SCIENTIFIC REPORTS

OPEN

Computational analysis of single-cell transcriptomics data elucidates the stabilization of Oct4 expression in the E3.25 mouse preimplantation embryo

Received: 4 December 2018
Accepted: 6 June 2019
Published online: 20 June 2019

Daniela Gerovska[1,2] & Marcos J. Araúzo-Bravo[1,2,3,4]

Our computational analysis focuses on the 32- to 64-cell mouse embryo transition, Embryonic day (E3.25), whose study in literature is concentrated mainly on the search for an early onset of the second cell-fate decision, the specification of the inner cell mass (ICM) to primitive endoderm (PE) and epiblast (EPI). We analysed single-cell (sc) microarray transcriptomics data from E3.25 using Hierarchical Optimal *k*-Means (HO*k*M) clustering, and identified two groups of ICM cells: a group of cells from embryos with less than 34 cells (E3.25-LNCs), and another group of cells from embryos with more than 33 cells (E3.25-HNCs), corresponding to two developmental stages. Although we found massive underlying heterogeneity in the ICM cells at E3.25-HNC with over 3,800 genes with transcriptomics bifurcation, many of which are PE and EPI markers, we showed that the E3.25-HNCs are neither PE nor EPI. Importantly, analysing the differently expressed genes between the E3.25-LNCs and E3.25-HNCs, we uncovered a non-autonomous mechanism, based on a minimal number of four inner-cell contacts in the ICM, which activates *Oct4* in the preimplantation embryo. *Oct4* is highly expressed but unstable at E3.25-LNC, and stabilizes at high level at E3.25-HNC, with *Bsg* highly expressed, and the chromatin remodelling program initialised to establish an early naïve pluripotent state. Our results indicate that the pluripotent state we found to exist in the ICM at E3.25-HNC is the *in vivo* counterpart of a new, very early pluripotent state. We compared the transcriptomics profile of this *in vivo* E3.25-HNC pluripotent state, together with the profiles of E3.25-LNC, E3.5 EPI and E4.5 EPI cells, with the profiles of all embryonic stem cells (ESCs) available in the GEO database from the same platform (over 600 microarrays). The shortest distance between the set of inner cells (E3.25, E3.5 and E4.5) and the ESCs is between the E3.25-HNC cells and 2i + LIF ESCs; thus, the developmental transition from 33 to 34 cells decreases dramatically the distance with the naïve ground state of the 2i + LIF ESCs. We validated the E3.25 events through analysis of scRNA-seq data from early and late 32-cell ICM cells.

The mouse preimplantation development begins with the division of the 1-cell zygote to progressively smaller cells, blastomeres, forming the morula, which at the 8-cell stage compacts. As a result of a first cell-fate decision, the morula ball forms a hollow and becomes blastocyst with outer cells forming the trophectoderm (TE) and interior cells, the inner cell mass (ICM). The ICM sets aside from outer cells in two successive waves of asymmetric cell division, 8–16-cell and 16–32-cell transitions, with Morris *et al.*[1] reporting an additional third one. A second cell-fate decision involves the segregation of the ICM, appearing morphologically as a homogeneous cell

[1]Computational Biology and Systems Biomedicine Group, Biodonostia Health Research Institute, Calle Doctor Beguiristain s/n, San Sebastián, 20014, Spain. [2]Computational Biomedicine Data Analysis Platform, Biodonostia Health Research Institute, Calle Doctor Beguiristain s/n, San Sebastián, 20014, Spain. [3]IKERBASQUE, Basque Foundation for Science, Calle María Díaz Harokoa 3, 48013, Bilbao, Spain. [4]CIBER of Frailty and Healthy Aging (CIBERfes), Madrid, Spain. Correspondence and requests for materials should be addressed to M.J.A.-B. (email: mararabra@yahoo.co.uk)

population, into embryonic epiblast (EPI), pluripotent and progenitor of the future body, and primitive endoderm (PE), which becomes a morphologically distinct monolayer by implantation at E4.5[2].

The mechanism governing the ICM cell specification in the early blastocyst is still unclear. Ohnishi *et al.*[3] looked for cell heterogeneity in the ICM at E3.25 that could indicate early onset of the second cell-fate decision, the time preceding the apparent segregation of PE and EPI at E3.5, but were not able to detect an early splitting transcriptomics event using state-of-the-art clustering techniques.

Oct4 is master regulator of pluripotency[4] and cornerstone not only in developmental biology but also in regenerative medicine for its role in somatic cellular reprogramming[5–7]. *Oct4* is expressed *de novo* in the ICM after being down-regulated in the early cleavage stages[8], however the mechanism that reactivates *Oct4* in the ICM remains unknown.

To obtain a more complete picture of the cell specification events occurring between 32- and 64-cell stage, we developed a new clustering algorithm, and used it to look for structure in the heterogeneity during the 32–64 cell wave of divisions, for transcriptomics events explaining the loss of totipotency in the ICM, and for the mechanism behind the reactivation of *Oct4*. We analysed the single single (sc) transcriptomics microarray data of Ohnishi *et al.*[3] with a Hierarchical Optimal *k*-Means (HO*k*M) algorithm, and succeeded in identifying two groups of cells in the E3.25 ICM, one of cells from embryos with less than 34 cells, and another of cells from embryos with more than 33 cells, which, we hypothesized, characterized two developmental sub-stages in the mouse preimplantation embryo (named later E3.25-LNCs, less than 34 cells, and E3.25-HNC, more than 33 cells, respectively). Then, we confirmed that the ICM split at E3.25 into E3.25-LNC and E3.25-HNC is not due to sex, karyotype aberration or miss-assignment of the studied single cells to TE. Next, we calculated the differently expressed genes (DEGs) between the two groups, E3.25-LNCs and E3.25-HNCs, and thus we found *Oct4* among the top-up-regulated genes in the E3.25-HNCs. It is worth mentioning that the number of all possible partitions of the 36 sc transcriptomics dataset of E3.25 from Ohnishi *et al.*[3] is the Bell number $B_{36} = 3.8197 \times 10^{30}$, and the number of all possible bi-partitions is the Stirling number of the second kind $S_{36,2} = 3.4360 \times 10^{10}$. It was staggering to discover that from this possible big number of partitions, $S_{36,2}$, the partition discovered by HO*k*M was associated with a stabilization of *Oct4* at high expression level. Previously, Ohnishi *et al.*[3] found bimodal expression of *Fgf4* within the E3.25 ICM cells, and suggested that as an early indication of future PE or EPI fate. We hypothesized that such bimodal expression of *Fgf4*, among others, could be traced to earlier developmental stages and performed a bifurcation analysis on the sc data of Ohnishi *et al.*[3], spanning the time from E3.25, E3.5 to E4.5, while introducing an additional bifurcation point (between E3.25-LNC and E3.25-HNC) during E3.25, owing to the newly identified E3.25 sub-stages. Thus, we found that *Fgf4*, among other PE and EPI markers, has bimodal expression earlier than E3.25-LNC. To compare the transcriptional similarity of the E3.25-LNC and E3.25-HNC with other developmental stages, we calculated PCA and violin plots of the top up-regulated genes in the E3.25-HNC, using the data of Ohnishi *et al.*[3] together with data from the same and earlier developmental stages of the same Affymetrix platform from other authors. To check whether the E3.25-HNCs are similar to PE or EPI cells, we performed scatter plots between the E3.25-HNC and E3.5 PE or E3.5 EPI cells of Ohnishi *et al.*[3]. We used the availability of Fgf4-KO data in the dataset of Ohnishi *et al.*[3] to compare the transcriptional similarity of the E3.25-LNC and E3.25-HNC with E3.25 Fgf4-KO, E3.5 Fgf4-KO and E4.5 Fgf4-KO. We have retrospectively validated the HO*k*M clustering algorithm on the E3.5 and E4.5 sc transcriptomics microarray data of Kurimoto *et al.*[9] to group the PE and EPI cells, and since their sc microarray dataset is produced on the same platform, we used it together with the E3.5 and E4.5 data of Ohnishi *et al.*[3] to find common and specific PE and EPI markers for E3.5 and E4.5. Then, we evaluated the closeness of all existent ESCs in the GEO database from the same Affymetrix platform to their *in vitro* ICM counterparts from Ohnishi *et al.*[3]. In order to validate the existence of developmental sub-stages around E3.25, we analysed the scRNA-seq data from Posfai *et al.*[10]. Finally, we built a simple cell-packing model to explain the strong transcriptional changes between ICM cells from embryos with less than 34 cells and ICM cells from embryos with more than 33 cells.

Results

E3.25 ICM cells segregate into lower- (LNC) and higher- number-of-cell (HNC) clusters. The Principal Component Analysis (PCA) (Fig. 1A) of transcriptomics data from oocyte till early post-implantation stages (Table 1) shows that first and second Principal Components reflect the embryo age and cell fate, respectively. Focusing on the E3.25-E3.5 stages and using only the ICM sc data from Ohnishi *et al.*[3] (Fig. 1B), the PCA revealed that E3.25 cells divide into two groups, one comprised mainly of cells from embryos with lower number of cells (LNCs), between 32 and 33, and cell 26 C41 IN from a 41-cell embryo, and another comprised of cells from embryos with higher number of cells (HNCs), between 34 and 50.

PCA explains only ~30% of the transcriptomics information (Fig. S1A,B). Therefore, to analyse the splitting event of the E3.25 ICM, accounting simultaneously for maximum amount of information and reducing the intrinsic noise of the sc data, we applied a Hierarchical Optimal *k*-Means (HO*k*M) clustering algorithm that determines the optimal number of cell subgroups and their members (see Methods). Ohnishi *et al.*[3] attempted to find an early splitting event between PE and EPI precursors at E3.25 with a conventional clustering method but detected no such event. With the HO*k*M clustering algorithm, we were open to identifying any splitting event, rather than one limited to PE and EPI precursors. The E3.25 ICM cells split best into two groups (optimal silhouette coefficient $k = 2$) (Fig. 1C) based on the number of cells of the embryo of origin: one, comprising all cells from LNC (32–33-cell) embryos, and another, comprising all cells from HNC (34–50-cell) embryos (Fig. 1D). We call these two clusters E3.25-LNC and E3.25-HNC from here on. Thus, we recognized a possible developmental sub-stage splitting event.

The ICM transcriptomics split at E3.25 into E3.25-LNC and E3.25-HNC is not due to sex, karyotype aberration or mis-assignment to ICM. In order to confirm that the split in the ICM at E3.25 into

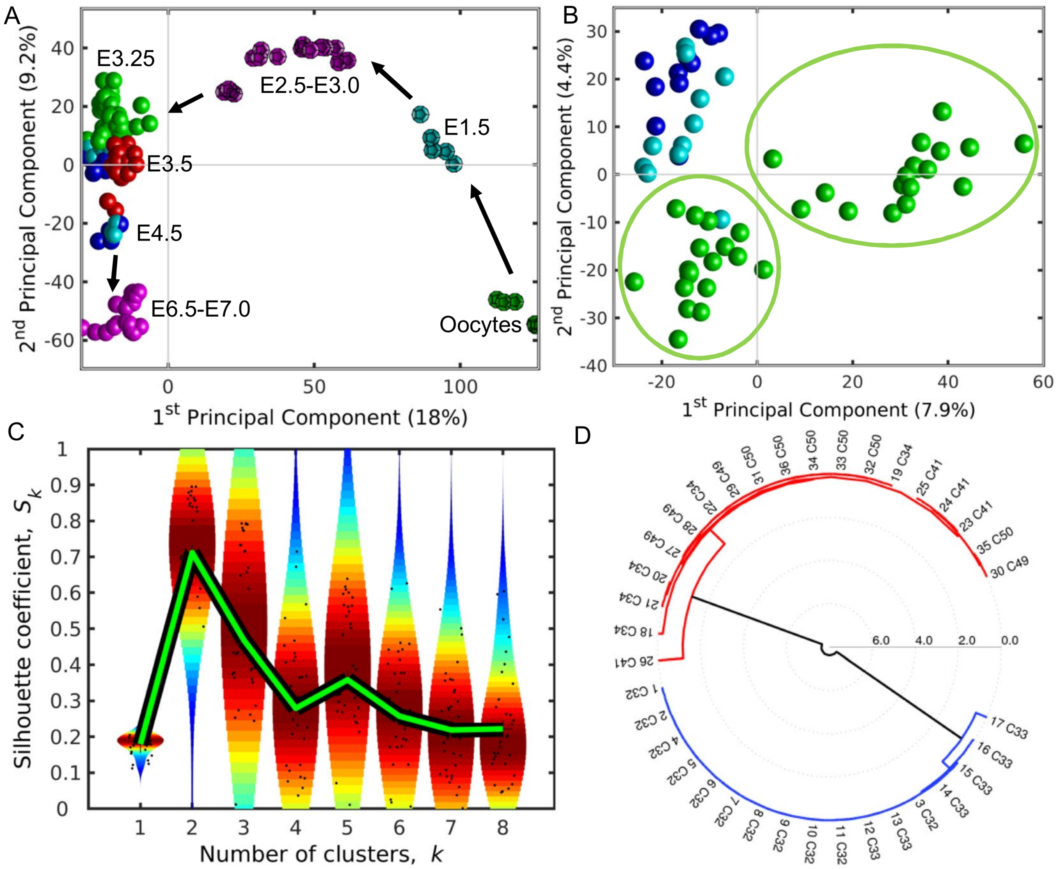


Figure 1. Hierarchical Optimal *k*-Means clustering (HO*k*M) reveals a splitting into two clusters of the gene expression in the ICM of the mouse embryo at E3.25. (**A**) Bi-dimensional Principal Component Analysis (PCA) of transcriptomics data from all *in vivo* wild type samples (Table 1). Dodecahedra and spheres mark bulk and single cells, respectively. Green, cyan and magenta dodecahedra mark bulk samples of oocytes, E1.5 and E2.5-E3.0 cells, respectively. Green, cyan and dark blue spheres mark the E3.25, EPI (E3.5 and E4.5) and PE (E3.5 and E4.5) cells of Ohnishi *et al.*[3], respectively; while the red spheres mark the E3.5 ICM cells of Kurimoto *et al.*[9], and the magenta spheres mark the Prestreak and Early/mid bud cells of Kurimoto *et al.*[56]. The arrows infer the direction of development. (**B**) Bi-dimensional PCA of the single cell transcriptomics data from E3.25 and E3.5 of Ohnishi *et al.*[3]. The green spheres mark the E3.25 cells, while the cyan and dark blue spheres mark the E3.5 EPI and E.3.5 PE cells, respectively. The two green ellipses encircle the two groups of E.25 cells, posteriorly classified by the HO*k*M method and named as E3.25-LNCs and E3.25-HNCs. (**C**) Violin plot of the silhouettes of the HO*k*M trajectories. The green line marks the position of the medium silhouette distributions. (**D**) Dendrogram of the optimal clustering ($k_o = 2$).

E3.25-LNC and E3.25-HNC is not due to factors like same sex of donor embryos, chromosomal mosaicism, or wrong allocation of trophectoderm (TE) cells as ICM, we checked these negative hypotheses.

