Defining the Three Cell Lineages of the Human Blastocyst by Single-Cell RNA-seq

Paul Blakeley\textsuperscript{1+}, Norah M.E. Fogarty\textsuperscript{1+}, Ignacio del Valle\textsuperscript{1}, Sissy E. Wamaitha\textsuperscript{1}, Tim Xiaoming Hu\textsuperscript{2,3}, Kay Elder\textsuperscript{4}, Philip Snell\textsuperscript{4}, Leila Christie\textsuperscript{4}, Paul Robson\textsuperscript{2,5} and Kathy K. Niakan\textsuperscript{1*}

1. The Francis Crick Institute, Mill Hill Laboratory, London, NW7 1AA, UK
2. Genome Institute of Singapore, A-STAR, 138672 Singapore
3. MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, OX1 3QX, UK
4. Bourn Hall Clinic, Bourn, Cambridge CB23 2TN, UK
5. The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06032, USA

+These authors contributed equally to this work
* Author for correspondence (e-mail: kathy.niakan@crick.ac.uk)

KEY WORDS

Human, mouse, epiblast, trophectoderm, embryonic stem cells, RNA-sequencing
SUMMARY

Here we provide fundamental insights into early human development by single-cell RNA-sequencing of human and mouse preimplantation embryos. We elucidate conserved transcriptional programs along with those that are human-specific. Importantly, we validate our RNA-sequencing findings at the protein level, which further reveals differences in human and mouse embryo gene expression. For example, we identify several genes exclusively expressed in the human pluripotent epiblast including the transcription factor KLF17. Key components of the TGF-β signaling pathway including NODAL, GDF3, TGFBRI/ALK5, LEFTY1, SMAD2, SMAD4 and TDGF1 are also enriched in the human epiblast. Intriguingly, inhibition of TGF-β signaling abrogates NANOG expression in human epiblast cells, consistent with a requirement for this pathway in pluripotency. Although key trophectoderm factors Id2, Elf5, and Eomes are exclusively localized to this lineage in the mouse, the human orthologues are either absent or expressed in alternative lineages. Importantly, we also identify genes with conserved expression dynamics including Foxa2/FOXA2, which we show is restricted to the primitive endoderm in both human and mouse embryos. Comparisons of the human epiblast to existing embryonic stem cells (hESCs) reveals conservation of pluripotency but also additional pathways more enriched in hESCs. Our analysis highlights significant differences in human preimplantation development compared to mouse and provides a molecular blueprint to understand human embryogenesis and its relationship to stem cells.
INTRODUCTION

The morphology of the preimplantation human embryo is remarkably similar to the mouse embryo. After fertilization, both undergo mitotic cell divisions, compaction and cavitation to form a blastocyst comprised of a trophectoderm (TE) layer and an inner cell mass (ICM). Despite these similarities, there are a number of significant distinctions such as the timing of cleavage divisions, blastocyst formation and implantation (Cockburn and Rossant, 2010; Niakan et al., 2012). Mouse embryos also undergo zygotic/embryo genome activation immediately after fertilization (Flach et al., 1982) whereas it remains unclear whether this occurs between the 4- to 8-cell stage or earlier in human embryos.

Three cell lineages comprise the blastocyst: pluripotent epiblast (EPI) cells that form the embryo proper, and extraembryonic TE cells and primitive endoderm (PE) cells that contribute to the placenta and yolk sac, respectively. The molecular mechanisms underlying the specification of these distinct lineages have been extensively studied in the mouse. In the mouse, the first cell fate decision, which segregates the ICM and TE, involves differential Hippo signaling at compaction (Nishioka et al., 2009). Differential FGF signaling at the blastocyst stage leads to the second cell fate decision, the segregation of the EPI and PE lineages within the ICM (Guo et al., 2010).

Comparatively little is known about mechanisms of lineage specification in human embryogenesis, though some gene expression patterns are shared with the mouse (Rossant, 2015). Like the mouse, human embryos express OCT4 in all cells until the blastocyst stage, when OCT4 is restricted to the EPI (Niakan and Eggan, 2013).
Importantly, we previously found that the restriction of OCT4 expression to the EPI correlates with the optimal time for human embryonic stem cell (hESC) derivation, suggesting that further understanding of lineage specification will also have importance for stem cell biology (Chen et al., 2009). However, differences between these species in the expression of lineage-associated factors have also been noted. For example, in the mouse Cdx2 is expressed at the morula stage, whereas CDX2 expression follows cavitation in the human blastocyst (Niakan and Eggan, 2013).

