torres-Padilla ME, Chambers I. Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. Development. 2014; 141:2173–81. [PubMed: 24866112]

19. Martinez Arias A, Brickman JM. Gene expression heterogeneities in embryonic stem cell populations: origin and function. Current opinion in cell biology. 2011; 23:650–6. [PubMed: 21982544]

20. Morgani SM, et al. Totipotent embryonic stem cells arise in ground-state culture conditions. Cell reports. 2013; 3:1945–57. [PubMed: 23746443]

21. Marks H, et al. The transcriptional and epigenomic foundations of ground state pluripotency. Cell. 2012; 149:590–604. [PubMed: 22541430]

22. Alexandrova S, et al. Selection and dynamics of embryonic stem cell integration into early mouse embryos. Development. 2016; 143:24–34. [PubMed: 26586221]

23. Martin Gonzalez J, et al. Embryonic Stem Cell Culture Conditions Support Distinct States Associated with Different Developmental Stages and Potency. Stem cell reports. 2016; 7:177–91. [PubMed: 27509134]

24. Macfarlan TS, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 2012; 487:57–63. [PubMed: 22722858]

25. Falco G, et al. Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. Developmental biology. 2007; 307:539–50. [PubMed: 17553482]

26. Boskovic A, et al. Higher chromatin mobility supports totipotency and precedes pluripotency in vivo. Genes & development. 2014; 28:1042–7. [PubMed: 24831699]

27. Ishiuchi T, et al. Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. Nature structural & molecular biology. 2015; 22:662–71.

28. Grun D, van Oudenaarden A. Design and Analysis of Single-Cell Sequencing Experiments. Cell. 2015; 163:799–810. [PubMed: 26544934]

29. Etzrodt M, Endele M, Schroeder T. Quantitative single-cell approaches to stem cell research. Cell stem cell. 2014; 15:546–58. [PubMed: 25517464]

30. Buganim Y, et al. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell. 2012; 150:1209–22. [PubMed: 22980981]

31. Guo G, et al. Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. Cell stem cell. 2013; 13:492–505. [PubMed: 24035353]

32. Leitch HG, et al. Naive pluripotency is associated with global DNA hypomethylation. Nature structural & molecular biology. 2013; 20:311–6.

33. Ficz G, et al. FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. Cell stem cell. 2013; 13:351–9. [PubMed: 23850245]

34. Habibi E, et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. Cell stem cell. 2013; 13:360–9. [PubMed: 23850244]

35. Zalzman M, et al. Zscan4 regulates telomere elongation and genomic stability in ESCs. Nature. 2010; 464:858–63. [PubMed: 20336070]

36. Amano T, et al. Zscan4 restores the developmental potency of embryonic stem cells. Nature communications. 2013; 4:1966.

37. Hirata T, et al. Zscan4 transiently reactivates early embryonic genes during the generation of induced pluripotent stem cells. Scientific reports. 2012; 2:208. [PubMed: 22355722]

38. Eckersley-Maslin MA, et al. MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. Cell reports. 2016; 17:179–92. [PubMed: 27681430]

39. Cahan P, Daley GQ. Origins and implications of pluripotent stem cell variability and heterogeneity. Nature reviews Molecular cell biology. 2013; 14:357–68. [PubMed: 23673969]

40. Wray J, et al. Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation. Nature cell biology. 2011; 13:838–45. [PubMed: 21685889]

41. Fazzio TG, Huff JT, Panning B. An RNAi screen of chromatin proteins identifies Tip60-p400 as a regulator of embryonic stem cell identity. Cell. 2008; 134:162–74. [PubMed: 18614019]

42. Hisada K, et al. RYBP represses endogenous retroviruses and preimplantation- and germ line-specific genes in mouse embryonic stem cells. Molecular and cellular biology. 2012; 32:1139–49. [PubMed: 22269950]

43. Suzuki A, et al. Loss of MAX results in meiotic entry in mouse embryonic and germline stem cells. Nature communications. 2016; 7 11056.

44. Aloia L, Di Stefano B, Di Croce L. Polycomb complexes in stem cells and embryonic development. Development. 2013; 140:2525–34. [PubMed: 23715546]

45. Schwartz YB, Pirrotta V. A new world of Polycombs: unexpected partnerships and emerging functions. Nature reviews Genetics. 2013; 14:853–64.

46. Gao Z, et al. PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Molecular cell. 2012; 45:344–56. [PubMed: 22325352]

47. Levine SS, et al. The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans. Molecular and cellular biology. 2002; 22:6070–8. [PubMed: 12167701]

48. Ogawa H, Ishiguro K, Gaubatz S, Livingston DM, Nakatani Y. A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells. Science. 2002; 296:1132–6. [PubMed: 12004135]

49. Zhao W, et al. Essential Role for Polycomb Group Protein Pcgf6 in Embryonic Stem Cell Maintenance and a Noncanonical Polycomb Repressive Complex 1 (PRC1) Integrity. The Journal of biological chemistry. 2017

50. Sanchez C, et al. Proteomics analysis of Ring1B/Rnf2 interactors identifies a novel complex with the Fbxl10/Jhdm1B histone demethylase and the Bcl6 interacting corepressor. Molecular & cellular proteomics : MCP. 2007; 6:820–34. [PubMed: 17296600]

51. Macfarlan TS, et al. Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes & development. 2011; 25:594–607. [PubMed: 21357675]

52. Peaston AE, et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell. 2004; 7:597–606. [PubMed: 15469847]

53. De Iaco A, et al. DUX-family transcription factors regulate zygotic genome activation in placental mammals. Nature genetics. 2017; 49:941–945. [PubMed: 28459456]

54. Hendrickson PG, et al. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. Nature genetics. 2017; 49:925–934. [PubMed: 28459457]

55. Wu J, et al. The landscape of accessible chromatin in mammalian preimplantation embryos. Nature. 2016; 534:652–7. [PubMed: 27309802]

56. Xu Y, et al. The p400 ATPase regulates nucleosome stability and chromatin ubiquitination during DNA repair. The Journal of cell biology. 2010; 191:31–43. [PubMed: 20876283]

57. Pradhan SK, et al. EP400 Deposits H3.3 into Promoters and Enhancers during Gene Activation. Molecular cell. 2016; 61:27–38. [PubMed: 26669263]

58. Eid A, Torres-Padilla ME. Characterization of non-canonical Polycomb Repressive Complex 1 subunits during early mouse embryogenesis. Epigenetics. 2016; 11:389–97. [PubMed: 27081692]