Firstly, we confirmed that the clustering of the cells at E3.25 is not due to the embryos of the two clusters E3.25-LNC and E3.25-HNC, 2 and 4 donor embryos, respectively, being 2 males and 4 females or vice versa. A way to infer the chromosomal sex (XX, XY) of the embryos is to examine the level of expression of the long non-coding RNA *Xist*, which is low for male and high for female. It was reassuring to find out that the sex of the donors to the two groups was mixed: 1 male (cells 1–11 of the embryo with 32 cells, C32) and 1 female (cells 12–17 of the embryo with 33 cells, C33) for the E3.25-LNC, and 2 male (cells 18, 19, 23–30) and 2 female (cells 20–22, 31–36) E3.25-HNC embryos (Fig. 2A). The Affymetrix Mouse Genome 430 2.0 Array platform has three probes, 1427262_at (*locus* ChrX 100658863-100659290), 1427263_at (*locus* ChrX 100655856-100656302) and 1436936_s_at (*locus* ChrX 100678088-100678555), for *Xist* and *Xist* 1436936_s_at is the most responding to the X chromosome inactivation process.

Secondly, we considered the possibility that the differences of gene expression could be due in part to factors like chromosomal mosaicism, aneuploidy, in which different cells of the same embryo have different karyotypes. We infer possible chromosomal mosaicism under the assumption that a ploidy aberration event affects the expression of the genes located on a chromosome. Therefore, we hypothesized that the statistical enrichment of a chromosome with DEGs between two sets is an indication that the chromosome has a ploidy change. We found that there are some cells with possible ploidy changes in some chromosomes through calculating the $-\log_{10}(p\text{-value}_{\mathrm{ChrEnr}})$ of the enrichment in each chromosome of the DEGs of each E3.25 cell in relation to the

Cellular stage	Number samples	Database	Database ID	Data type	Reference
Metaphase II oocyte 1-cell stage 2-cell stage 4-cell stage 8-cell stage E3.0 morula	3 3 3 3 3 3	GEO	GSE41358	Bulk	Cao *et al.*[53]
E3.25 blastocysts[a] E3.5 PE & EPI[a] E4.5 PE & EPI[a] E3.25 FGF4-KO[a] E3.5 FGF4-KO[a] E4.5 FGF4-KO[a]	36 22 8 17 8 10	Array Express	E-MTAB-1681	Single cell	Ohnishi *et al.*[3]
8-cell stage (E2.5)	4	GEO	GSE41925	Bulk	Choi *et al.*[54]
E3.5 ICM	20	GEO	GSE4309	Single cell	Kurimoto *et al.*[9]
2-cell stage (E1.5) 4- to 8-cell stage (E2.5)	2 2	GEO	GSE7309	Bulk	Maekawa *et al.*[55]
Prestreak (PS) Early/mid bud (E/MB)	8 8	GEO	GSE11128	Single cell	Kurimoto *et al.*[56]
Trophoblast stem cells	6	GEO	GSE47719	Bulk	Kubaczka *et al.*[11]
ESCs	665	GEO	[b]*	[b]*	[b]*

Table 1. Datasets from the Affymetrix Mouse Genome 430 2.0 Array used in the different transcriptomics analyses. [a]The staging of the embryos selected for single cell transcriptomics analysis was calculated as the average total number of cells counted in all the embryos of the same clutch, excluding samples with the maximum and minimum cell numbers. [b]Collection of ESCs (bulk and single cells) from different GSEs and authors.

pool of remaining E3.25 cells (Fig. 2B). Next, we checked whether the distributions of the $-\log_{10}(p\text{-value}_{\text{ChrEnr}})$ between the E3.25-LNC and E3.25-HNC are different, and we found no statistically significant difference between the two distributions, obtaining a $p\text{-value}_{\text{KS}} = 0.808$ for the two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test (Fig. 2C). Therefore, we concluded that the differences of gene expression between the E3.25-LNC and E3.25-HNC are not due to chromosomal mosaicism.

Thirdly, we considered possible wrong allocation of embryo cells to TE and ICM. We compared the E3.25 cells from Ohnishi *et al.*[3] with trophoblast stem (TS) cells from the same platform from Kubaczka *et al.*[11] (Table 1). The PCA (Fig. 2D) shows that all E3.25 cells of Ohnishi *et al.*[3] group far away and differ considerably in their 1st Principal Component from the TS cells. Additionally, the PCA shows the clear separation between the E3.25-HNC and E3.25-LNC. We compared the expression of TE markers, taken from Wu *et al.*[12] and Kubaczka *et al.*[11], in the E3.25 cells of Ohnishi *et al.*[3] and in TS cells of Kubaczka *et al.*[11], and we did not find any evidence for any of the E3.25 cells to have been mis-assigned to ICM (Fig. 2E).

Previously, Ohnishi *et al.*[3] studied the possibility of a batch effect in their sc transcriptomics data, however the results were inconclusive. Therefore, we assume that the differences in transcriptomics expression between the E3.25-LNC and E3.25-HNC are not due to batch effects.

Cell junction and plasma membrane *Bsg*, *Ctnnb*1 and *Fgfr*1 genes are highly expressed in E3.25-HNCs.

Next, we searched for the specific genes driving the E3.25 ICM optimal clustering performing a Differentially Expressed Genes (DEGs) analysis between E3.25-LNC and E3.25-HNC. We found 399 DEGs for the highly expressed transcripts in E3.25-HNC (HNC-h-DEGs), the top-ranked depicted in Fig. 3A, and the complete set in Fig. S3. The top-ranked HNC-h-DEG is basigin *Bsg/CD147/EMMPRIN* (Fig. 3A,B), known to regulate the canonical Wnt/beta-catenin signalling pathway[13] and thought to be regulated by hypoxia[14]. The fourth top-ranked HNC-h-DEG is the key mediator of the Wnt pathway *Ctnnb1* (*β-catenin*) (Fig. 3A,B), which, together with the signal transducer *Ywhaz*, is one of the main hubs of the protein binary interaction network of the HNC-h-DEGs, the major one being Oct4 (Fig. 4A). Among the top-ranked HNC-h-DEGs, 15th, is *Fgfr1*, required for the lineage establishment and progression within the ICM[15] (Fig. 3A,B). Interestingly, among the top-ranked HNC-h-DEGs is *Gapdh*, assumed widely and in Ohnishi *et al.*[3] as a housekeeping gene. The expression of all 5 probes of the Affymetrix Mouse Genome 430 2.0 Array platform of *Gapdh* are given in Fig. S4. The probe that we detected as HNC-h-DEG is AFFX-GapdhMur/M32599_M_at, while probe AFFX-GapdhMur/M32599_3_at is the one that behaves as a housekeeping gene.

The expression of *Oct4* and several chromatin remodels is stabilized at high level in E3.25-HNC.

The Gene Ontology (GO) analysis of the HNC-h-DEGs revealed that among statistically significantly enriched GO terms are chromatin remodellers such as the INO80 and the SWI/SNF complex, and cell-cell interaction terms such as adherent junction, focal adhesion and bi-cellular tight junction (Fig. 4B). A detailed list of all found GO terms (Fig. S6) and their corresponding genes are provided in Tables S1–3. The HNC-h-DEGs involved in chromatin-remodelling complexes, together with their roles, are enlisted in Table S4 and annotated in Figs 3A and S3. Among the top-ranked DEGs, we discovered one of the pluripotency regulation masters: *Pou5f1* (*Oct4*), which has much lower expression in E3.25-LNCs than in E3.25-HNCs. Though *Pou5f1* has some

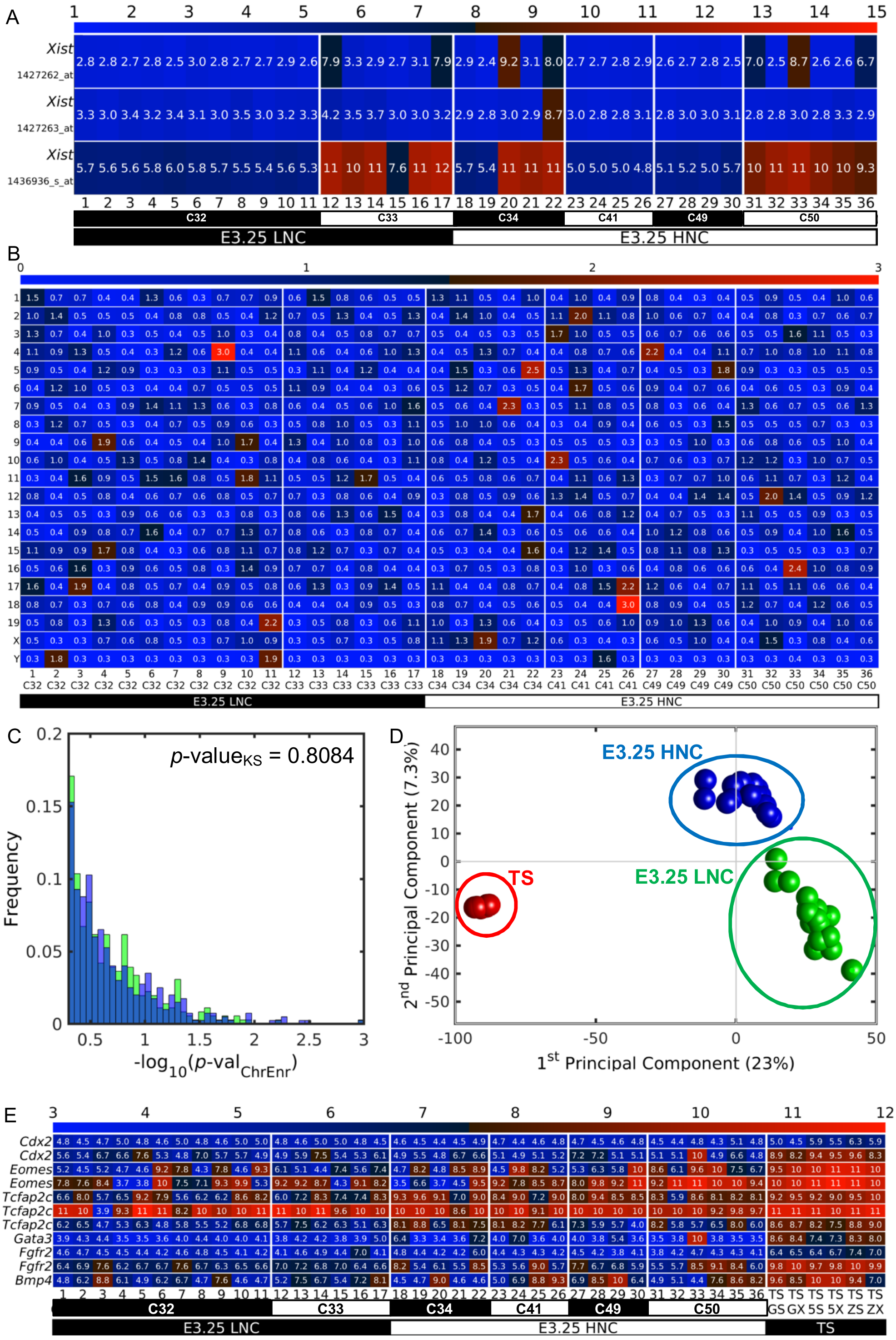


Figure 2. The ICM split at E3.25 into E3.25-LNC and E3.25-HNC is not due to sex, karyotype aberration or mis-assignment to ICM. (**A**) Heat map of the expression of the three probes targeting the long non-coding RNA *Xist* in the single cells from E3.25. The colour bar codifies the gene expression in $\log_2$ scale. Higher gene expression corresponds to redder colour. (**B**) Heat map of the -log10(p-value$_{\text{ChrEnr}}$) of the statistical significance of the enrichment of each chromosome of the DEGs, obtained by the Jack-knife method, between each E3.25 single cell and the pool of the remaining E3.25 cells. The colour bar codifies the $-\log10$(p-value$_{\text{ChrEnr}}$). Higher $-\log10$(p-value$_{\text{ChrEnr}}$) corresponds to redder colour. (**C**) Histogram of the distributions of the -$\log_{10}$(p-value$_{\text{ChrEnr}}$) between the E3.25-LNC (green bars) and E3.25-HNC (blue bars). Over-imposed is the p-value$_{\text{KS}}$ = 0.808 for the two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test between for the two distributions. (**D**) PCA of all E3.25 cells of Ohnishi *et al.*[3] and the TS cells of Kubaczka *et al.*[11] (Table 1). The E3.25 LNC, E3.25 HNC and TS cells are marked with green, blue and red spheres. (**E**) Heat map of the expression of the probes targeting TE markers in the single cells from E3.25 of Ohnishi *et al.*[3], and in the TS cells of Kubaczka *et al.*[11],

namely TS GS (EGFP cell line in TS medium), TS GX (TS EGFP cell line in TX medium), TS 5S (TS L5 cell line in TS medium), TS 5 × (TS L5 cell line in TX medium), TS ZS (TS LaCZ cell line in TS medium), TS ZX (TS LaCZ cell line in TX medium). The colour bar codifies the gene expression in $\log_2$ scale. Higher gene expression corresponds to redder colour.

oscillatory expression spikes characteristic of a "salt and pepper" expression pattern (Fig. 3A), it discriminates very well the E3.25-LNC and E3.25-HNC populations (Fig. 3B). In all E3.25-HNCs, *Oct4* is stabilized at very high expression level. The violin plots of the expression of *Oct4* across all early developmental stages (Fig. S5) reveal a continuously growing expression since oocyte until the 8-cell stage, which is perturbed to lower expression in 32- and 33-cell embryos in order to come back again to high expression during the following developmental stages. The expressions of the *Oct4* common regulators *Sox2* and *Nanog* correlate slightly with that of *Oct4* (Fig. S5). At E3.25 *Nanog* shows the previously reported "salt and pepper" expression pattern, with a small population following the low expression of *Oct4* in the 33- and 34-cell embryos, whereas the expression of *Sox2* remains almost continuously high, with exception of some cells from 32-, 41-, and 50-cell embryos (Fig. S5). The enrichment of pluripotency genes among the HNC-h-DEGs prompted an elucidation of the strength of the pluripotency network in the E3.25-HNCs through an intersection of the HNC-h-DEGs with the interactomes of the master regulator of pluripotency Oct4 and the main pluripotency players (Sox2 and Nanog) (see Methods). Among the proteins shared by the Oct4-Sox2 interactome and the HNC-h-DEGs, we found four genes (Fig. S7): *Cct3, Hnrnpu, Parp1*, and *Ssrp1*, whose function is described in Table S5. Particularly, *Hnrnpu* and *Parp1* are *Oct4* transcriptional regulators[16,17]. Among the proteins shared by the Oct4-Nanog interactome and the HNC-h-DEGs, we found four common genes (Fig. S7): *Arid1a*, *Mbd3*, *Msh6* and *Tet1*, all of them chromatin remodellers (Table S4).