Advances in single cell RNA-sequencing (RNA-seq) transcriptomics approaches have provided significant insights into the transcriptional programs underlying human embryogenesis (Piras et al., 2014; Xue et al., 2013; Yan et al., 2013). While previous studies have compared the transcriptomes of human and mouse preimplantation embryos (Piras et al., 2014; Xue et al., 2013), there is a limited focus on lineage specification. Additional studies used microarray analysis of whole embryos, however cellular heterogeneity complicates the identification of cell-type specific gene expression (Madissoon et al., 2014; Xie et al., 2010; Zhang et al., 2009). Furthermore, few of these studies have validated their computational analyses with independent approaches. This is particularly important because of the known technical variability and stochastic expression in single-cell RNA measurements (Brennecke et al., 2013; Kim and Marioni, 2013) in addition to the threshold for expression having not yet been firmly established (Hebenstreit et al., 2011).

Here we integrated our own human single-cell RNA-seq dataset with published human datasets and compared this to a published mouse single-cell dataset, allowing us to unravel novel temporal-, lineage- and species-specific factors. We developed a
computational pipeline to cluster single-cells into developmental stages based on their global gene expression profiles and showed that the major wave of embryo genome activation occurs between the 4- to 8-cell stage in human and the zygote to late 2-cell stage in mouse. Our analysis revealed that temporal expression dynamics of key developmental regulators and their co-expressed genes are largely distinct in human versus mouse. Significantly, we resolved lineage specific gene expression in humans including expression of a number of key components of the TGF-β signaling pathway in the EPI. Treating human embryos with a potent TGF-β signaling inhibitor resulted in downregulation of NANOG, suggesting that this pathway is necessary to maintain the pluripotent EPI. Our analysis also uncovered factors with conserved expression in human and mouse embryos such Foxa2/FOXA2, which was restricted to the PE. However, while we identified the transcription factor KLF17 as exclusively expressed in the human EPI, we found that the mouse EPI factors Esrrb, Klf2 and Bmp4 are absent from the human EPI. Moreover, a number of key mouse TE factors including Elf5 and Eomes were absent in the human TE and conversely human TE factors CLDN10, PLAC8 and TRIMAL were absent in the mouse. We found that although hESCs expressed many EPI-enriched genes, they also expressed genes that are absent in in vivo pluripotent cells. Altogether, we present a comprehensive comparison of human and mouse preimplantation development that reveals previously unappreciated differences in gene expression and highlight the importance of further analyzing human preimplantation development rather than assuming equivalence to the mouse.
RESULTS

Comparative transcriptomics analysis throughout human and mouse preimplantation development reveals temporal differences in gene expression.

To unravel similarities and differences between human and mouse embryogenesis, we compared their preimplantation transcriptomes using single-cell RNA-seq analysis. We used previously published human (Yan et al., 2013) and mouse (Deng et al., 2014) single-cell RNA-seq datasets as both include deep transcriptome profiling at comparable developmental stages, allowing comparative analysis of gene expression over time.

To normalize for sequencing depth and transcript length, the Reads Per Kilobase of exon model per Million mapped reads (RPKM) method (Mortazavi et al., 2008) was applied to both datasets. For subsequent analysis of temporal changes in gene expression, genes were retained in both datasets if they were expressed in at least one sample, using an RPKM > 5 threshold. This has been shown to reliably capture putative functional mRNAs (Hebenstreit et al., 2011) and is a more stringent threshold compared to RPKM ≥ 0.1 that was previously used (Yan et al., 2013). To investigate gene expression pattern variation between cells at a given stage and across time, we used principal components analysis (PCA) to identify single-cell samples with similar global gene expression patterns in human zygote, 2-cell, 4-cell, 8-cell, morula and late-blastocyst samples (Fig. 1A). As a comparison, we also performed a PCA of mouse zygote, early 2-cell, late 2-cell, 4-cell, 8-cell, morula, early-blastocyst and late-blastocyst samples. While the plot of our de novo PCA of mouse samples closely resembles that previously reported (Deng et al., 2014) our PCA plot of the
human samples is distinct from Yan et al., suggesting that this might be due to different RPKM thresholds applied to the data.