59. Miyanari Y, Torres-Padilla ME. Control of ground-state pluripotency by allelic regulation of Nanog. Nature. 2012; 483:470–3. [PubMed: 22327294]

60. Guo G, et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell. 2010; 18:675–85. [PubMed: 20412781]

61. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods--a bioconductor package providing PCA methods for incomplete data. Bioinformatics. 2007; 23:1164–7. [PubMed: 17344241]

62. Burton A, et al. Single-Cell Profiling of Epigenetic Modifiers Identifies PRDM14 as an Inducer of Cell Fate in the Mammalian Embryo. Cell reports. 2013

63. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008; 9:R137. [PubMed: 18798982]

64. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. [PubMed: 23104886]

65. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. bioRxiv. 2014

66. Liu Z, Kraus WL. Catalytic-Independent Functions of PARP-1 Determine Sox2 Pioneer Activity at Intractable Genomic Loci. Molecular cell. 2017; 65:589–603 e9. [PubMed: 28212747]

67. de Dieuleveult M, et al. Genome-wide nucleosome specificity and function of chromatin remodellers in ESCs. Nature. 2016; 530:113–6. [PubMed: 26814966]

68. Kundu S, et al. Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. Molecular cell. 2017; 65:432–446 e5. [PubMed: 28157505]

69. Farcas AM, et al. KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. eLife. 2012; 1:e00205. [PubMed: 23256043]

70. Morey L, Aloia L, Cozzuto L, Benitah SA, Di Croce L. RYBP and Cbx7 define specific biological functions of polycomb complexes in mouse embryonic stem cells. Cell reports. 2013; 3:60–9. [PubMed: 23273917]

71. Krepelova A, Neri F, Maldotti M, Rapelli S, Oliviero S. Myc and max genome-wide binding sites analysis links the Myc regulatory network with the polycomb and the core pluripotency networks in mouse embryonic stem cells. PloS one. 2014; 9:e88933. [PubMed: 24586446]

72. Ramachandran P, Palidwor GA, Porter CJ, Perkins TJ. MaSC: mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data. Bioinformatics. 2013; 29:444–50. [PubMed: 23300135]

73. Ramirez F, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic acids research. 2016; 44:W160–5. [PubMed: 27079975]

74. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9:357–9. [PubMed: 22388286]

75. Chung D, et al. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLoS computational biology. 2011; 7:e1002111. [PubMed: 21779159]

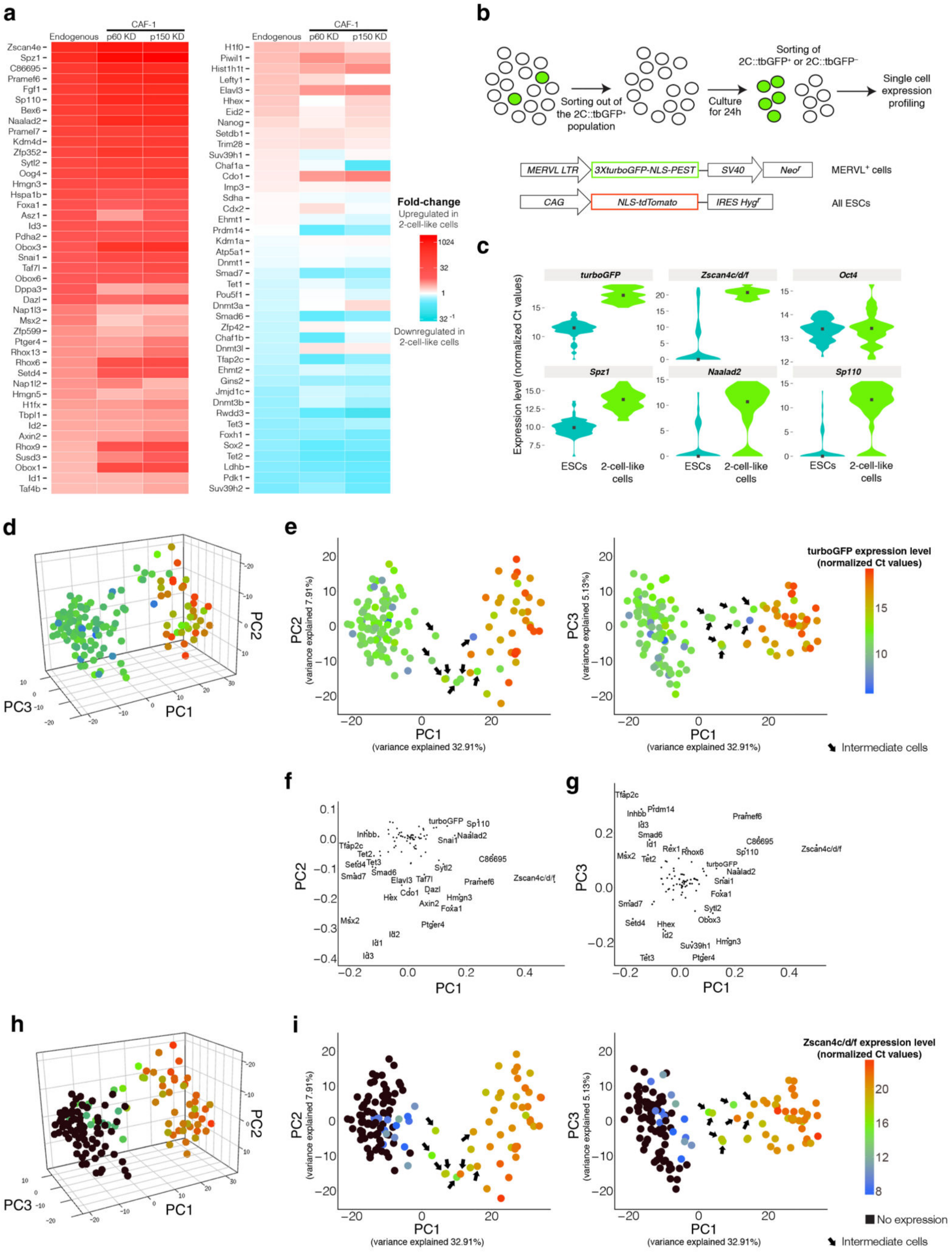

Figure 1. Zscan4$^+$ cells exhibit an intermediate expression profile between ESCs and 2-cell-like-cells.

a. Heatmap showing the changes in expression levels of the genes selected for the single cell analysis in endogenous, p60 KD-induced and p150 KD-induced 2-cell-like-cells. Fold-changes are calculated based on bulk RNA-seq data 27 and colour coded relative to ESCs.

b. Experimental design.

c. Violin plots showing the distribution of the expression levels of individual cells for the indicated genes. Higher values correspond to higher expression levels and a Ct value of 0 indicates that no amplification was detected. Median is indicated by a square. Note that turboGFP and Spz1 possess no exon junctions and therefore their readings reflect to some extent the genomic background.

d-e. Different viewpoints of the Principal Component Analysis of the single cell expression dataset. Each point corresponds to a single cell, which is coloured according to the expression level of turboGFP. PC1 separates ES and 2-cell-like-cells. Black arrows indicate turboGFP^{-} cells with an intermediate expression profile between that of ESCs and that of 2-cell-like-cells.

f-g. Principal Component projection of the 93 genes used for the analysis, showing the contribution of each gene to the first three principal components. Only the most influential genes are labelled.

h-i. PCA as in d-e, but individual cells (dots) are colour-coded according to their levels of Zscan4c/d/f expression. Black dots correspond to no expression.