To check the distribution of the HNC-h-DEGs across the chromosomes, we performed chromosomal landscaping analysis. We discovered that chromosome 17, on which *Oct4* is located, is statistically significant enriched by HNC-h-DEGs (Fig. 4C). To visualize the distribution of the HNC-h-DEGs along chromosome 17, we zoomed-in and performed clustering of the HNC-h-DEG *loci*. We found three major clusters in the A3.3, B1 and E4 cytobands (Fig. 4D). The cluster in the A3.3 cytoband contains the phosphoglycolate phosphatase *Pgp*; the transmembrane protein *Tmem8*; the serine/arginine-rich splicing factor 3 *Srsf3*, and the Zinc finger *Zfand3 loci*. The cluster in the B1 cytoband contains the members of the histocompatibility 2 complex, *H2-D1*, *H2-K1;* the zinc transporter from the endoplasmic reticulum/Golgi apparatus to the cytosol *Slc39a7*, the epigenetic regulator *Ehmt2*; the member of the LSm family of RNA-binding proteins *Lsm2*; and the master regulator of pluripotency *Pou5f1 loci*. The cluster in the E4 cytoband contains the transcription factor *Zfp36l2*; the potential transcriptional regulator *Lrpprc*; the serine peptidase *Prepl*; the calmodulin gene family member *Calm2;* the epithelial cell adhesion molecule *Epcam* that plays a role in ESCs proliferation and differentiation[18]; the chromatin remodellers *Msh2* and *Msh6*, and the member of the ubiquitin protein ligase complex SCF (SKP1-cullin-F-box) *Fbxo11 loci*. Among the genes in non-clustered regions are *Tgif1* that counterbalances the activity of core pluripotency factors, Oct4, Sox2, and Nanog, in mouse ESCs[19], and the chromatin modifier *Satb1* that regulates cell fate through *Fgf* signalling in the early mouse embryo[20].

Fgf4, *Sox2* and *Dvl1* segregate into two different trajectories before E3.25-LNC.

We checked for the transitions of the transcriptomics dynamics through a bifurcation analysis from E3.25 to E4.5 stage (E3.25-LNC, E3.25-HNC, E3.5, E4.5), and determined the first genes associated to dynamic segregation. Such analysis was based on clustering the gene expression of the most highly variable expressed genes across each embryonic day (see Methods). We classified the gene trajectories into 15 categories based on the number of bifurcations of expression for each of the four embryonic stages in order to determine the moment of segregation. The four most interesting trajectory categories are: (2-2-2-2) with two bifurcation points at each of the 4 embryonic stages that corresponds to a very early segregation time (before E3.25-LNC), (1-2-2-2) that corresponds to an intermediate segregation time E3.25-HNC, (1-1-2-2) with no bifurcation before E3.5 but bifurcating at E3.5 and remaining bifurcated till E4.5, and (1-1-1-2) corresponds to late segregation time E4.5 (Fig. 5).

Interestingly, category (2-2-2-2) with bifurcation time before E3.25-LNC is the most populated with 2271 genes (Fig. 5A). This indicates that the ICM cells that eventually become PE and EPI progenitors at E4.5 carry considerable heterogeneity from before E3.25-LNC. The topmost gene, with the widest bifurcation from this category, is the EPI marker *Fgf4*, with two microarray probes among the 12 top-ranked transcripts (Fig. 5B). The violin plots of the gene expression distribution of *Fgf4* across all the preimplantation developmental stages (Fig. S5) confirm that *Fgf4* shows a splitting pattern ever since the 2-cell stage. Among the top genes in the list is also the *Pou5f1* partner *Sox2*, the member of the Dishevelled family of proteins which regulate both canonical and non-canonical (planar cell polarity pathway) Wnt signalling *Dvl1*, and the PE lineage marker *Cubilin* (*Cubn*).

E3.25-HNC embryos are more developed than E3.25-LNC.

The PCA suggested that E3.25-LNC are more similar to earlier than E3.25 developmental stages, while E3.25-HNC are more similar to later than E3.25 developmental stages (Fig. 1B). To confirm that, we used as representative members of each cluster their top-ranked DEGs (E3.25 HNC-h-DEGs with heat map in Fig. S3; and E3.25 LNC-h-DEGs with heat map in Fig. S2), and we analysed their transcriptomics distributions across other developmental stages using violin plots (Fig. 6A,B). All the stages before the 33-cell embryos are dissimilar in terms of the distribution spread of the E3.25 HNC-h-DEGs, with more dispersed distributions than the distributions after the 34-cell stage. The transcriptomics distributions of the E3.25 HNC-h-DEGs after the 34-cell stage are similar both in terms of median values and spread (Fig. 6A). The violin plots of the top E3.25 LNC-h-DEGs (Fig. 6B) depict their transcriptomics

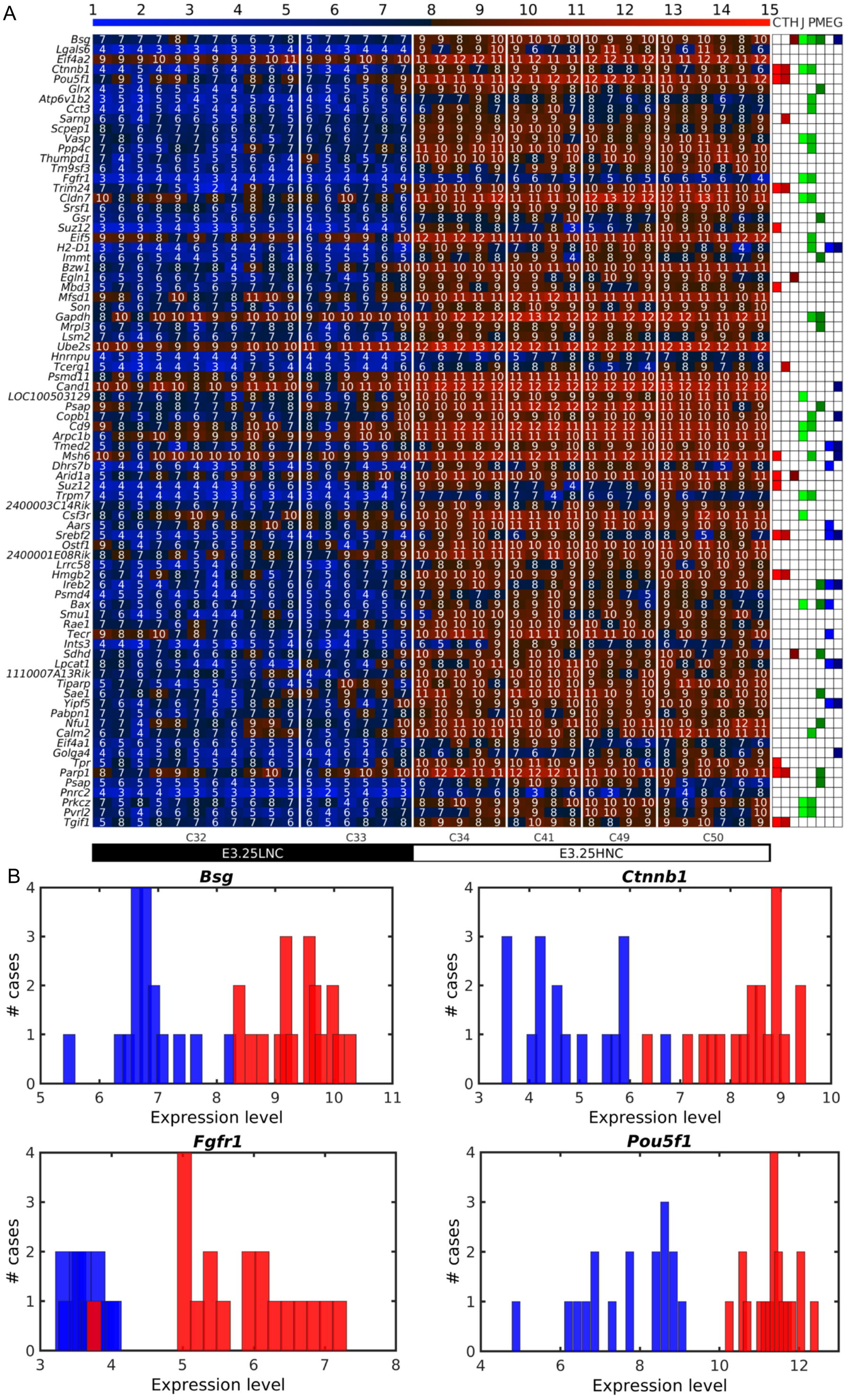


Figure 3. Expression of *Oct4* and several chromatin remodellers is stabilized at high level in E3.25-HNCs. (**A**) Heat map of the expression of the 80 top-ranked E3.25 HNC-h-DEGs in decreasing order of significance. The colour bar codifies the gene expression in $\log_2$ scale. Higher gene expression corresponds to redder colour. The table to the right annotates GO terms: C (Chromatin remodellers), T (Transcription factor activity), H (Hypoxia), J (Cell junction), P (Plasma membrane), M (Mitochondrion), E (Endoplasmic reticulum), G (Golgi apparatus). (**B**). Histograms showing the ability of the top-ranked HNC-h-DEGs (*Bsg*, *Ctnnb1*, *Fgfr1* and *Oct4*) to discriminate between the E3.25-LNC (blue bars) and E3.25-HNC (red bars) populations.

distributions as unique and not similar to any of the distributions associated to other developmental stages, being slightly closer to the 4-cell stage rather than to the 8-cell stage, as it might be expected. Thus, E3.25-HNCs are more related to later developmental stages than the E3.25-LNCs.

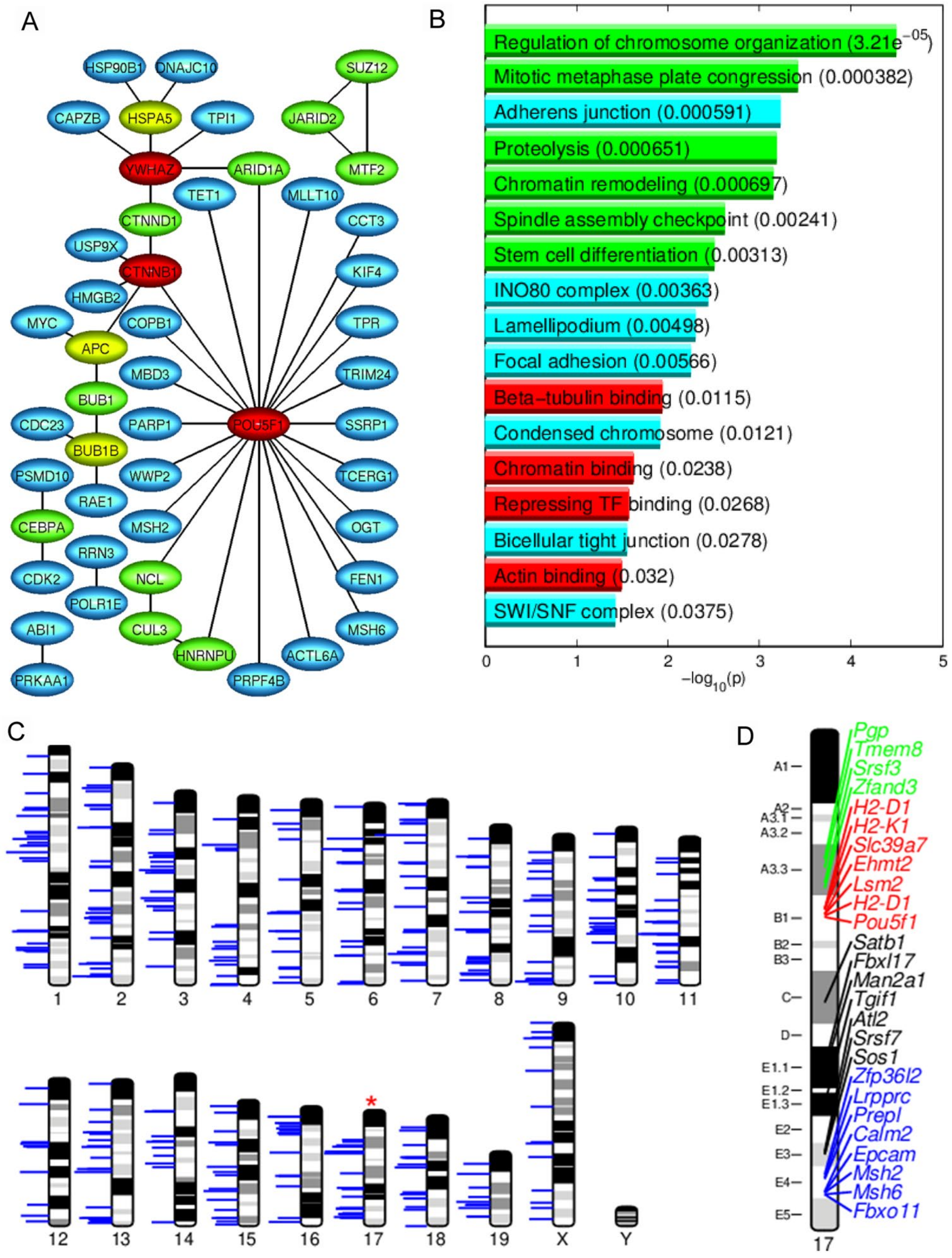


Figure 4. *Oct4* plays a central role in the network of the E3.25 HNC-h-DEGs. (**A**) Protein binary interaction network of the HNC-h-DEGs. The node colour codifies incidence number (blue, green, yellow and red for incidences 1, 2, 3 and more than 4, respectively). (**B**) Bar plot of the -$\log_{10}$(p-value) of the significant enriched GO terms of HNC-h-DEGs. Longer bars correspond to higher statistical significance of the enrichment (p-values inside parentheses). The red, green and cyan bars correspond with molecular function, biological process and cellular compartment GO terms, respectively. (**C**) Chromosomal landscaping of the HNC-h-DEGs. The horizontal blue lines mark the *loci* of the HNC-h-DEGs, and their length is proportional to the average level of expression of each HNC-h-DEG across all the HNCs. The red asterisk marks the chromosome with statistically significant enrichment of HNC-h-DEGs, hypergeometric distribution p-value $< \alpha_{LAN} = 0.01$. (**D**) Zoom-in on chromosome 17, marking the HNC-h-DEG gene names and their *loci*. The green, red and blue labels mark the clusters on the A3.3, B1 and E4 cytobands, the black labels mark the *loci* that did not pass the criterion for uni-dimensional clustering of *loci*.