The human and mouse PCA plots showed that the majority of single cells clustered according to their developmental stage. The compact cluster of the human zygote, 2-cell and 4-cell stage samples suggests that they are closer transcriptionally compared to later stages. Conversely in mouse, cells at the zygotic and early 2-cell stage clustered together, resulting in a clear distinction between late 2-cell versus zygotic/early 2-cell stage. Therefore, the PCA suggests that the timing of embryo genome activation in human occurs between the 4- to 8-cell stage, consistent with previous experiments (Braude et al., 1988; Tesarík et al., 1987). Later in development, the human late-blastocyst samples clustered distinctly from the morula samples (Fig. 1A), suggesting that the human late-blastocyst are more divergent in global gene expression.

To further understand developmental gene expression dynamics, we used k-means clustering to group genes with similar expression profiles in the human and mouse time-course data across development (Fig. 1B, Figs S1, S2 and Tables S1, S2). We focused our analysis on genes with a fold change of more than two between any two developmental stages in each species. To determine the optimum number of k-means clusters, we used the Bayesian Information Criterion (BIC) score of the human data, (Fig. S3A) and therefore used 50 clusters in subsequent analyses.

The 50 k-means clusters of co-expressed genes were further grouped by hierarchical clustering (Fig. S3B, C). Here, we observed two general patterns in both datasets. The
first comprises genes that were highly expressed in the zygote and rapidly down-regulated in subsequent stages, perhaps indicating maternal transcripts. The second comprises genes that were largely absent in the zygote and subsequently up-regulated during or after zygotic/embryo genome activation. In mouse, clusters that were largely absent in the zygote were first up-regulated at the 2-cell stage (n = 10 clusters). In contrast, in human embryos we first observed up-regulation at the 4-cell stage (n = 7 clusters), followed by the 8-cell stage (n = 14 clusters). This is consistent with the onset of embryo genome activation at the 2-cell and between the 4- to 8-cell stages in mouse and human, respectively.

To distinguish potentially conserved clusters of co-expressed genes, we selected key pluripotency-associated factors Pou5f1/POU5F1, Sox2/SOX2 and Nanog/NANOG and followed their temporal expression dynamics (Fig. 1B). Pou5f1/POU5F1 (human cluster 49 and mouse cluster 4) shows an up-regulation of expression from the 4-cell to the blastocyst stage. However, the genes co-expressed within these clusters were distinct between the species. For example, human POU5F1 was co-expressed with the TE marker GATA3, the epigenetic regulator BMI1, and the pluripotency factors ZFP42/REX1 and FOXD3, which were all absent in the corresponding mouse cluster. By contrast, Pou5f1 was co-expressed with the epigenetic regulator Jarid2 and the pluripotency factor Tbx3, which were absent in the corresponding human cluster. SOX2 expression was up-regulated from the 4-cell to the blastocyst stage in human (cluster 14) and was co-expressed with a number of genes including KLF3, FZD7, ELF2 and HNF4A. Mouse Sox2 expression was highly up-regulated at the blastocyst stage (cluster 36) and interestingly was co-expressed with a number of TE-associated genes including Gata2, Id2, Elf5 and Eomes. While human NANOG expression was
up-regulated at the 4- to 8-cell stage (cluster 7), mouse Nanog expression was up-regulated earlier between the zygotic and 2-cell stage (cluster 1). Intriguingly, Id1/ID1, Klf4/KLF4 and Bambi/BAMBI were co-expressed with Nanog/NANOG in both species, suggesting that they may belong to a conserved gene regulatory network. The NANOG cluster also contains a number of additional key developmental regulators such as the endoderm transcription factor GATA6, the epigenetic regulator EED and the pluripotency factors DPPA2 and DPPA4. In all, largely distinct sets of genes co-expressed with these key pluripotency factors suggests alternative molecular programs operating between these species.

To further resolve gene expression dynamics, we generated boxplots of RPKM values across time (Fig. 1C). Importantly, the boxplots allow greater insight into variance of gene expression patterns. For some genes such as Gata6/GATA6 we observed similar expression dynamics in both human and mouse embryos. However, the expression dynamics of most genes diverged between these species. For example, while Klf4 was expressed from the earliest stages of mouse development and maintained thereafter, KLF4 was first up-regulated at the 8-cell stage in human. Similarly, Esrrb was expressed in mouse zygotes and maintained throughout preimplantation development whereas ESRRB was expressed in human morulas and subsequently in blastocysts. Pou5f1 transcripts were present in the mouse zygote and initially down-regulated, followed by up-regulation at the 8-cell stage. Interestingly, POU5F1 was not present in the human embryo in appreciable levels until the 8-cell stage, suggesting that maternal transcripts present in human and mouse zygotes may differ significantly. Altogether, this suggests that there are significant differences in gene expression dynamics across time in mouse and human embryos.
Lineage-specific gene expression in human and mouse blastocysts.