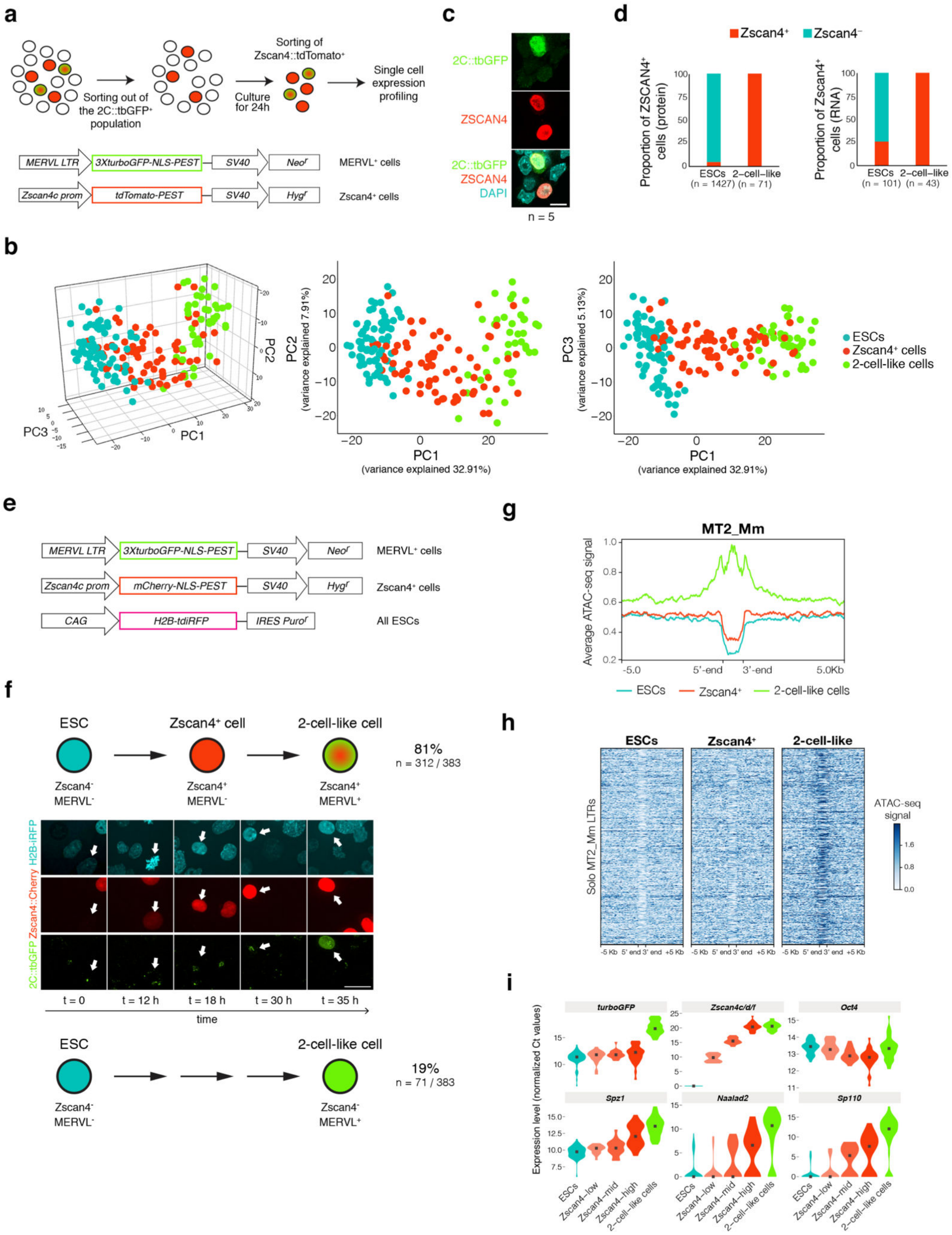

Figure 2. Two-cell-like cells arise primarily from Zscan4$^+$ cells

a. Experimental design.

b. Projection of the individual expression profiles of Zscan4$^+$ cells onto the principal components of the ES and 2-cell-like-cell dataset from Figure 1d-e. Each dot represents a single cell and is coloured according to whether it corresponds to an ES, a Zscan4$^+$ cell or a 2-cell-like-cell as indicated.

c. Immunofluorescence of 2C::turboGFP and endogenous ZSCAN4. A representative confocal single section image from 5 different cell culures is shown. Scale bar, 10 μm.

d. Quantification of the number of cells expressing ZSCAN4 protein (left), based on immunofluorescence, and mRNA (right), based on the single cell expression analysis, according to whether they are ES or 2-cell-like-cells. n = number of cells.

e. Design of the reporter cell line used for the time-lapse shown in Fig. 2f.

f. Time-lapse analysis of 2-cell-like and Zscan4$^+$ cells emergence. Spontaneously arising 2-cell-like-cells were classified depending on whether they arose from a Zscan4::mCherry$^+$ cell or not. Results shown are pooled from 11 independent cultures. Scale bar, 20 μm. n = number of cells.

g. Average ATAC-seq signal intensity over all annotated *MT2_Mm* LTRs in ESCs, Zscan4$^+$ and 2-cell-like-cells.

h. Heatmaps showing ATAC-seq signal intensity over 10 Kb genomic windows centred on 914 solo *MT2_Mm* LTRs.

i. Violin plots showing the distribution of the expression levels of individual cells for the indicated genes. Higher values correspond to higher expression levels and a Ct value of 0 indicates that no amplification was detected. Median is indicated by a square. Note that turboGFP and Spz1 possess no exon junctions and therefore their readings reflect to some extent the genomic background.