Interestingly, E3.25-HNCs are dissimilar to both PE and EPI (Fig. 6A) based on a comparison of the expression of the E3.25-HNC-DEGs. Anyway, we checked for a possible global similarity of E3.25-HNCs with EPI or PE using all transcripts. We made scatter plots, calculating the average pool of all the cells of E3.25-LNC, E3.25-HNC, E3.5 PE and E3.5 EPI. The split of E3.25-LNC and E3.25-HNC is preserved after performing the average of the respective single cells (Fig. 6C). They have, not very high for transcriptomics studies, correlation coefficient $\rho = 0.970$ and 464 DEGs. Both E3.25-HNC and E3.25-LNC express in a similar way the *Fgf4* and *Nanog*

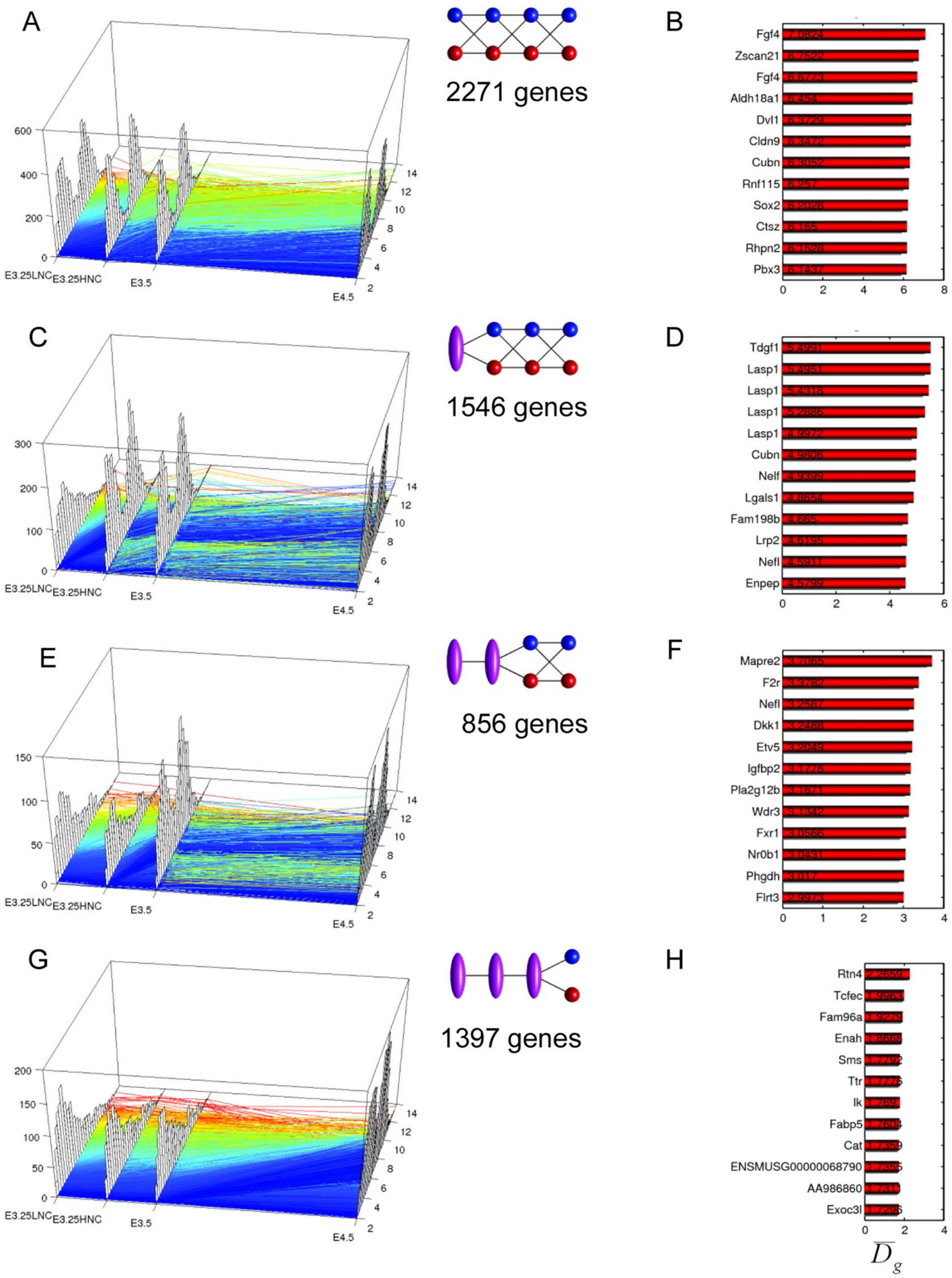


Figure 5. Dynamics of the transcriptomics bifurcations across E3.25-LNC, E3.25-HNC, E3.5 and E4.5. (**A**,**C**,**E**,**G**) Transcriptomics trajectories ordered from earlier to later bifurcation decision. Each trajectory colour corresponds to the level of expression at the earliest time of the trajectory. Bluer colour corresponds to lower expression at the earliest measured time. The histogram of the transcriptomics level of the trajectories is shown on the *z*-axis at each embryonic day. The two-point passing stages represent a bifurcation of the gene trajectory at such stage, whereas the one-point passing stage represents a continuous distribution of states that cannot be segregated into two modes. (**B**,**D**,**F**,**H**) Top 12 transcripts with the maximum separation of the bifurcation of their corresponding trajectories, defined by the difference $\overline{D}_g$ of the bifurcations means.

EPI markers, and the *Gata4 and Sox17* PE markers. E3.25-HNCs are closer in terms of correlation coefficient ρ and number of DEGs to both PE and EPI E3.5 populations, compared to the E3.25-LNCs (Fig. 6F,G). It is noteworthy that E3.25-HNCs are close to both PE (ρ = 0.975) and EPI (ρ = 0.980) but not to any of the two lineages

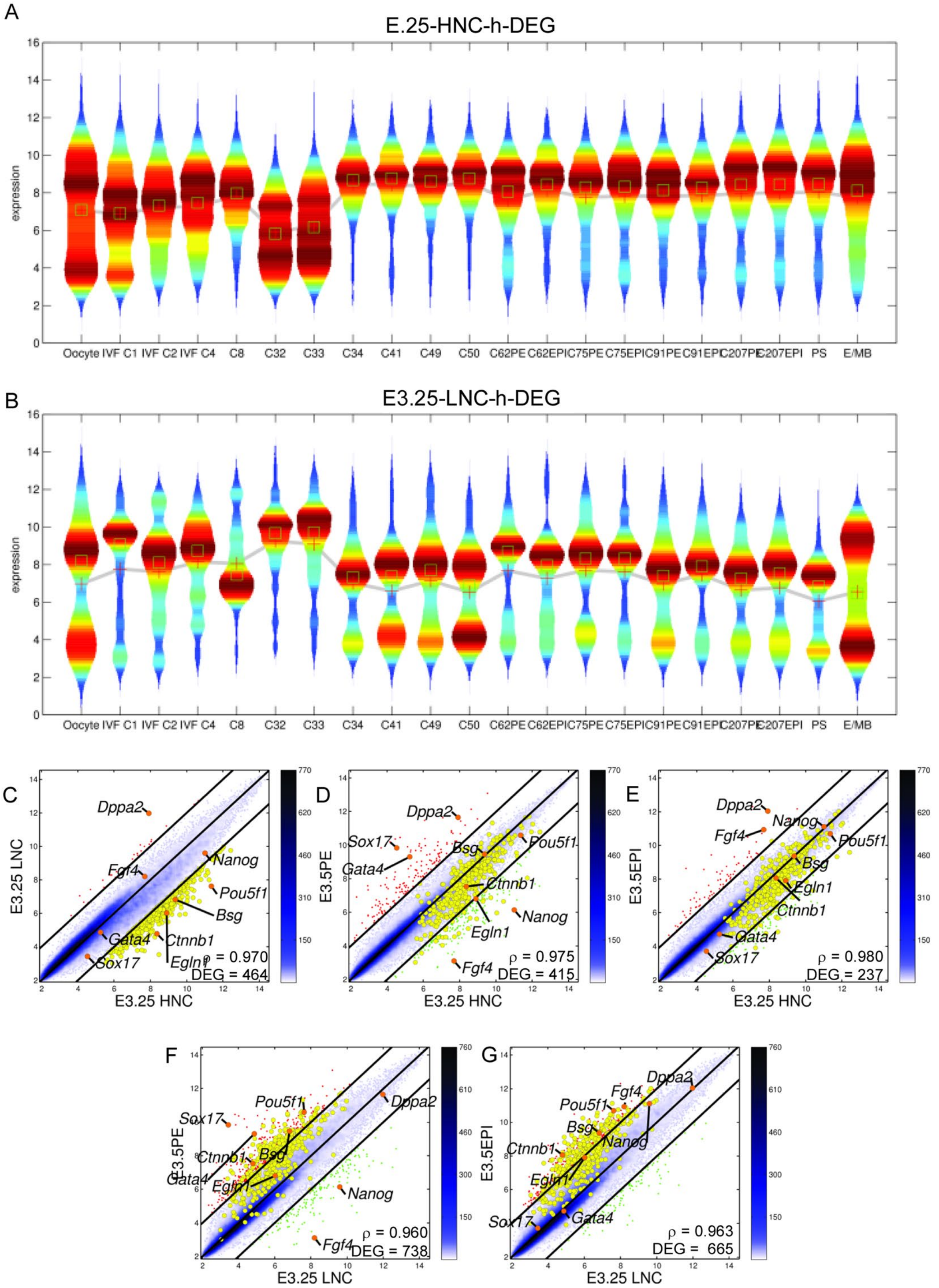


Figure 6. E3.25-HNCs are more developed than E3.25-LNCs. Violin plots of the distribution of the expression of (**A**) E3.25 HNC-h-DEGs and (**B**) E3.25 LNC-h-DEGs across different developmental stages. The mean and median are shown as red crosses and green squares, respectively. Pairwise scatter plots of (**C**) E3.25-LNC *vs* E3.25-HNC, (**D**) E3.5-PE *vs* E3.25-HNC, (**E**) E3.5-EPI *vs* E3.25-HNC, (**F**) E3.5-PE *vs* E3.25-LNC, (**G**) E3.5-EPI *vs* E3.25-LNC. The black lines are the boundaries of the 2-fold changes in gene expression levels between the paired samples. Transcripts up-regulated in ordinate samples compared with abscissa samples, are shown with red dots; those down-regulated, with green. The positions of some markers are shown as orange dots. The colour bar indicates the scattering density. Darker blue colour corresponds to higher scattering density. The transcript expression levels are $\log_2$ scaled. ρ is the Pearson's correlation coefficient. The E3.25 HNC-h-DEGs are over-imposed as yellow dots.

specifically. Thus, the higher transcriptomics similarity of the E3.25-HNC embryos with older-than-E3.25 embryos is not related to the cell fate but rather to the developmental stage, early or late.

E3.25 HNC-h-DEGs share transcriptomics features with E3.5 *Fgf4*-KO. Fgf4 is the main determinant of the EPI - PE segregation[21–23]. In an attempt to see whether the set of genes defining the HNC cluster, E3.25 HNC-h-DEGs, is related in some way to the PE and EPI cell specification in the ICM, we compared the E3.25 HNC-h-DEGs with *Fgf4* knockouts (*Fgf4*-KO) from Ohnishi *et al.*[3] (Fig. S8). We found that the HNC-h-DEGs are lowly expressed in E3.25 *Fgf4*-KO and E4.5 *Fgf4*-KO cells, however they are highly expressed in E3.5 *Fgf4*-KO, thus, the transcriptomics profile of E3.5 *Fgf4*-KO resembles the profile of E3.25-HNCs. For the subset of E3.25 HNC-h-DEGs, the E3.25 *Fgf4*-KO cells are similar to the E3.25-LNCs. This suggests that both E3.25 *Fgf4*-KO and E3.25-LNCs have still undetermined ICM fate. The E3.25 *Fgf4*-KOs originate from low and high number-of-cell embryos, which hints, that lack of Fgf4 prevents inner cells from even high-number-of-cell embryos (E3.25 *Fgf4*-KO with >33 cells) to take secure ICM fate, and holds back the inner cells to the undetermined ICM cell-fate stage (E3.25 with <34 inner cells). Interestingly, the effect of the *Fgf4*-KO at E4.5 reverts the transcriptomics profile based on the subset of the E3.25 HNC-h-DEGs to the E3.25-LNC state, whereas the E3.5 *Fgf4*-KO preserves the E3.25-HNC transcriptomics state.

Additionally, we followed the evolution of the transcriptomics profiles based on the subset of E3.25 HNC-h-DEGs in the PE and EPI cells at E3.5 and E4.5, and compared them with the *Fgf4*-KOs. Three groups of cells (E3.5PE, E3.5EPI and E3.5 *Fgf4*-KO) have similar transcriptomics profiles based on the subset of E3.25-HNC-h-DEGs (Fig. S8A). The similarity between E3.5 PE and E3.5 EPI is noteworthy, and can be explained by the fact that at E3.5 the PE and EPI cells can still revert their fate to the opposite fate. The subset of E3.25 HNC-h-DEGs is related to a chromatin remodelling process that pushes forward the totipotent state (E3.25-LNC) to a more advanced pluripotent state (E3.25-HNC and E3.5 (PE + EPI)) and it seems that *Fgf4*-KO does not affect this process. This is confirmed by the absence of *Fgf4* in the E3.25 HNC-h-DEGs and by the very early bifurcation of *Fgf4* before E3.25. Strikingly, at E4.5 the profile of the E3.25 HNC-h-DEGs in the *Fgf4*-KO cells is similar to the profile of the E3.25-LNC, even though the embryos at E4.5 have very high number of cells, much higher than 33, pointing that Fgf4 affects the machinery that drives E3.25-LNC to more advanced/differentiated pluripotent state in the HNC at E3.25 and E4.5 but not at E3.5.

E3.25-HNCs are more enriched in PE and EPI markers than E3.25-LNCs. To study whether the E3.25-LNCs and E3.25-HNCs differ in terms of lineage differentiation, we predicted early PE and EPI lineage markers from the consensus of the E3.5 transcriptomics single-cell datasets of Ohnishi *et al.*[3], E3.5O and E4.5O, and Kurimoto *et al.*[9], E3.5K. Since the Kurimoto *et al.*[9] E3.5 ICM dataset was not initially classified into PE and EPI, we applied the HO*k*M algorithm to classify it. We obtained that the optimal number of clusters is $k = 2$, with one cluster comprising cells 3–5, 8, 10, 11, 15, 17 and 20, and another comprising cells 1, 2, 6, 7, 9, 12–14, 18 and 19 (Fig. 7A). For validation purposes, we applied the HO*k*M algorithm to the already classified PE and EPI cells from E3.5 and E4.5 of Ohnishi *et al.*[3]. Even though the HO*k*M algorithm is an unsupervised clustering method, it classified with 100% accuracy the two EPI - PE datasets, namely E3.5O and E4.5O (Fig. 7B,C).

After the correct EPI - PE classification, we obtained the consensus shared markers. We searched for the statistically significant DEGs between the EPI *vs* PE clusters of each dataset, and performed the intersections of the corresponding lists of estimated PE and EPI markers (Fig. 7D,E). Among the genes shared by the three common PE intersections are the known markers *Sox17*, *Gata6* and *Cubn*. Among the E3.5O ∩ E3.5K *exclusive* PE intersection, corresponding to earlier markers, are known markers such as a second probe of *Sox17*, and *Fgfr2*, but also transcripts such as two probes of *LOC640502* that correspond to *Uap1* (UDP-N-acetylglucosamine pyrophosphorylase 1) and that has not been reported as a PE marker. Among the genes of the intersection of the three EPI sets, is the known marker *Fgf4* (with two probes). Among the E3.5O ∩ E3.5K *exclusive* EPI intersection is the known marker *Nanog*.

The only E3.5-specific PE and EPI markers among the HNC-h-DEGs are *Cpne8* and *Ubxn2a*, respectively. The E3.5-specific EPI marker[24] Ubxn2a (UBX domain protein 2A)/Ubxd4, regulates the cell surface number and stability of alpha3-containing nicotinic acetylcholine receptors[25] and is located in both the endoplasmic reticulum and cis-Golgi compartments interacting with the ubiquitin-proteasome system. The calcium-dependent membrane-binding phospholipid protein Cpne8 may regulate molecular events at the interface of the cell membrane playing a role in membrane trafficking.