To resolve lineage-specific gene expression in human blastocysts we initially used several unbiased approaches to distinguish cell-type specific gene expression in the late-blastocyst samples from Yan et al. We performed a PCA on the human late-blastocyst samples (Fig. 2A), which shows that projection onto the first two principal components was sufficient to group the human cells into two or three clusters. In parallel, we performed unsupervised hierarchical clustering, which reveals that the same samples cluster similarly into three groups (Fig. S4A).

Given the limited number of single-cell EPI and PE samples analysed above, we sought to increase the number of biological replicates to improve statistical power to detect differential gene expression. We performed RNA-seq of additional samples (n = 30 cells; 7 embryos) followed by PCA combined with the time-course dataset from Yan et al. (Fig. S4B). As expected, the additional samples clustered closer to the late-blastocyst stage samples from the Yan et al. dataset. A PCA of the blastocyst samples revealed that while the additional EPI-assigned samples were intermingled with the Yan et al. EPI samples, the PE and TE samples were distinct on the PC2 and PC3 axes (Fig. 2B). These differences may be due to the inherent difficulty of matching developmental stages, differences in the single-cell cDNA synthesis and library preparation protocols or divergent genetic backgrounds. However, the samples do largely cluster into three lineage groups in the PCA as well as by an unsupervised hierarchical clustering (Fig. 2C).
To determine which lineage(s) these groups may correspond to, we generated a list of differentially expressed genes using NOISeq (Tarazona et al., 2011), a data adaptive, non-parametric approach. This approach is well suited for single-cell RNA-seq analysis as these data may not always conform to the same distributional assumptions as RNA-seq data from pooled cells (Kharchenko et al., 2014). NOISeq identified genes enriched in the presumptive EPI including \textit{NANOG}, \textit{ETV4}, \textit{PRDM14}, \textit{FOXD3}, \textit{POU5F1} and \textit{SOX2} (Fig. 3A and Table S3). By contrast the presumptive TE samples were enriched for genes including \textit{GATA2}, \textit{GATA3}, \textit{CDX2} and \textit{KRT18}, whereas the PE samples were enriched for \textit{GATA4}, \textit{GATA6}, \textit{SOX17} and \textit{COL4A1}. We also performed an independent test using DESeq (Anders and Huber, 2010), which fits a negative binomial model and observed considerable overlap of differentially expressed genes predicted by these two independent statistical methods (Fig. S4C, D and Table S4). A heatmap of a subset of lineage-associated genes revealed that most of the human blastocyst samples exclusively expressed genes enriched in one of the lineages, suggesting that at this stage the cells were specified. As before, we found significant differences in the lineage assignments of several blastocyst samples when we compared our assignments to those of Yan et al. (Fig. S4E). This further suggests that the RPKM threshold initially applied to determine expressed genes influences the conclusions drawn from subsequent analyses.

To distinguish pathways differently enriched in either the EPI or TE we performed a comparative analysis of signaling pathways operating in these two lineages. Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) of human TE-enriched genes identified MAPK signaling, transmembrane transport of small molecules and metabolism of lipids and proteins among the most significantly enriched terms (Fig.
3B and Fig. S5). By contrast, in the human EPI GSEA showed that stem cell maintenance and TGF-β signaling were most significantly enriched. Altogether, this is consistent with appropriate lineage assignments for each human blastocyst cell as the pathways identified reflect expected biological characteristics of these lineages.