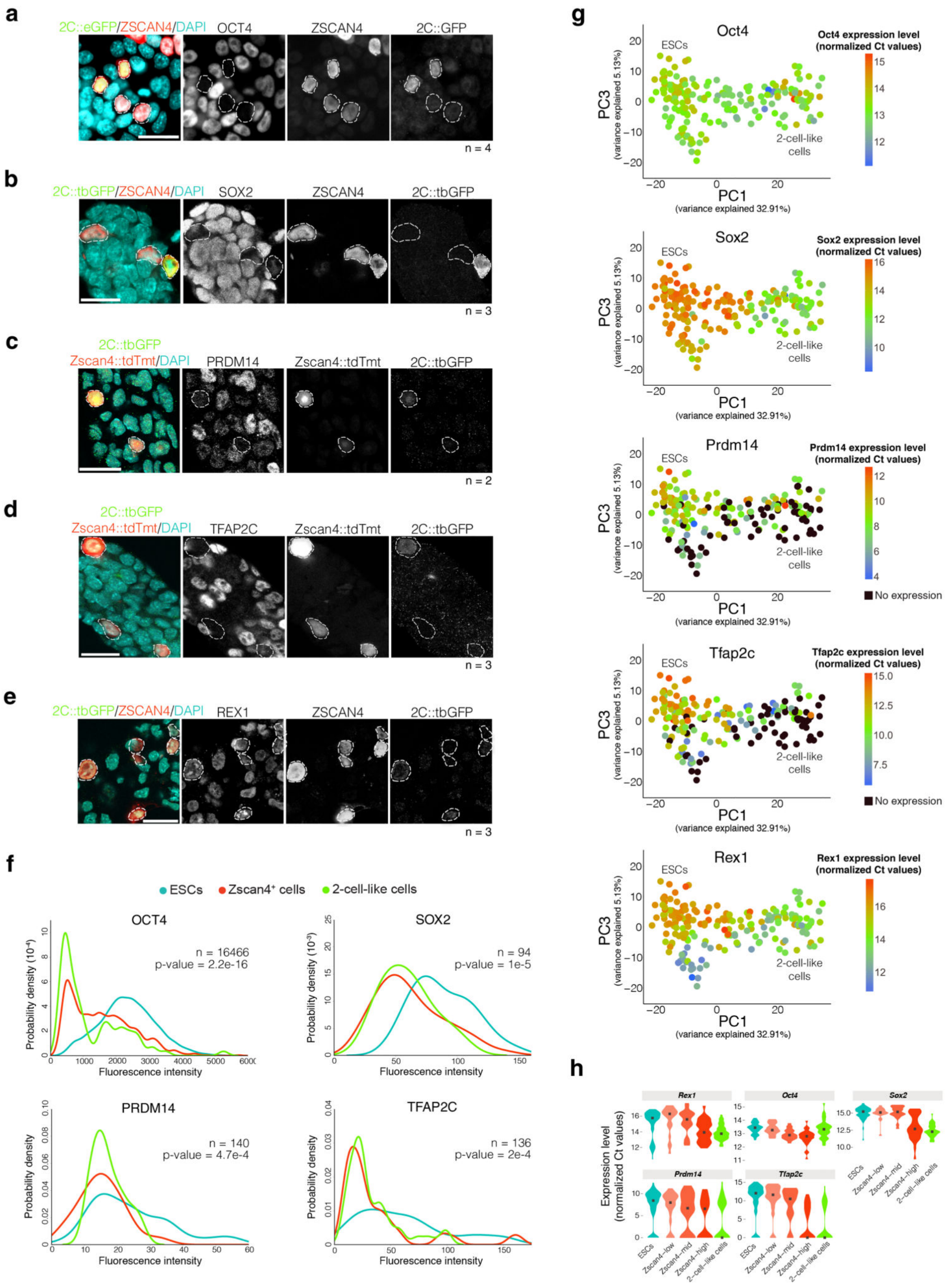

Figure 3. Zscan4$^+$ cells downregulate pluripotency factors

a-e. Immunofluorescence stainings using antibodies for the indicated proteins together with 2C::tbGFP or 2C::eGFP, and Zscan4 or Zscan4::tdTomato, as indicated. The merge image shows the DAPI (cyan), Zscan4 (red) and 2C-reporter (green) expression. Grayscale images of the respective TF, Zscan4 and 2C-reporter channel are shown on the right. Dashed lines contour Zscan4$^+$ and 2-cell-like-cells. Scale bars, 20 μm. n = number of independent cultures.

f. Quantification of the data in panels a-e. Density plots of the distribution of mean fluorescence intensities for the indicated TF. The number of cells quantified for each graph is indicated. P-values were calculated using the Mann-Whitney U test. n = number of cells.

g. Principal Component analysis of the single cell dataset from Figure 2b, with single cells colour-coded according to their expression level of the indicated transcription factor. Black dots signify no expression. As in Figure 1 and 2, PC1 separates ES from 2-cell-like-cells, while PC3 segregates naive versus primed pluripotency.

h. Violin plots showing the distribution of the expression levels of individual cells for the indicated genes. Higher values correspond to higher expression levels and a Ct value of 0 indicates that no amplification was detected. Median is indicated by a square.