We built a 2D map of the EPI - PE space based on *consensus* predicted markers, assigning to each cell the transcriptomics mean across all PE and EPI markers (Fig. 7F). The wild type PEs locate as high PE - low EPI. E3.25-LNCs and E3.25-HNCs are mainly high EPI - low PE, where E3.25-HNCs are with slightly higher PE and EPI expression. In the *Fgf4*-KO case, E3.25 KOs are depleted in both PE and EPI markers (low PE - low EPI), while E3.5 KOs and E4.5 KOs are low PE-high EPI. The violin plots showing the variability of the expression of the PE and EPI markers for each cell type, confirm that the E3.25-HNC express slightly higher both PE and EPI markers than the E3.25-LNC, while the *Fgf4*-KOs express scarcely PE markers and highly EPI markers at E3.5 and E4.5 but not at the early E3.25 stage (Fig. 7G,H).

E3.25-HNC transcriptomics profiles are closest to 2i + LIF ESCs compared to all other ESCs. To assess the pluripotency fingerprint of the E3.25-E4.5 inner cells of Ohnishi *et al.*[3], we compared their transcriptomics profiles with those of all the available ESCs in the GEO database from Affymetrix Mouse Genome 430 2.0 Arrays. The *in vitro* ESCs cluster away from the *in vivo* E3.25-E4.5 inner cells (Fig. 8A). To find the closest *in vitro* pluripotent counterparts of each developmental stage from E3.25 to E4.5, we averaged the single-cell transcriptomics profiles of each stage, and calculated the distance to each ESC profile.

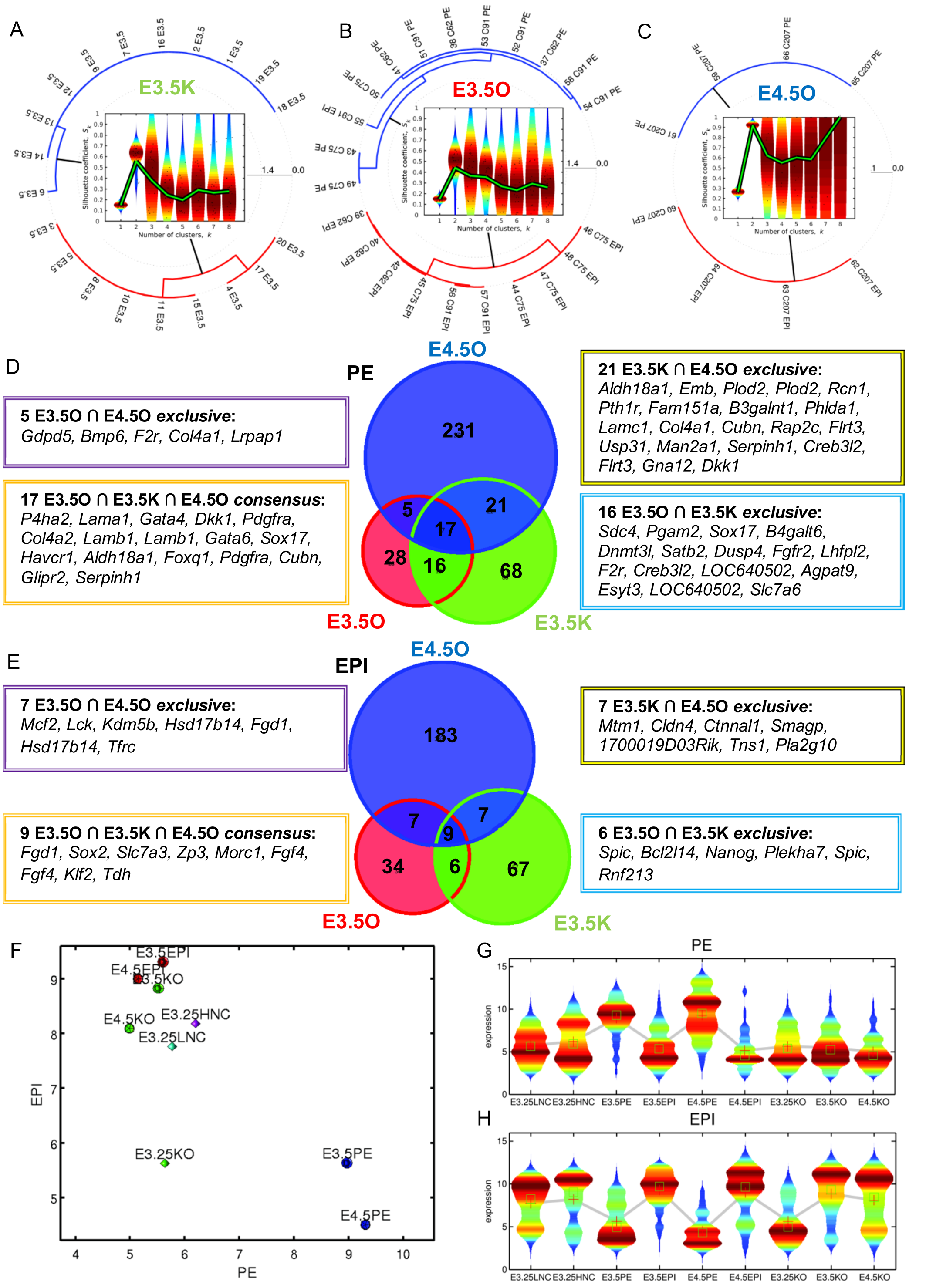
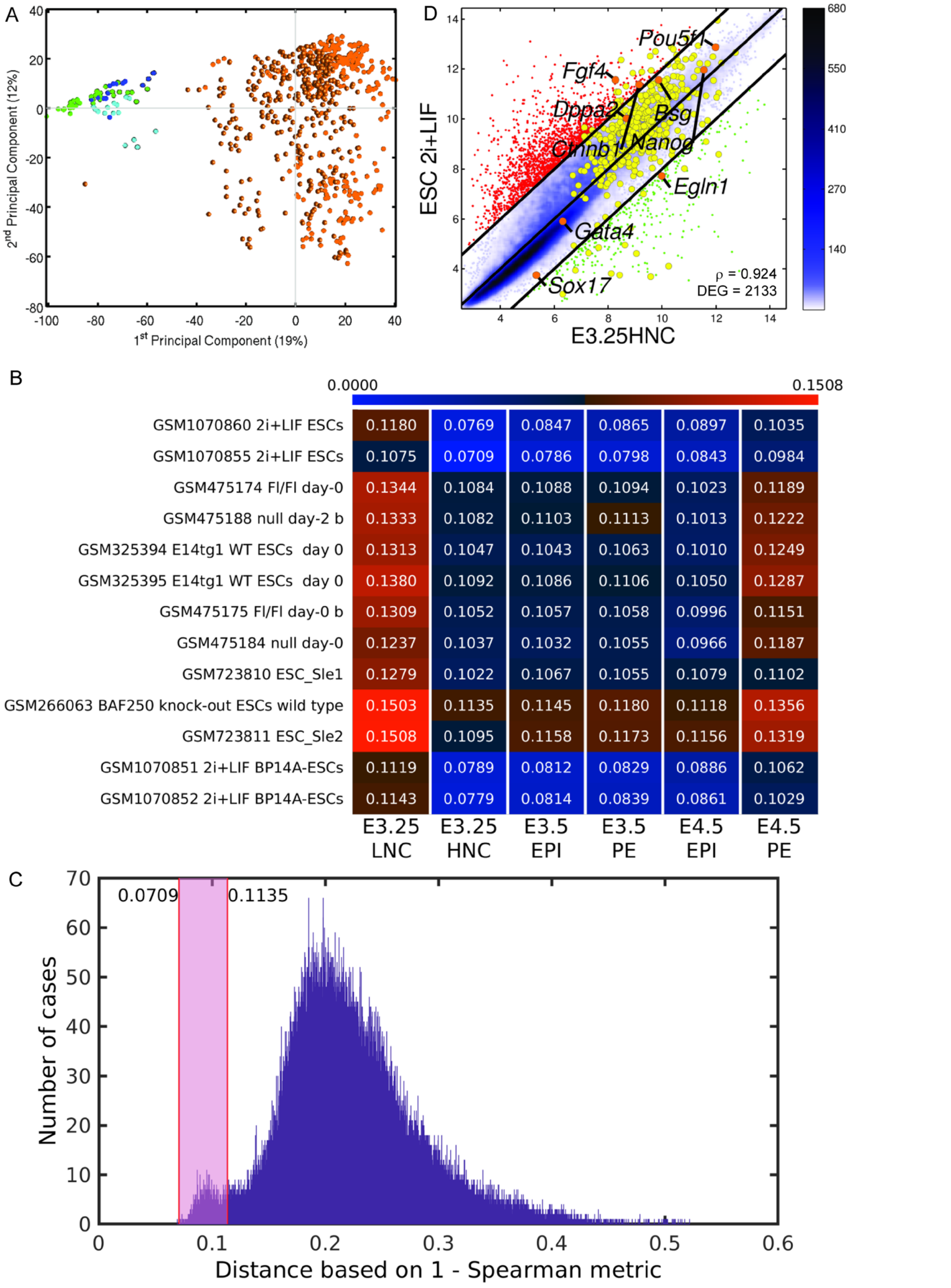


Figure 7. Map of samples on the EPI *vs* PE space built from the consensus of EPI and PE markers predicted through HO*k*M. Polar dendrograms of the HO*k*M of (**A**) the E3.5 data from Kurimoto *et al.*[9] E3.5K(urimoto), and from Ohnishi *et al.*[3] (**B**) E3.5O(hnishi) (**C**) E4.5O(hnishi). Violin plots of the silhouettes of the HO*k*M trajectories are presented in the centre of each dendrogram. The over-imposed green line in the violin plots marks the position of the medium silhouette distributions. Euler-Venn diagrams of the transcripts shared by the (**D**) PE and (**E**) EPI populations. (**F**) Map of single-cell transcriptomics data on the EPI - PE space. Violin plots of the distributions of the consensus (**G**) PE and (**H**) EPI markers. The mean and median are shown as red crosses and green squares, respectively.

The shortest distance between the set of E3.25-E4.5 inner cells and the ESCs is between the E3.25-HNCs and 2i + LIF ESCs (Fig. 8B). The developmental transition from 33 to 34 cells (E3.25-LNC to E3.25-HNC) decreases dramatically the distance to the naïve 2i + LIF ESC. To analyse the significance of these small



	E3.25 LNC	E3.25 HNC	E3.5 EPI	E3.5 PE	E4.5 EPI	E4.5 PE
GSM1070860 2i+LIF ESCs	0.1180	0.0769	0.0847	0.0865	0.0897	0.1035
GSM1070855 2i+LIF ESCs	0.1075	0.0709	0.0786	0.0798	0.0843	0.0984
GSM475174 Fl/Fl day-0	0.1344	0.1084	0.1088	0.1094	0.1023	0.1189
GSM475188 null day-2 b	0.1333	0.1082	0.1103	0.1113	0.1013	0.1222
GSM325394 E14tg1 WT ESCs day 0	0.1313	0.1047	0.1043	0.1063	0.1010	0.1249
GSM325395 E14tg1 WT ESCs day 0	0.1380	0.1092	0.1086	0.1106	0.1050	0.1287
GSM475175 Fl/Fl day-0 b	0.1309	0.1052	0.1057	0.1058	0.0996	0.1151
GSM475184 null day-0	0.1237	0.1037	0.1032	0.1055	0.0966	0.1187
GSM723810 ESC_Sle1	0.1279	0.1022	0.1067	0.1055	0.1079	0.1102
GSM266063 BAF250 knock-out ESCs wild type	0.1503	0.1135	0.1145	0.1180	0.1118	0.1356
GSM723811 ESC_Sle2	0.1508	0.1095	0.1158	0.1173	0.1156	0.1319
GSM1070851 2i+LIF BP14A-ESCs	0.1119	0.0789	0.0812	0.0829	0.0886	0.1062
GSM1070852 2i+LIF BP14A-ESCs	0.1143	0.0779	0.0814	0.0839	0.0861	0.1029

Figure 8. Comparison of E3.25-E4.5 stage inner cells (Ohnishi *et al.*[3]) with all existing ESC Affymetrix Mouse Genome 430 2.0 Array transcriptomics profiles. (**A**) PCA. Green circles and dodecahedra mark E3.25 LNCs and HNCs, respectively. Cyan circles and dodecahedra mark E3.5 and E4.5 PE, respectively. Blue circles and dodecahedra mark E3.5 and E4.5 EPI, respectively. Orange circles mark ESCs. (**B**) Heat map of the 1 - Spearman correlation distances of the top-closest ESC to the different pre-implantation developmental stages. (**C**) Histogram of the distribution of the distance between the transcriptomics profiles of each of the E3.25–E4.5 single cells of Ohnishi *et al.*[3] and each one of all of the ESC Affymetrix Mouse Genome 430 2.0 Array transcriptomics profiles. The shadowed rectangle marks the range of distances between the top-closest ESC and E3.25 HNC. (**D**) Pairwise scatter plots of ESC 2i + LIF *vs* E3.25-HNC. The black lines are the boundaries of the 2-fold changes in gene expression levels between the paired samples. Transcripts up-regulated in ordinate samples compared with abscissa samples, are shown with red dots; those down-regulated, with green. The positions of some markers are shown as orange dots. The colour bar indicates the scattering density. Darker blue colour corresponds to higher scattering density. The expression levels are $\log_2$ scaled. ρ is the Pearson's correlation coefficient. The E3.25 HNC-h-DEGs are over-imposed as yellow dots.

distances, we calculated the empirical distribution of the distance between the transcriptomics profiles of each of the E3.25–E4.5 single cells of Ohnishi *et al.*[3] and each one of the ESC Affymetrix Mouse Genome 430 2.0 Array transcriptomics profiles (Fig. 8C). The range of distances between the transcriptomics profiles of the top-closest ESC and the E3.25-HNCs encompasses a small mode associated to enrichment of short distances. Most of the HNC-h-DEGs are similarly expressed in the E3.25-HNCs and the 2i + LIF ESCs (Fig. 8D). The egl-9 family hypoxia-inducible factor 1, *Egln1*, is among the few HNC-h-DEGs down-regulated in the 2i + LIF ESCs (Fig. 8D). Previous sc transcriptomics studies on identifying the closest counterpart of ESCs in the early embryo[26,27] found that the closest counterpart of their 2i-LIF ESCs was the E4.25 EPI. However, the qRT-PCR study of Boroviak *et al.*[26] is based on profiling of only 96 genes and does not cover the E3.25 embryo stage (only E1.5, E2.5, E3.5, E4.0, E4.5, E5.5). Our study is the first that looks for such ESC counterpart at E3.25; moreover, we use the whole transcriptomics genome to make the study. Semrau *et al.*[27] also found through scRNA-seq analysis and comparison with EPI from E3.5, E4.5 that their 2i-LIF ESCs are closest to E4.45 EPI, however their study also lacks comparison with E3.25 inner cells.