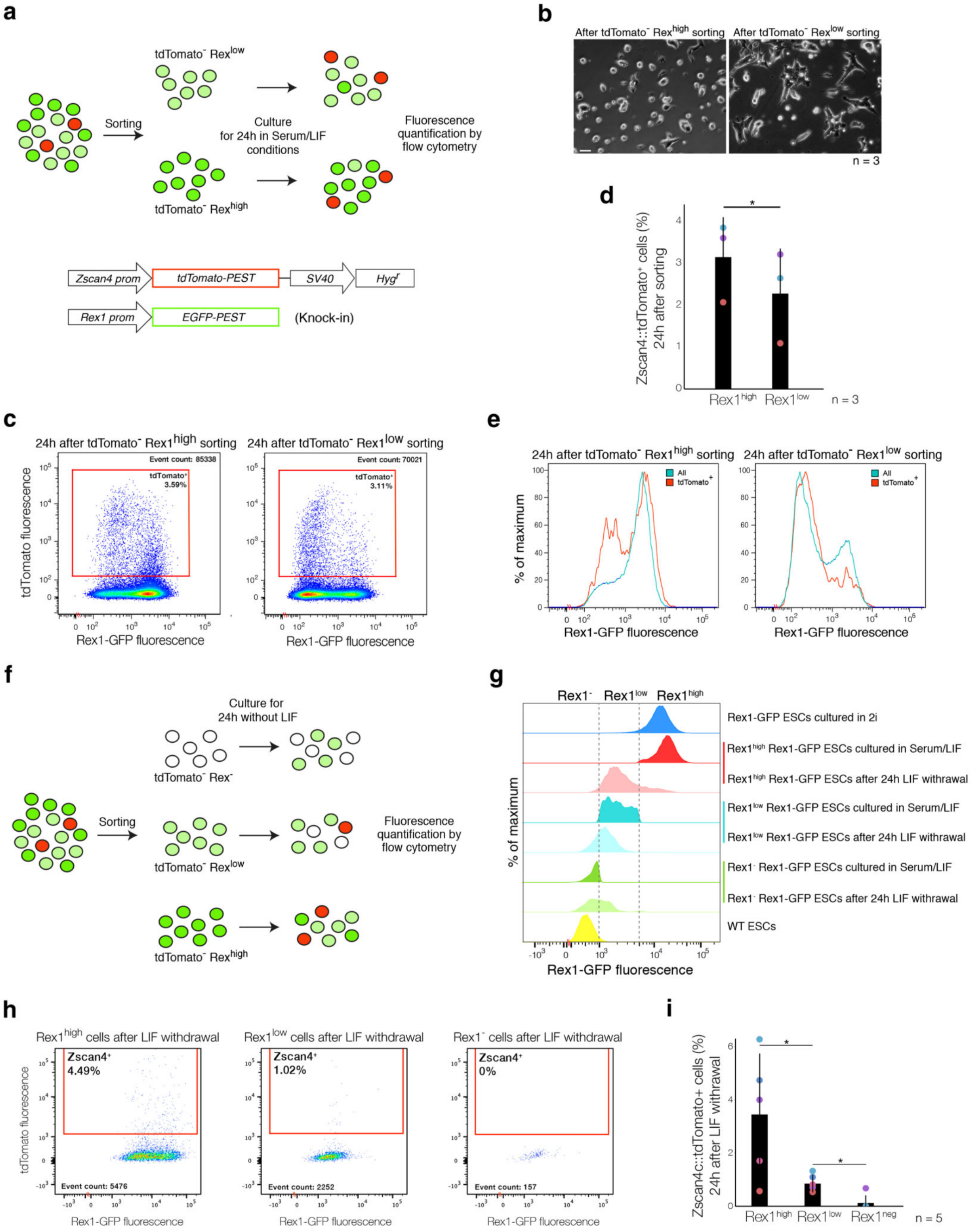

Figure 4. Entry into the Zscan4$^+$ state and downregulation of pluripotency factors is unrelated to differentiation

a. Experimental design. Rex1^{high} and Rex1^{low} tdTomato$^-$ cells were sorted based on their Rex1::EGFP fluorescence and cultured for 24 hours in serum/LIF over feeders. After this culture period, the number of Zscan4$^+$ cells was determined by FACS.

b. Phase-contrast image of Rex1^{high} and Rex1^{low} cells 24h after sorting. Scale bar, 10 μm. n = number of independent cultures.

c. Representative scatter plot from data from 4d, showing *Zscan4::tdTomato* and *Rex1*-EGFP fluorescence measurements of individual cells as assayed by FACS.

d. Quantification of the Zscan4::tdTomato$^+$ population in the Rex1^{high} and the Rex1^{low} fractions as in C. Error bars indicate the s.d. of 3 independent experiments from different cultures. * $P < 0.05$, paired student's t-test.

e. Density plots of Rex1-EGFP fluorescence for the Rex1^{high} and the Rex1^{low} populations after 24 h of culture under serum/LIF conditions. The entire population is indicated in blue and Zscan4$^+$ cells are indicated in red. Plots are representative from 3 independent cultures.

f. Experimental design. Following one week of culture under serum/LIF on feeders, tdTomato$^+$ cells were removed and Rex1^{high}, Rex1^{low} and Rex1$^-$ cells were replated without LIF over gelatin for 24 hours, after which the proportion of Zscan4$^+$ cells was determined by FACS.

g. Density plots of Rex1-EGFP fluorescence for the Rex1^{high}, Rex1^{low} and Rex1$^-$ populations after sorting and 24 hours after LIF withdrawal. Plots are representative from experiments shown in 4i.

h. Scatter plots showing Zscan4::tdTomato and *Rex1*-EGFP fluorescence FACS measurements of individual cells. Plots are representative from experiments shown in 4i.

i. Quantification of the data presented in H. Error bars indicate s.d. of 5 independent experiments from different cell cultures. * $P < 0.05$; paired student's t-test.

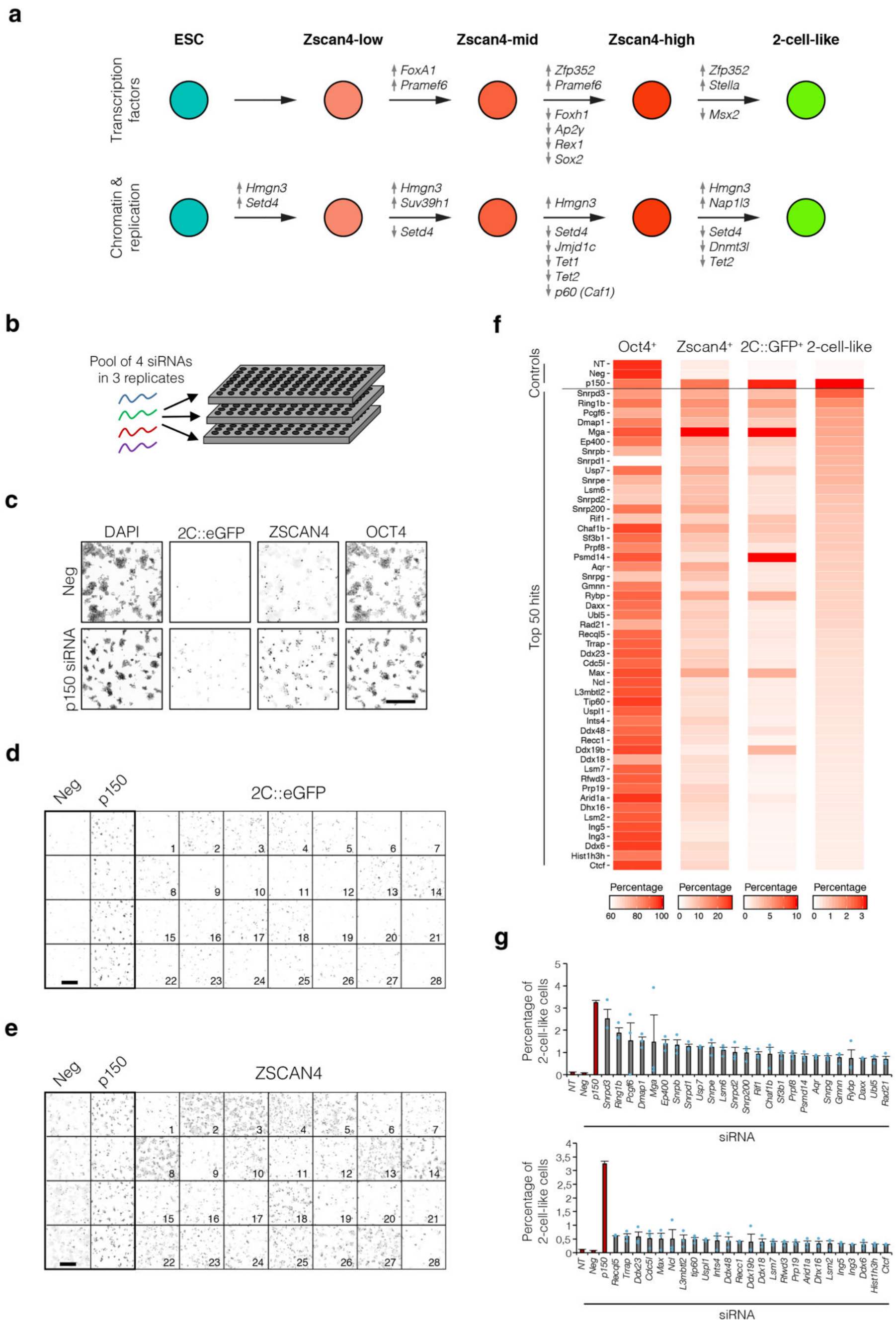

Figure 5. A chromatin modifier siRNA screen identifies regulators of 2-cell-like-cells emergence

a. Schematic representation of significantly and differentially expressed genes between individual stages of the transition from the ES to the 2-cell-like state (Supplementary Table 9). Changes were considered significant if they exhibited at least 2-fold changes across cells between individual states and a P < 0.05 (Mann-Whitney U test).

b. Design of siRNA screen.

c. Representative inverted dynamics images for the negative (scramble siRNA cells “neg”; n=270 wells) and positive (p150 siRNA transfected cells “p150”; n=270 wells) controls of the screening are shown. Nine random images from one well were combined. Scale bar, 500 µm.

d, e. Representative images of the eGFP (d) and ZSCAN4 (e) immunostaining for selected hits from the screening are shown in inverted dynamics. Scale bar represents 500µm. Images correspond to the following siRNA in order, as indicated: 1/Snrpd3 2/Ring1b 3/Pcgf6 4/ Dmap1 5/Ep400 6/Snrpb 7/Snrpd1 8/Usp7 9/Snrpe 10/Lsm6 11/Snrpd2 12/Snrp200 13/ Rif1 14/p60 15/Sf3b1 16/Prpf8 17/Psmd14 18/Aqr 19/Snrpg 20/Gmnn 21/Dnmt3b 22/Ubl5 23/ Rad21 24/Recql5 25/Trrap 26/Ddx23 27/Cdcl5 28/Ncl.

f. Heatmap showing the effect on OCT4, ZSCAN4 and 2C::eGFP expression of the top 50 hits of the screening, ranked according to the percentage of 2-cell-like-cells induced upon siRNA. Data are combined from the primary and the secondary validation screening, and includes additional hits that were tested during the secondary screening. 2-cell-like-cells are defined as cells that are positive for eGFP and ZSCAN4 but negative for OCT4 immunostaining.

g. Quantification of 2-cell-like-cells (e.g. 2C::eGFP positive, ZSCAN4 positive and OCT4 negative) induced upon siRNA of the top 50 targets (grey) from the primary and secondary screening. The mean values ± s.d. from triplicate wells are shown. Each dot corresponds to measurements of independent wells. Negative (NT, non-transfected; and Neg, scramble siRNA) and positive (p150) controls are shown in red.