ScRNA-seq study shows that *Oct4* expression is stabilized at high level in the ICM of late stage 32-cell embryos. To validate our findings based on microarrays, we searched for a single-cell RNA-seq dataset that encompasses the developmental time around E3.25 and found that of Posfai *et al.*[10], GSE84892. Unfortunately, the information on the exact number of cells of the embryos from the early and late 32-cell stages, from which the single cells for the RNA-seq analysis were dissected is not preserved (personal communication). However, (1) The early and late 32-cells have been harvested 72 h and 78 h post fertilization, and correspond to E3.0 and E3.25, respectively, (2) The early and late 32-cell stage differ in the degree of formation of blastocoel, and (3) Most of the early 32 TE cells are "specified and committed and few cells are capable of forming ICM" and among the late 32 cells, "all TE cells are specified and committed", and for both early and late 32 cells, "ICM specified, but not committed, TE fate can be induced by forced outside exposure or by inactivating Hippo signaling" (Fig. 7 in Posfai *et al.*[10]). This situation resembles what we observed from our analysis of the dataset of Ohnishi *et al.*[3] that "E3.25-HNC embryos are more developed than E3.25-LNC" but not yet committed to PE or EPI lineage. Therefore, we searched for the DEGs between the early and late 32 cell stage ICM cells from Posfai *et al.*[10]. The top 80 up-regulated genes in the late 32-cell ICM (L32ICM-h-DEGs) are shown in the heat map of Fig. 9, together with their expression in the ICM cells during early and late 16-cell stage and in the ICM of the 64-cell stage of Posfai *et al.*[10]. The intersection between the top 80 L32ICM-h-DEGs and the top 80 HNC-h-DEGs includes *Pou5f1*/*Oct4*, *Eif4a2* and *Trim24*, all of them among the several top ones in both L32ICM-h-DEGs and HNC-h-DEGs. The expression of *Oct4* in the ICM RNA-seq data from Posfai *et al.*[10] follows that of the Affymetrix array data presented in the *Oct4* violin plot in Fig. S5. *Oct4* is higher expressed and stabilizes at high level in the late 32 ICM cells (Figs 3A and 9). Prior to the 32-cell stage, *Pou5f1*/*Oct4* is high and stable in both microarray and RNA-seq data (Figs S5 and 9). Interestingly, together with *Oct4*, among the top 80 L32ICM-h-DEGs, we found the core pluripotency factors *Klf4* and *Sox2*, as well as the pluripotency associated transcript 9, *Platr9*/2410114N07Rik. Additionally, we found *Gata6*, observed in the majority of ICM cells in early blastocysts and later confined to the PE progenitors at the mid-blastocyst stage[21]. We checked the expression of L32ICM-h-DEGs in the ENCODE project and found that these genes have the highest expression in either testis adult (*Trim24*, *Pcbp2*, *Aven*, *1810044D09Rik*, *Csd/Ybx3*, *Pou5f1*, *Yif1b*) and ovary adult (*Yy1*, *Pum1*), or placenta adult (*Morf4l2*, *Trap1a* with expression restricted to placenta, *Snx5*), i. e. are genes expressed in either the germ cells of the embryo proper or the extra-embryonic placenta. Analogously to the HNC-h-DEGs, we performed an intersection of the L32ICM-h-DEGs (fold-change 1.5) with the interactomes shared by Oct4 and Sox2 (Oct4-Sox2), and the interactomes shared by Oct4 and Nanog (Oct4-Nanog) (Fig. S9). Among the proteins shared by the Oct4-Sox2 interactome and the L32ICM-h-DEGs, we found two genes (Fig. S8): *Parp1*, and *Ssrp1*, both of them members of the intersection of the HNC-h-DEGs with the Oct4-Sox2 interactomes. Among the proteins shared by the Oct4-Nanog interactomes and the L32ICM-h-DEGs, we found five common genes (Fig. S9): *Cnot1*, *Esrrb*, *Mbd3*, *Sall4* and *Sox2*, with *Mbd3* being common with intersection of the HNC-h-DEGs with the Oct4-Nanog interactome. *Cnot1*, CCR4-NOT transcription complex, subunit 1, is a component of the core pluripotency circuitry conserved in mouse and human ESCs[28]. *Sall4*, spalt like transcription factor 4, is an important pluripotency factor that is required for stem cell maintenance. *Essrb*, estrogen related receptor, beta, enhances induction of naïve pluripotency, when transduced with *Oct4*, *Sox2*, and *Klf4* (OSK) into murine fibroblasts[29]. *Esrrb* is shown to unlock silenced enhancers for reprogramming to naïve pluripotency[30].

Minimum 34 cells are required for a kernel without contacts with the outer shell. We found a clear cut boundary in the transcriptomics profile of the ICM cells from LNC and HNC blastocysts at estimated E3.25. Several of the 80 top-ranked HNC-h-DEGs are cell-cell signalling (*Cnntb1*, *Cldn7, Srsf1*), plasma membrane proteins (*Bsg, Cd9, Tmed2*), or/and hypoxia-inducible (*Bsg*, *Sdhd*, *Egln1*) (Fig. 3A); a comprehensive annotation of all HNC-h-DEGs is given in Fig. S3. Therefore, we expect that the strong transcriptomics change arising in embryos with 34 cells are due to the cell signalling caused by the cell packaging. After compaction at the 8-cell stage, the blastomeres are not spherical but resemble a rhombic dodecahedron; however, to study the packing of the blastomeres, we approximated them by spheres. We built compact models under the hypothesis that at E3.25 the embryo is approximately a sphere with all its cells compactly packed to minimize the embryo's volume. Using a random close packing (rcp), the expected packing density is $\eta \approx 0.64$[31], whereas the demonstration of the Kepler conjecture[32] shows that the optimal arrangement of any packaging of spheres has a maximum average packing density $\eta_{max} = \pi/3\sqrt{2} \approx 0.74$. Such density is achieved using hexagonal close-packed (hcp) or face-centred cubic (fcc) regular lattices. The maximum number of near neighbours (coordination number) of hcp and fcc regular lattices is 12. Using a hcp model to fill a spherical embryo, we found that for

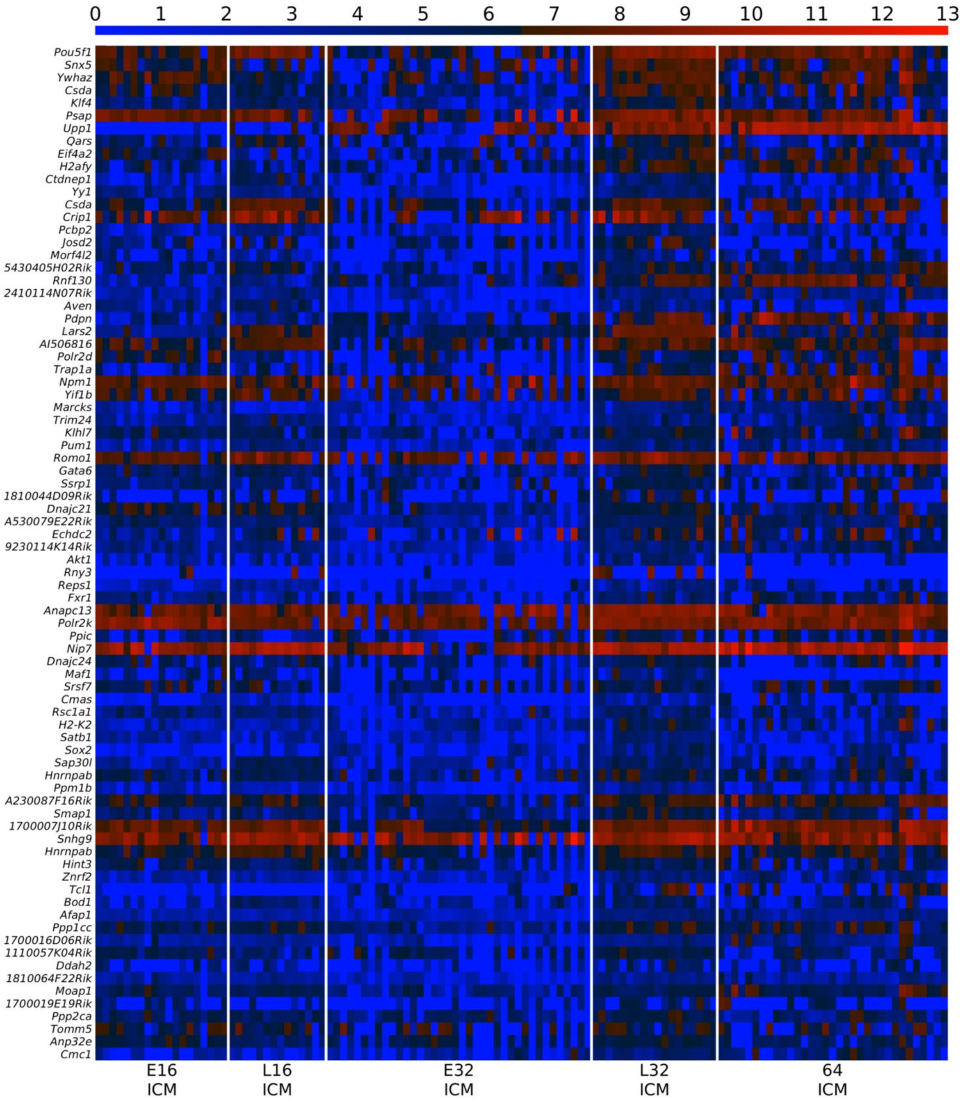


Figure 9. Expression of *Oct4* is stabilized at high level in the late 32-cell ICM. Heat map of the expression of the 80 top-ranked L32ICM-h-DEGs in the ICM cells from early 16 cells to 64 cells from the dataset of Posfai *et al.*[10]. The L32ICM-h-DEGs are the DEGs up-regulated in the late 32-cell ICM in comparison with the early 32-cell ICM of Posfai *et al.*[10]. The colour bar codifies the gene expression in $\log_2$ scale. Higher gene expression corresponds to redder colour.

blastocysts with ≤ 33 cells (Fig. 10A,C), the kernel of cells without contacts with external cells consists of a maximum of 5 cells (Fig. 10E), while for all blastocysts with ≥ 34 cells (Fig. 10B,D), the kernel of cells without contacts with external cells has a minimum of 6 cells (Fig. 10F). The 5-cell kernel forms a pyramid (Fig. 10E), with the 4 base-cells having a coordination number (within the other kernel cells) of 3, and the vertex cell having a coordination number of 4; thus, there are $4 \times 3 + 4 = 16$ contacts (within kernel cells), that, since the maximum coordination number is 12, implies that $16/(5 \times 12) = 26.7\%$ of the kernel contacts are with kernel cells, and 73.3% with inner but non-kernel cells. The 6-cell kernel forms an octahedron (Fig. 10F), whose cells have a coordination number (within the kernel cells) of 4; thus, there are $4 \times 6 = 24$ contacts, which implies that $24/(6 \times 12) = 33.3\%$ of the kernel contacts are with kernel cells, and 66.7% with inner but non-kernel cells. These results suggest that the stabilization of *Oct4* expression at high level, and the activation of the epigenetic (DNA unmethylation and chromatin remodelling events), requires a kernel of cells with at least a 33.3% of the contacts established inside the kernel; thus, suggesting a minimal interaction competitive cell network in which a third of the internal contacts counteracts the signals from non-kernel cells. Therefore, we hypothesize that in such networks the signal among the kernel cells should be at least twice more intensive than the signal coming from the outside for the kernel cells. Ctnnb1, Pou5f1 and Dppa2 are the driving forces in the embryo development at E3.25, and in the segregation of the E3.25 ICM cells into groups corresponding to two different developmental stages (Fig. 10G).

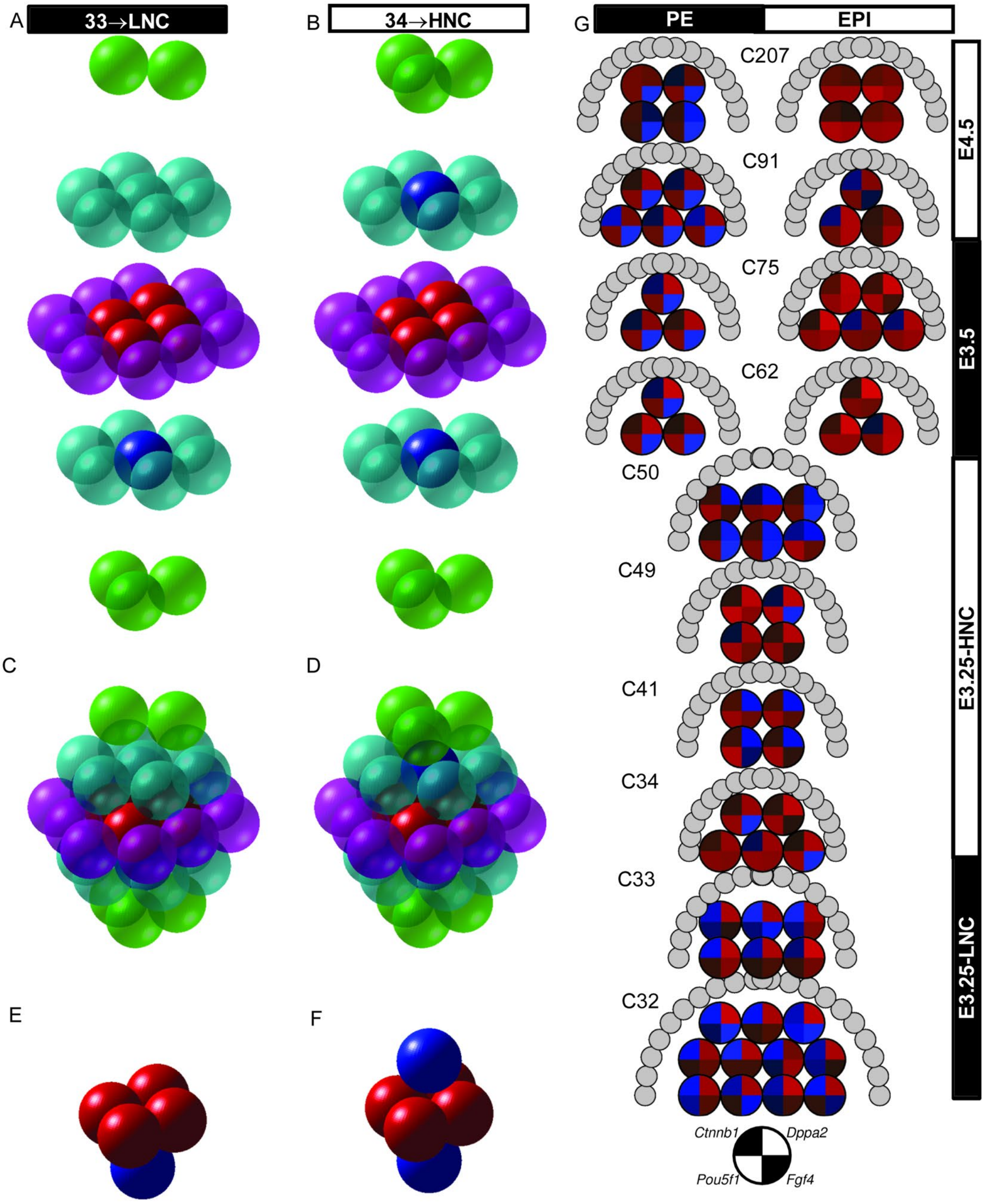


Figure 10. Simplified spatial-temporal mouse embryo model. 3D embryo model based on hexagonal close-packing for embryos with 33 and 34 cells, and layer reconstruction of: (**A**) 33-cell and (**B**) 34-cell embryos. Packing of the: (**C**) 33-cell and (**D**) 34-cell embryos. (**E**) 5-cell and (**F**) 6-cell kernels formed of cells without external contacts. The transparent and solid spheres represent external and kernel cells, respectively. (**G**) Temporal representation of ICM cells from blastocysts at different number-of-cell stages and for different cell types. The gene expression of each cell is represented by a pie-chart of the expression of *Pou5f1*, *Fgf4*, *Ctnnb1* and *Dppa2*. The gene expression is represented in $\log_2$ scale. Higher gene expression corresponds to redder colour, and lower expression to bluer colour. The genes and their corresponding positions in the pie-chart are represented by the black and white pie-chart in the bottom. The semicircle of grey circles surrounding the top part of each group of single cells represents the blastocyst outer cells.