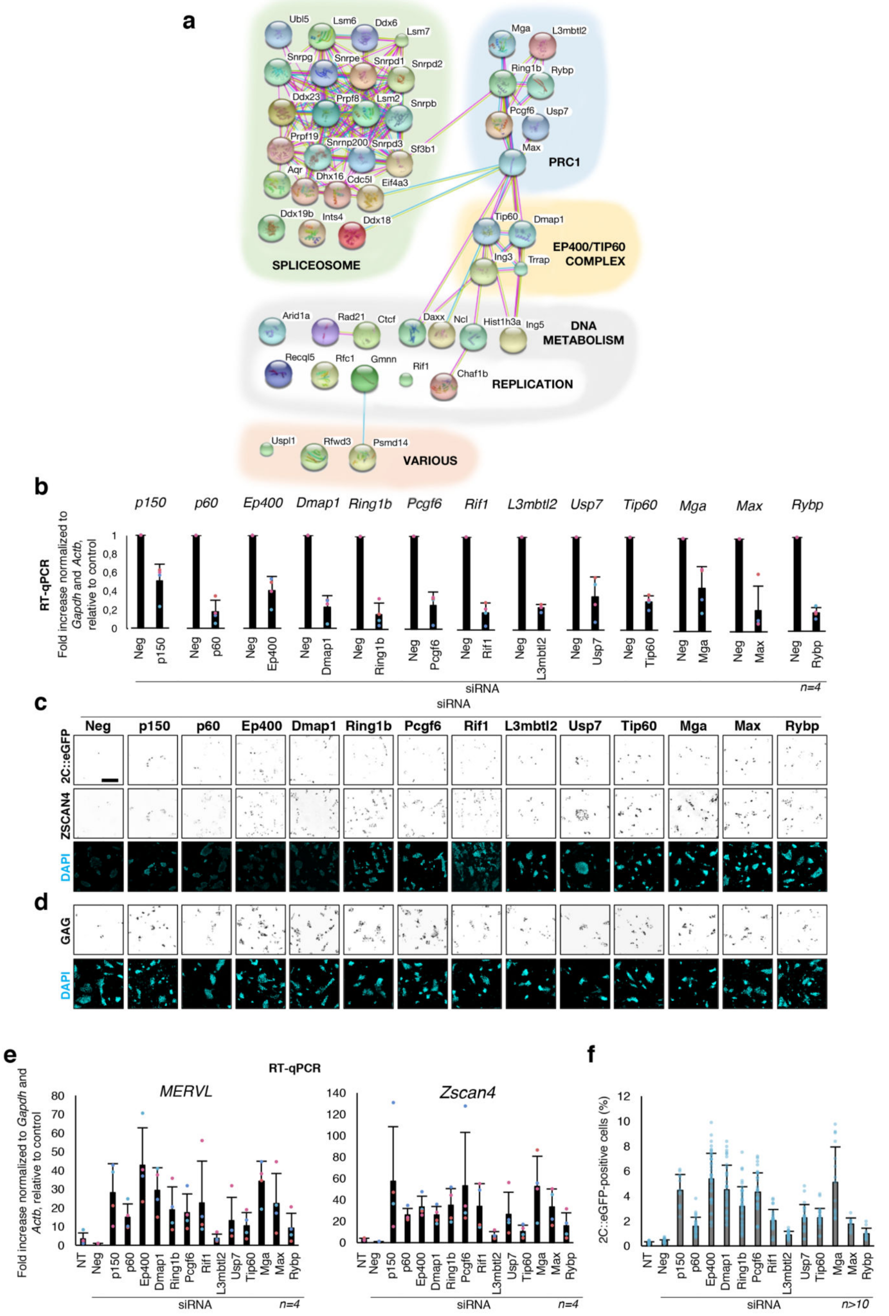

Figure 6. The PRC1.6 and Ep400/Tip60 complexes are inhibitors of 2-cell-like-cells emergence.

a. Protein interaction network for the validated top 49 hits from the screening.

b. RT-qPCR analysis of the indicated transcripts after the transfection of the corresponding siRNAs in the 2C::eGFP reporter cell line. Shown are mean values ± s.d. of two technical replicates from 4 independent cultures performed on different days.

c-d. Immunostaining of the 2C::eGFP reporter cell line with antibodies to eGFP and ZSCAN4 (*c*) or eGFP and the GAG coded by the endogenous MERVL loci (*d*) after transfection of the indicated siRNA. Images are shown in inverted dynamics. Nuclei were counterstained with DAPI (cyan). Scale bar, 200μm.

e. RT-qPCR analysis of MERVL and Zscan4 expression in the 2C::eGFP reporter cell line after the transfection of the indicated siRNA. Shown are mean values ± s.d. of two technical replicates from 4 different cell cultures performed on different days.

f. Quantification of eGFP-positive cells (%) by FACS after transfection of the indicated siRNA. Mean ± S.D. from independent culture measurements is shown. Each dot indicates measurements from independent cell cultures in panels b, e and f.

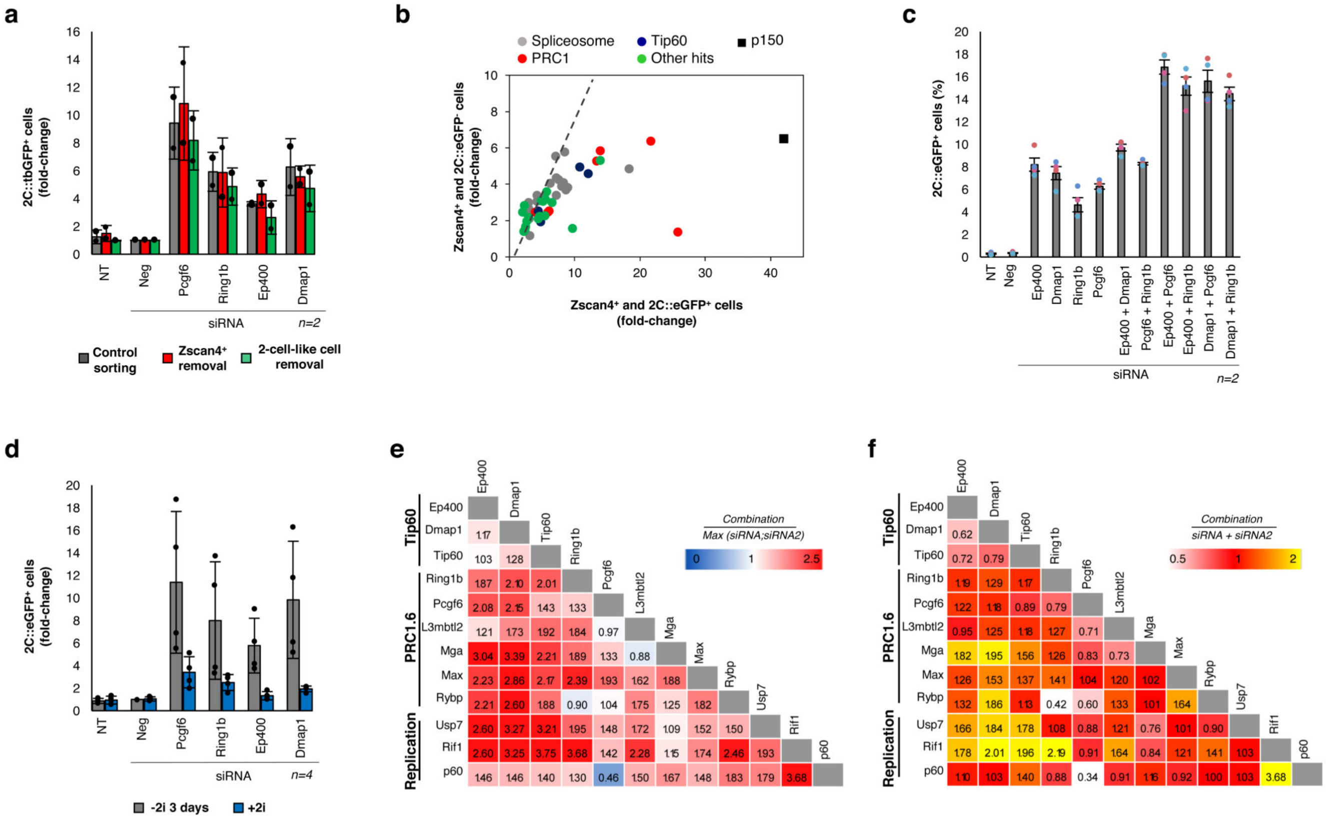

Figure 7. PRC1.6 and Ep400/Tip60 regulate 2-cell-like-cells emergence synergistically

a. Induction of 2-cell-like-cells upon siRNA for Pcgf6, Ring1b, Ep400 or Dmpa1 in an initial population of cells devoid of 2-cell-like or of Zscan4-positive cells. In control, ‘sorted’ cells, cells were passed through the FACS sorter, but Zscan4+ and 2-cell like cells were not removed. Quantification of tbGFP-positive cells by FACS was performed 2 days after siRNA transfection. Mean ± s.d. of 2 different transfected wells (dots) is shown.

b. Scatter plot of the top 50 hits of the screening showing the fold change of the percentage of ZSCAN4 and 2C::eGFP double positive cells (x-axis) and of cells positive for ZSCAN4 and negative for 2C::eGFP (y-axis). Each dot is colour-coded based on its molecular complex. The dashed grey line indicates a 1:1 ratio.

c. Quantification of eGFP-positive cells (%) by FACS after transfection of the indicated single or combined siRNA. Shown are mean values ± s.d. of two independent cell cultures.

d. Effect of 2i treatment on 2-cell-like-cell induction upon siRNA for Pcgf6, Ring1b, Ep400 or Dmap1Dmpa1. eGFP-positive cells (%) was quantified by FACS 48h after transfection. Cells were grown in 2i and were either kept in 2i for the whole experiment or 2i was removed 1 day before transfection. Mean ± s.d. of the indicated number of independent cell cultures (shown by individual dots) is shown.

e-f. Summary of the combinatorial siRNA analyses on 2-cell-like-cells induction. Pairs of siRNA were co-transfected in the 2C::eGFP reporter cell line and the percentage of 2C::eGFP positive cells was measured by FACS. Combinatorial additive (e) or synergistic effects (f) were assessed.

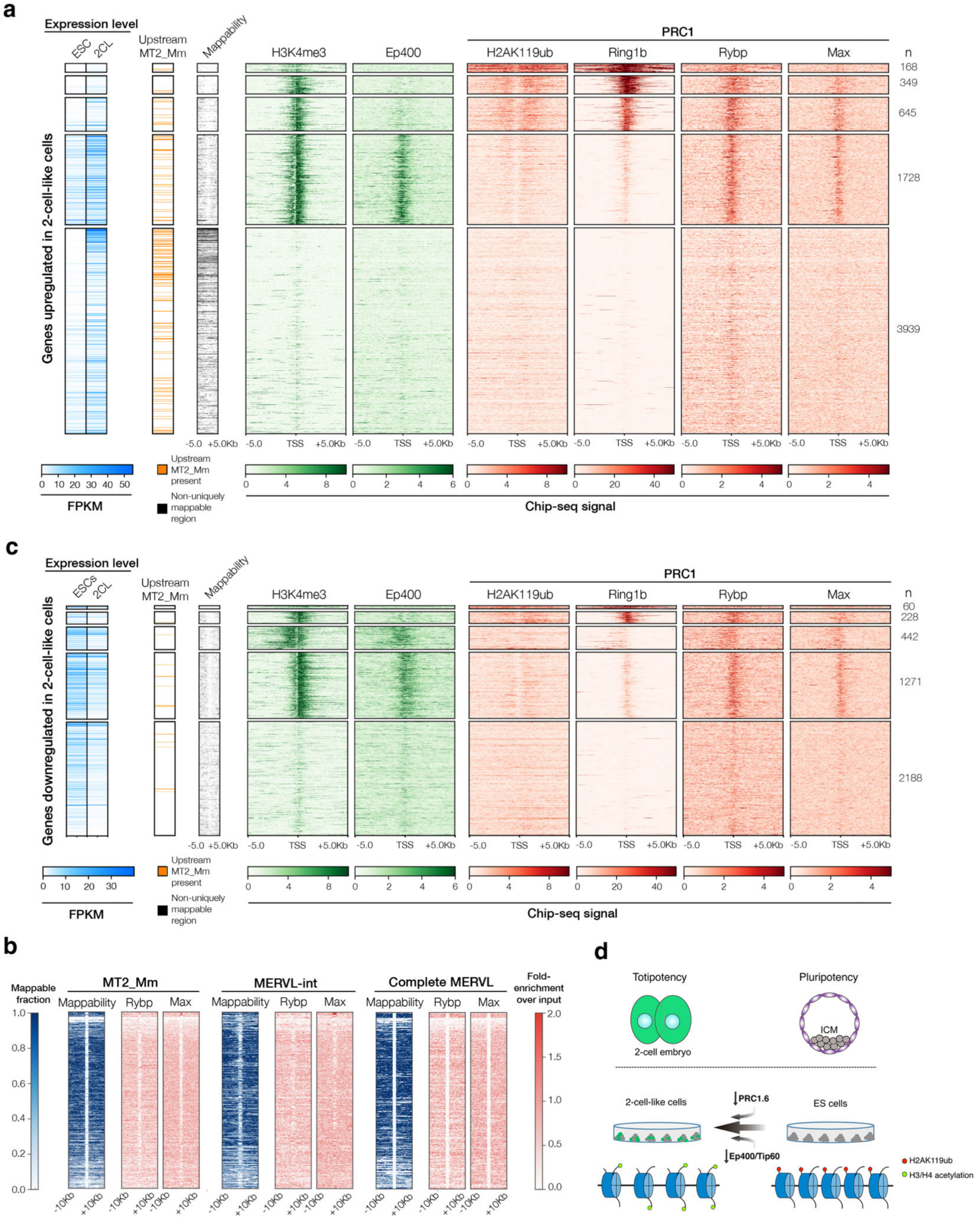

Figure 8.

a, c. Heatmap of the ChIP-seq profiles (for references see Online Methods) of the indicated chromatin modifiers or histone modifications over all TSSs of the (a) 2313 upregulated (6829 TSSs) and the (c) 951 downregulated (4189 TSSs) genes in ESCs grown in serum/LIF conditions. Regions of low mappability are depicted in black, occurrence of an *MT2_Mm* within the 50kb upstream of the respective TSS is indicated in orange; expression level in ES and 2-cell-like-cells is shown in blue.

b. Heatmap of Rybp and Max enrichment (fold-enrichment over input) in 20kb around complete, internal, or solo LTR *MERVLs* as indicated. The mappability score for the corresponding regions is depicted.

d. Working model for 2-cell-like-cells induction from mouse ES cell cultures based on the siRNA screening results. Downregulation of various subunits of the PRC1.6 and Ep400/Tip60 complexes leads to the induction of the 2-cell-like state, which is accompanied by a global reduction of H2AK119ub levels and increased histone acetylation24,27. The precise interplay of the two complexes as well as the extent to which these mechanisms may operate *in vivo* to repress the totipotent regulatory program remains to be established.