Discussion

The HO*k*M clustering algorithm enabled us to identify two groups of ICM cells during the E3.25 32-64-cell embryonic stage, corresponding to two subsequent developmental stages, E3.25-LNC and E3.25-HNC, rather than to the lineage progenitors PE or EPI. The E3.25-LNC and E3.25-HNC show different degrees of potency, a hypothesis supported by the functional study of Suwińska *et al.*[33], which showed that blastomeres lose totipotency after the fifth cleavage division (5th cell division = 32 cells). Suwińska *et al.*[33] showed that both inner and outer blastomeres of 16-cell embryo developed into normal, fertile mice, while only inner cells from the 32-cell embryo (nascent blastocyst) and containing newly-formed trophectoderm could implant and develop in utero. As we observe, the E3.25-HNC cells are neither PE or EPI, which is in concordance with conclusion of the functional study of Tarkowski *et al.*[34] that "up to the early blastocysts stage, the destiny of at least some blastomeres, although they have begun to express markers of different lineage, is still labile". The genes defining the E3.25-LNC and E3.25-HNC stages of different degree of potency, indicate that during the cell divisions ocurring between the 32- and 64-cells, the development of the embryo to 34 cells triggers a dramatic event in which *Oct4* expression is finally stabilized at high level in the ICM to establish the pluripotent state and initialize the chromatin remodelling program. Here, we propose that the stabilization of *Oct4* to high expression is a non-cell autonomous process that requires an "inside" environment, i.e. a minimal number of four inner cell contacts in conjunction with a heterogeneous environment formed by the niche of a kernel of at least six inner cells covered by a crust of outer TE cells. The involvement of such non-autonomous cell process prevented the uncovering of the mechanism of establishing of early embryonic pluripotency until now.

We hypothesize that at 34–64 cells (E3.25-HNC) compaction and adhesion events are intensified compared to 32–33 cells (E3.25-LNC) since the HNC-h-DEGs are enriched in glycoproteins such as *Bsg, Psap*, *Cd9*, *Epcam*, *Scarb2, Sep15* and *Dnajc10*, and as it has been reported, cell surface glycoproteins play an important role in the development of preimplantation mouse embryo during compaction and trophoblast adhesion[35]. Furthermore, several E3.25 HNC-h-DEGs such as *Ctnnb1*, *Ccni* and *Epcam* are coding cell-adhesion related proteins[36,37], and some are coding transmembrane proteins such as *Cldn7*, *Tm9sf3*, and *Mbtps1*. Additionally, the creating of the "inside" environment at the 34–64-cell stage activates hypoxia signalling, shown by the up-regulation of several hypoxia-inducible genes (*Bsg*, *Acaa2*, *Sdhd*, *Atp1b1*, *Egln1*, *Hsp90b1*).

When the embryo acquires 34 cells, the cell-cell contacts between kernel cells increase to 33.3%. This is also the stage when *Ctnnb1* is highly expressed, indicating activation of the Wnt pathway. It is known that Oct4 is downstream of the Wnt pathway, being a direct target of Ctnnb1[38]. Wnt signalling promotes maintenance of pluripotency by leading to the stabilization of Oct4[39], and through interaction with Klf4, Oct4 and Sox2, Ctnnb1 enhances expression of pluripotency circuitry genes promoting cellular reprogramming[40]. Wnt/β-catenin triggers the activation of *Oct4* expression since it enhances iPSCs induction at the early stage of reprogramming, however, Wnt/β-catenin is not required for pluripotent stem cell self-renewal[40]. We observed that *Oct4* expression is stabilized in E3.25-HNCs, together with activation of several chromatin remodellers. Additionally, a negative regulator of Wnt is Dkk1 (Dickkopf1), which we found to be among the 22 common PE markers for E3.5 and E4.5, and which is a feedback target gene for Wnt. Among the genes whose expression follows the activation of the Wnt pathway and are reported in the literature (www.stanford.edu/~rnusse/wntwindow.html), are also *Sox17* (in the intersection of E3.5 and E4.5 PE markers), *Fgf4* and *Nanog* (in the intersection of E3.5 and E4.5 EPI markers), as well as *Son*, which is a E3.25 HNC-h-DEG.

The E3.25-LNC have "salt and pepper" expression pattern of *Oct4*. They have still undetermined ICM fate whereas the E3.25-HNC have robust expression of *Oct4* and have confidently taken the ICM fate. Such observation has been confirmed through our reanalysis of the scRNA-seq data of early and late 32-cell ICM. The E3.25-HNCs are similar to both E3.5 PE and E3.5 EPI, which is consistent with the fact that E3.25-HNCs are ICM, and therefore pluripotent; however expressing simultaneously PE and EPI markers. The late 32-cell ICM cells have up-regulated genes high-expressed in adult testis, ovary, and placenta, thus they have neither EPI nor PE determined fate. The E3.25-HNCs are also ICM cells that have not taken yet the EPI or PE fate. *Dppa2*, highly expressed E3.25-LNC is an early marker of reprogramming to iPSCs involved in the early stochastic reprogramming phase[41]. In E3.25-LNC, *Dppa2* is highly expressed. The dichotomy between high expression of *Dppa2* in the E3.25-LNC and "salt and pepper" expression pattern in E3.25-HNC, and the converse expression pattern of *Oct4*, could produce different binding patterns in preimplantation embryos at E3.25 as it has been observed in ESCs, where Dppa2 binds promoters but is absent from enhancers[42], binding predominantly to promoters with low activity (high enriched in H3K4me3 and low enriched in H3K27), with a promoter binding pattern uncorrelated with H3K27ac, features only shared with Kdm5b and Polycomb repressor proteins. These features are different from other pluripotency-inducing factors, such as Oct4 and Esrrb, which predominantly bind moderately active enhancers.

Our comprehensive scanning of all the ESC Affymetrix Mouse Genome 430 2.0 Array transcriptomics profiles in GEO database revealed that the closest *in vitro* counterpart of the *in vivo* E3.25-HNCs are 2i + LIF ESCs, not excluding the possibility for the discovery of closer *in vitro* counterparts in the future.

Concluding, we propose a model to explain the establishment of pluripotency from totipotency at the preimplantation based on a non-cell autonomous process in which when the embryo reaches 34 cells at E3.25, the 33.3% of the cell-cell kernel contacts reach a critical surface to activate the canonical Wnt pathway, which stabilizes its downstream Oct4 target, that in its turn activates the chromatin remodelling to promote the expression of the pluripotency circuitry genes. The transcriptomics heterogeneity underlying the EPI - PE segregation is running well before the 32-cell stage (E3.25-LNC). Both PE (*Cubn*) and EPI (*Fgf4*, *Sox2*) markers, among 2271 genes in total, follow bifurcated trajectories ever since before E3.25-LNC. We propose that *Oct4* stabilization, combined with the broad, already existing heterogeneity, leads to the distinguishable PE and EPI transcriptomics profiles of the two lineages at E3.5.

Methods

Microarray data processing. We collected data from the Affymetrix Mouse Genome 430 2.0 Array platform from bulk and single-cell analysis from the GEO and ArrayExpress databases. The annotation records of the collected data are presented in Table 1. The data were normalized with the Robust Multi-array Analysis (RMA)[43] using the BioConductor software package (www.bioconductor.org). The PCA and the hierarchical clustering of genes and samples was performed with one minus correlation metric and the unweighted average distance (UPGMA) (also known as group average) linkage method.

Hierarchical Optimal *k*-Means (HO*k*M) clustering. To account simultaneously for maximal amount of information and reduction of the intrinsic noise of the single-cell transcriptomics profiles, we designed a new clustering algorithm, Hierarchical Optimal *k*-Means (HO*k*M) that determines simultaneously, in an optimal way, the number of clusters of cells, and the cells belonging to each cluster. HO*k*M is a hybrid method, combining the classical *k*-means and the hierarchical clustering methods. HO*k*M uses the classical *k*-means to cluster the single cell gene expression profiles building a bidimensional dimensional "fingerprint", and the hierarchical clustering to re-cluster the clustering results of the "fingerprint" with a classical *k*-means clustering method. HO*k*M is based on the generation of a bidimensional "fingerprint" for each tentative number of clusters, *k*. Such "fingerprint" is the cluster assignment of each cell (for a collection of features for different thresholds of signal variability, and for different number of replicates) by the classical *k*-means clustering algorithm. Next, each "fingerprint", corresponding to a *k*, is clustered through a hierarchical clustering algorithm, and the fitness of the clustering is evaluated through a silhouette coefficient. Finally, HO*k*M selects simultaneously an optimal *k* (the one resulting in the maximum silhouette coefficient) and an optimal clustering.

Following is a detailed description of the HO*k*M clustering algorithm:

(1) Preprocessing. To eliminate low responsive signal, HO*k*M filters out all the probes with maximal signal across all the samples below a detection threshold $\theta_d = 5$.

(2) For each k of a range of tentative number of clusters k, $k_{min} \leq k \leq k_{max}$, HO*k*M builds a $N_S \times (N_V \cdot N_R)$ "fingerprint" bidimensional matrix F_k, where N_S is the number of samples (transcriptomics profiles of cells), N_V is the number of thresholds of signal variability θ_V, and N_R is the number of replicates used in the classical *k*-means algorithm. The element $F_k(s,v{\cdot}r)$ is the cluster assigned by the r-th replicate of the classical *k*-means clustering algorithm to the v-th slice of variability passing the threshold of variability θ_V for the s-th cell-transcriptomics sample. Here, we used $k_{min} = 1$, $k_{max} = 8$.
HO*k*M considers the dynamics of the signal variability by eliminating progressively low variable signals for a variability threshold θ_V, $\theta_{VMin} \leq \theta_V \leq \theta_{VMax}$, at a step θ_{VSte}, i. e. HO*k*M filters out all the probes, whose difference between maximal and minimal signal across all the samples is greater than θ_v; thus, creating sets of slices of probes with progressively higher signal-to-noise ratio. Here, we used $\theta_{VMin} = 2$, $\theta_{VMax} = 9$, $\theta_{VSte} = 0.1$. For each filtered set of probes, HO*k*M applies the classical *k*-means clustering algorithm using N_R different random starting means of clusters; thus, creating a collection of partitions with different θ_V and different starting means of clusters. Here, we used $N_R = 200$. Such collection of partitions is the "fingerprint" F_k of k.

(3) For each k, HO*k*M performs a hierarchical clustering on the s dimension of F_k with a standardized Euclidean metric and a Ward (inner squared distance, minimum variance algorithm) linkage method. HO*k*M cuts the dendrogram produced by the hierarchical clustering at the distance that produces k clusters. The datasets belonging to each of the k clusters are taken as the k partition of the data.

(4) For each k partition of the hierarchical clustering, HO*k*M calculates the silhouette coefficient s_i (which is a measure of how similar that dataset is to datasets in its own cluster, compared to datasets in other clusters) for each dataset $i \leq k$, $-1 \leq \mathrm{s}_i \leq 1$:

$$s_i = \frac{min\left(b_{i,j}\right) - a_i}{max\left(a_i,\ min\left(b_{i,j}\right)\right)},\quad i,\ j \leq k, \tag{1}$$

where $a_i = (1/(|C_i| - 1))\sum_{j \in C_i, i \neq j} d(i,\ j)$ is the average distance from the i-th dataset to the other datasets in its cluster, and $b_{i,j} = \underset{i \neq j}{min}(1/|C_j|)\sum_{j \in C_j} d(i,\ j)$ is the average distance from the i-th dataset to datasets j in another cluster[44], where i is a sample point in the cluster C_i, and $d(i, j)$ is the distance between sample points i and j.
For the limit case $k = 1$, for which the silhouette coefficient is undefined, we define an extrapolation of the silhouette coefficient that is a measure of the cluster compaction:

$$s_1 = \frac{max(d_1) - a_1}{max(a_1,\ max(d_1))}, \tag{2}$$

where d_1 is a vector with the distances between all the datasets; thus $max(d_1)$ is the maximal distance among the cloud of datasets. Since for $k = 1$ the *k*-means clustering is not needed, HO*k*M uses $N_R = 1$ and as a result the silhouette coefficient is less dispersed.

(5) The silhouette coefficient S_k for the k partition is the average of the silhouette coefficients s_i of all the datasets. HO*k*M takes as an optimal number of clusters k_o, the k that produces the highest S_k, and as optimal clustering, the partition produced by the hierarchical clustering of k_o.

We validated the HO*k*M algorithm by applying it to the already classified PE and EPI cells from E3.5 and E4.5 of Ohnishi *et al.*[3] and classified the cells with 100% accuracy (Fig. 7B,C).

Transcriptomics dynamics through bifurcation analysis. To eliminate low responsive signals, for each group of single-cell transcriptomics datasets we filtered out all the probes whose maximal signal across all the samples was below a detection threshold $\theta_d = 5$. To eliminate low variable signals, we filtered out all the probes whose difference between maximal and minimal signal across all the samples was over the signal variability threshold $\theta_V = 8$. Thus, we selected a set of 16,647 responding probes *RP* with higher signal to noise ratio and reduced the number of probes to be processed in the subsequent stages. For each of the selected probes, and for each group of datasets of the embryonic stages $ES = \{$E3.25-LNC, E3.25-HNC, E3.5, E4.5$\}$, we performed a hierarchical clustering with the Euclidean metric and the single (shortest distance) linkage method. For each of the resulting $|RP| \times |ES| = 16,647 \times 4 = 66,588$ hierarchical clusterings, the clustered tree was cut into not more than k_{max} branches. Since we were interested in the analysis of two-fate decisions, we performed a bifurcation analysis, thus, we used $k_{max} = 2$. Each probe p was represented by the mean of the data of each cluster. To avoid spurious bifurcations, when the mean values of two clusters were smaller than a proximity threshold $\theta_d = 3$, the two clusters were fused into a unique one, and the cluster mean was recalculated as the mean of all the samples. This process was repeated for the four ES datasets. Thus, we classified each probe for each embryonic stage ES into one or two clusters which were represented by the average of probes belonging to them. The representative value of each probe in each cluster was set by the average of the expression values of the probe in all the cells belonging to that cluster. Since we have $d = |ES| = 4$ passing times, $d_i \subset ES = \{$E3.25-LNC, E3.25-HNC, E3.5, E4.5$\}$, and one to two bifurcations of transcriptomics values for each passing time, the total number of possible transcriptomics trajectories is $(k_{max})^d - 1 = 15$. For each probe p belonging to these 15 trajectory categories, and for each passing time d_i, we calculated the difference D_p of the bifurcations means. In case of no bifurcation ($k = 1$) at d_i, D_p was assigned a zero. The averaged difference $\overline{D_p}$ across the d passing times was used to rank the transcripts in decreasing order of bifurcation for each of the 15 trajectory types.

Selection of statistically significant DEGs. To find the statistically significant DEGs between two groups of samples, we calculated the mean values of each probe across all the samples of each of the two groups. Next, we filtered out all the probes whose absolute value of difference of mean values between the two groups was less than a selection threshold $\theta_{DEG} = 4$ (that corresponds to a fold change of 2 in $\log_2$ scale), and we applied the Student's t-test. The multitest effect influence was tackled through control of the False Discovery Rate (FDR) using the Benjamini-Hochberg method for correcting the initial p-values with significance threshold $\alpha_{DEG} = 0.001$.

Chromosomal landscaping of statistically significant DEGs. To evaluate the chromosomal enrichment of lists of DEGs, we mapped each DEG onto its corresponding chromosome, and we compared the number of DEGs mapped onto a chromosome with the number of genes on each chromosome. We estimated the statistical significance of the enrichment by the p-value calculated using the hypergeometric distribution using a with significance threshold. The multitest effect influence was tackled through control of the False Discovery Rate (FDR) using the Benjamini-Hochberg method for correcting the initial p-values with significance threshold $\alpha_{LAN} = 0.01$. To analyse the way the gene *loci* group on a chromosome, we designed a one-dimensional clustering technique that searches for groups of gene *loci* located at a distance smaller than $D_{clu} = 2,000,000$ bps, and considers them as a cluster if they are populated by a minimum number $N_{min} = 4$ genes.

Chromosomal mosaicism analysis based on transcriptomics data. We infer possible chromosomal mosaicism under the assumption that a ploidy aberration event affects the expression of the genes located on a chromosome. Therefore, we hypothesized that the statistical enrichment of a chromosome with DEGs between two sets is an indication that the chromosome has a ploidy change. We applied the Jack-knife method between the transcriptomics profile of each single cell and the transcriptomics profile of the pool of cells remaining after the single cell was withdrawn from the total pool of E3.25 cells. The chromosomal enrichment was evaluated by the chromosomal landscaping method where the statistical significance was estimated by the p-value$_{\mathrm{ChrEnr}}$ calculated using the hypergeometric distribution. To visualize these p-values$_{\mathrm{ChrEnr}}$, we used the $-\log10(p\text{-value}_{\mathrm{ChrEnr}})$ transformation. To check whether the transcriptomics split at E3.25 is associated to possible chromosomal mosaicism, we calculated the statistical significance between the p-values$_{\mathrm{ChrEnr}}$ of the E3.25-LNC and E3.25-HNC distributions using the two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test.

GO statistically significant enrichment analysis of DEGs. The GO terms were taken from the curated collection of molecular signatures (gene set collection C5) of version 3.0 of the Molecular Signatures Database (MSigDB)[45]. The significance of DEGs was analysed using an enrichment approach based on the hypergeometric distribution to estimate the significance (p-value) of the gene set enrichment. The multitest effect influence was tackled through control of the False Discovery Rate (FDR) using the Benjamini-Hochberg method for correcting the initial p-values with significance threshold $\alpha_{GO} = 0.05$.

Building of the interaction network of the E3.25 HNC-h-DEGs. For the list of E3.25 HNC-h-DEGs, we took from the BioGRID database[46] (3.4.125 release) all the known binary interactions involving simultaneously two members of the E3.25 HNC-h-DEGs. Then, we used the neato layout program of the GraphViz graph visualization software (http://www.graphviz.org/) to calculate an approximation of the interaction network layout. The network was streamlined using in-house implemented functions in Matlab.

Intersection of the HNC-h-DEGs with Oct4, Sox2 and Nanog interactomes. The Oct4 interactome was built as the union of the Oct4 interactomes from Pardo *et al.*[47], van den Berg *et al.*[48], Ding *et al.*[49] and Esch *et al.*[50]. The Sox2 interactome was taken from Fang *et al.*[51], and the Nanog interactome from Gagliardi *et al.*[52].

Finding *in vitro* counterparts of the E3.25-E4.5 inner cells in the GEO database. We downloaded all 665 ESC transcriptomics profiles from the Affymetrix Mouse Genome 430 2.0 Array platform from the GEO database available at the time, and RMA normalized them together with the *in vivo* data from Table 1. We averaged the single-cell transcriptomics profiles of each stage and differentiated cell type PE and EPI where applicable (E3.25-LNC, E3.25-HNC, E3.5PE, E3.5EPI, E4.5PE, E4.5EPI) and calculated the distance (using the 1 - Spearman correlation metric) to each ESC profile. From each of the six individual stage (and cell type) comparisons, we selected the ten closest ESC profiles. Next, we made the union of the six groups and visualized the results with a heat map.

ScRNA-seq data processing. The SRA database for GSE84892, SRP079965, contains 292 single-cell RNA-seq data from 16, 32 and 64 cell stage mouse embryos, out of which 122 were classified as ICM by Posfai *et al*.[10]. The single-end RNA-seq data were produced on Illumina HiSeq 2000 platform. We aligned the raw data to the GRCm38 genome using Tophat, and calculated the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) using Cufflinks and in-house software in Python 2.7. Variance stabilization was performed using $\log_2$ scaling. We selected the 80 top high DEGs up-regulated in the late 32-cell ICM cells in relation to the early 32-cell ICM cells corresponding to a fold change of 2.2.

References

1. Morris, S. A. *et al*. Origin and formation of the first two distinct cell types of the inner cell mass in the mouse embryo. *Proc. Natl. Acad. Sci. USA* **107**, 6364–6369 (2010).
2. Zernicka-Goetz, M., Morris, S. A. & Bruce, A. W. Making a firm decision: multifaceted regulation of cell fate in the early mouse embryo. *Nat. Rev. Genet.* **10**, 467–477 (2009).
3. Ohnishi, Y. *et al*. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell Biol.* **16**, 27–37 (2014).
4. Jaenisch, R. & Young, R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell.* **132**, 567–582 (2008).
5. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell.* **126**(4), 663–676 (2006).
6. Kim, J. B. *et al*. Direct reprogramming of human neural stem cells by OCT4. *Nature.* **461**(7264), 649–653 (2009a).
7. Kim, J. B. *et al*. Oct4-induced pluripotency in adult neural stem cells. *Cell.* **136**(3), 411–419 (2009b).
8. Rosner, M. H. *et al*. A POU-domain transcription factor in early stem cells and germ cells of the mammalian embryo. *Nature.* **345**(6277), 686–692 (1990).
9. Kurimoto, K. *et al*. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.* **34**, e42 (2006).
10. Posfai, E. *et al*. Position- and Hippo signaling-dependent plasticity during lineage segregation in the early mouse embryo. *Elife*. **6** (2017).
11. Kubaczka, C. *et al*. Derivation and maintenance of murine trophoblast stem cells under defined conditions. *Stem Cell Reports.* **2**(2), 232–242 (2014).
12. Wu, G. *et al*. Initiation of trophectoderm lineage specification in mouse embryos is independent of Cdx2. *Development.* **137**(24), 4159–4169 (2010).
13. Sidhu, S. S. *et al*. EMMPRIN regulates the canonical Wnt/beta-catenin signaling pathway, a potential role in accelerating lung tumorigenesis. *Oncogene.* **29**(29), 4145–4156 (2010).
14. Ke, X. *et al*. Hypoxia upregulates CD147 through a combined effect of HIF-1α and Sp1 to promote glycolysis and tumor progression in epithelial solid tumors. *Carcinogenesis.* **33**(8), 1598–1607 (2012).
15. Kang, M., Garg, V. & Hadjantonakis, A. K. Lineage establishment and progression within the inner cell mass of the mouse blastocyst requires FGFR1 and FGFR2. *Dev Cell.* **41**(5), 496–510.e5 (2017).
16. Vizlin-Hodzic, D., Johansson, H., Ryme, J., Simonsson, T. & Simonsson, S. SAF-A has a role in transcriptional regulation of Oct4 in ES cells through promoter binding. *Cell Reprogram.* **13**(1), 13–27 (2011).
17. Roper, S. J. *et al*. ADP-ribosyltransferases Parp1 and Parp7 safeguard pluripotency of ES cells. *Nucleic Acids Res.* **42**(14), 8914–8927 (2014).
18. González, B., Denzel, S., Mack, B., Conrad, M. & Gires, O. EpCAM is involved in maintenance of the murine embryonic stem cell phenotype. *Stem Cells* **27**(8), 1782–1791 (2009).
19. Lee, B. K. *et al*. Tgif1 counterbalances the activity of core pluripotency factors in mouse embryonic stem cells. *Cell Rep.* **13**(1), 52–60 (2015).
20. Goolam, M. & Zernicka-Goetz, M. The chromatin modifier Satb1 regulates cell fate through Fgf signalling in the early mouse embryo. *Development.* **144**(8), 1450–1461 (2017).
21. Chazaud, C., Yamanaka, Y., Pawson, T. & Rossant, J. Early lineage segregation between epiblast and primitive endoderm in mouse blastocysts through the Grb2-MAPK pathway. *Dev. Cell.* **10**, 615–624 (2006).
22. Guo, G. *et al*. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell.* **18**, 675–685 (2010).
23. Yamanaka, Y., Lanner, F. & Rossant, J. FGF signal-dependent segregation of primitive endoderm and epiblast in the mouse blastocyst. *Development.* **137**(5), 715–724 (2010).
24. Gerovska, D. & Araúzo-Bravo, M. J. Does mouse embryo primordial germ cell activation start before implantation as suggested by single-cell transcriptomics dynamics? *Mol. Hum. Reprod.* **22**(3), 208–225 (2016).
25. Rezvani, K. *et al*. UBXD4, a UBX-containing protein, regulates the cell surface number and stability of alpha3-containing nicotinic acetylcholine receptors. *J. Neurosci.* **29**(21), 6883–6896 (2009).
26. Boroviak, T., Loos, R., Bertone, P., Smith, A. & Nichols, J. The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification. *Nat Cell Biol.* **16**(6), 516–528 (2014).
27. Semrau, S. *et al*. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat Commun.* **8**(1), 1096 (2017).
28. Zheng, X. *et al*. Cnot1, Cnot2, and Cnot3 maintain mouse and human ESC identity and inhibit extraembryonic differentiation. *Stem Cells.* **30**(5), 910–922 (2012).
29. Sone, M. *et al*. Hybrid cellular metabolism coordinated by Zic3 and Esrrb synergistically enhances induction of naive pluripotency. *Cell Metab.* **25**(5), 1103–1117 (2017).
30. Adachi, K. *et al*. Esrrb unlocks silenced enhancers for reprogramming to naive pluripotency. *Cell Stem Cell.* **23**, 266–275 (2018).
31. Torquato, S., Truskett, T. M. & Debenedetti, P. G. Is random close packing of spheres well defined? *Phys. Lev. Lett.* **84**, 2064–2067 (2000).
32. Hales, T. C. A proof of the Kepler conjecture. *Annals of Mathematics. Second Series* **162**(3), 1065–1185 (2005).
33. Suwińska, A., Czołowska, R., Ozdzeński, W. & Tarkowski, A. K. Blastomeres of the mouse embryo lose totipotency after the fifth cleavage division: expression of Cdx2 and Oct4 and developmental potential of inner and outer blastomeres of 16- and 32-cell embryos. *Dev Biol.* **322**(1), 133–144 (2008).
34. Tarkowski, A. K., Suwińska, A., Czołowska, R. & Ożdżeński, W. Individual blastomeres of 16- and 32-cell mouse embryos are able to develop into foetuses and mice. *Dev Biol.* **348**(2), 190–198 (2010).

35. Surani, M. A., Kimber, S. J. & Handyside, A. H. Synthesis and role of cell surface glycoproteins in preimplantation mouse development. *Exp. Cell Res.* **133**(2), 331–339 (1981).
36. Kim, W. T., Seo Choi, H., Min Lee, H., Jang, Y. J. & Ryu, C. J. B-cell receptor-associated protein 31 regulates human embryonic stem cell adhesion, stemness, and survival via control of epithelial cell adhesion molecule. *Stem Cells.* **32**, 2626–2641 (2014).
37. Hagmann, H. *et al.* Cyclin I and p35 determine the subcellular distribution of Cdk5. *Am. J. Physiol. Cell Physiol.* **308**(4), C339–C347 (2015).
38. Li, J., Li, J. & Chen, B. Oct4 was a novel target of Wnt signaling pathway. *Mol. Cell Biochem.* **362**(1-2), 233–220 (2012).
39. Kelly, K. F. *et al.* β-catenin enhances Oct-4 activity and reinforces pluripotency through a TCF-independent mechanism. *Cell Stem Cell.* **8**(2), 214–227 (2011).
40. Zhang, P. *et al.* Regulation of induced pluripotent stem (iPS) cell induction by Wnt/β-catenin signaling. *J. Biol. Chem.* **289**(13), 9221–9232 (2014).
41. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell.* **150**, 1209–1222 (2012).
42. Engelen, E. *et al.* Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry. *Nat. Commun.* **6**, 7155 (2015).
43. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research.* **31**, e15 (2013).
44. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics.* **20**, 53–65 (1987).
45. Subramanian, A. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**(43), 15545–15550 (2005).
46. Stark, C. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**(Database issue), D535–D539 (2006).
47. Pardo, M. *et al.* An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell Stem Cell.* **6**(4), 382–395 (2010).
48. van den Berg, D. L. *et al.* An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell.* **6**(4), 369–381 (2010).
49. Ding, J., Xu, H., Faiola, F., Ma'ayan, A. & Wang, J. Oct4 links multiple epigenetic pathways to the pluripotency network. *Cell Res.* **22**(1), 155–167 (2012).
50. Esch, D. *et al.* A unique Oct4 interface is crucial for reprogramming to pluripotency. *Nat. Cell Biol.* **15**(3), 295–301 (2013).
51. Fang, X. *et al.* Landscape of the SOX2 protein–protein interactome. *Proteomics* **11**, 921–934 (2011).
52. Gagliardi, A. *et al.* A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *The EMBO Journal.* **32**, 2231–2247 (2013).
53. Cao, F., Fukuda, A., Watanabe, H. & Kono, T. The transcriptomic architecture of mouse Sertoli cell clone embryos reveals temporal–spatial-specific reprogramming. *Reproduction.* **145**, 277–288 (2013).
54. Choi, I., Carey, T. S., Wilson, C. A. & Knott, J. G. Transcription factor AP-2γ is a core regulator of tight junction biogenesis and cavity formation during mouse early embryogenesis. *Development.* **139**(24), 4623–4632 (2012).
55. Maekawa, M., Yamamoto, T., Kohno, M., Takeichi, M. & Nishida, E. Requirement for ERK MAP kinase in mouse preimplantation development. *Development.* **134**, 2751–2759 (2007).
56. Kurimoto, K. *et al.* Complex genome-wide transcription dynamics orchestrated by Blimp1 for the specification of the germ cell lineage in mice. *Genes Dev.* **22**(12), 1617–1635 (2008).

Acknowledgements

D.G. and M.J.A.-B. have been supported by Grants DFG113/18 from Diputación Foral de Gipuzkoa, Spain, Ministry of Economy and Competitiveness, Spain, MINECO Grant BFU2016-77987-P and Instituto de Salud Carlos III (AC17/00012) Grant co-funded by the European Union (Eracosysmed/H2020 Grant Agreement No. 643271). The authors are grateful to Dr Michele Boiani and an anonymous reviewer for their critical reviewer suggestions that helped improving the manuscript.

Author Contributions

D.G. and M.J.A.-B. designed and implemented the study, and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-45438-y.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.