We next sought to determine lineage-specific expression in the mouse blastocyst samples. A PCA plotting both PC1 against PC2, and PC2 against PC3, distinguished a cluster of six cells from the remaining cells (Fig. 2D). NOISeq analysis identified 354 genes significantly enriched in these six ICM samples, including Nanog, Pou5f1 and Sox2, compared to 143 genes in the remaining presumptive TE samples, which included Gata3, Krt8 and Krt18 (Fig. 3A and Table S5). Hierarchical clustering of these samples using the set of differentially expressed genes indicates a subset of EPI- and PE-associated genes that were simultaneously expressed in the six ICM-designated samples including Nanog, Esrrb, Sox2, Gata6, Sox17 and Gata4 (Fig. 2E). This suggests that while the mouse samples used in this study may have displayed morphological features of blastocyst formation, the ICM cells had not yet undergone lineage specification to EPI or PE. Interestingly, GSEA of mouse ICM-enriched genes revealed stem cell maintenance, embryonic development and regulation of WNT signaling were among the most significantly enriched terms (Fig. 3B). By contrast, cell-cell adhesion, lipid metabolic process, transport of small molecules and EGFR1 pathway were significant terms for mouse TE enriched genes.

Although the samples clustered into distinct lineages, within each group there was heterogeneity in levels of gene expression between individual cells. For example, in the human EPI cells, which expressed consistently high NANOG and DPPA5, we see
variable expression of \textit{POU5F1} and \textit{SOX2} (Fig. 2B). Similarly, the variability in \textit{Pou5f1} expression in the mouse ICM was also captured in the single-cell transcriptomics analysis. However, \textit{PDGFRA}, \textit{COL4A1} and \textit{RSPO3} were consistently expressed in the human PE, suggesting that these may be informative markers of this lineage. The observed heterogeneity in gene expression between single cells, even for key transcriptional regulators, highlights the need for including multiple replicate samples when studying lineage-specific gene expression.

\textbf{Comparison of lineage-specific gene expression in human and mouse blastocysts.}

Next, we investigated genes that were conserved in their lineage-specific expression. Key TE-associated genes \textit{Cdx2/CDX2}, \textit{Gata3/GATA3} and \textit{Krt18/KRT18} were more highly expressed in this lineage in both human and mouse (Fig. 3C). Comparative analysis between human EPI and the mouse ICM revealed 54 orthologous genes enriched in these lineages relative to their respective TE (Fig. 3D, F), including core pluripotency factors \textit{Nanog/NANOG}, \textit{Pou5f1/POU5F1} and \textit{Sox2/SOX2}. Moreover, a number of additional genes thought to function in regulating pluripotency were also conserved including \textit{Prdm14/PRDM14}, \textit{Klf4/KLF4}, \textit{Dppa4/DPPA4}, \textit{Hesx1/HESX1}, \textit{Dppa2/DPPA2}, \textit{Tdgf1/TDGF1} and \textit{Gdf3/GDF3} (Fig. 3F). This suggests that there may be additional genes within the overlapping set that have a conserved role in the pluripotent EPI but whose function has not yet been explored. The PE-associated genes \textit{Sox17/SOX17}, \textit{Pdgfra/PDGFR\textalpha}, and \textit{Gata4/GATA4} also showed conserved enrichment in the human PE and the mouse ICM (Fig. 3E).

Although we identified a number of genes with conserved expression, we also observe important differences. Significantly, although \textit{Elf5}, \textit{Eomes} and \textit{Id2} were
highly enriched in the mouse TE, ELF5 and EOMES were completely absent from any of the lineages in human and ID2 was most abundantly expressed in the PE and absent from most TE cells (Fig. 4A). We also observed genes highly enriched in the human TE, which were not expressed in mouse TE including Cldn10/CLDN10, Triml1/TRIML1 and Plac8/PLAC8, demonstrating key differences in TE gene expression between human and mouse. Furthermore, we find that Tcfap2c, a key transcriptional regulator in the mouse TE, had a different expression pattern in the human (Fig. 4B). Our RNA-seq analysis detected Tcfap2c transcripts in the mouse zygote, with levels remaining high as development proceeds. In contrast, abundant expression of the Tcfap2c orthologue TFAP2C was first detected at the 8-cell stage in human embryos. Lineage-specific analysis showed that, as expected, Tcfap2c was enriched in the mouse TE. In contrast, in the human blastocyst TFAP2C was expressed at similar levels in both the TE and EPI. Immunofluorescence analysis confirmed that Ap2γ, the protein product of Tcfap2c, was specifically localized to Cdx2-positive TE cells in mouse and absent from Nanog-positive cells within the ICM (Fig. 4C and Fig. S5). By contrast, AP2γ was detected in both CDX2-positive TE cells and NANOG-positive EPI cells in human blastocysts (Fig. 4C and Fig. S6).

We investigated the conservation of PE-associated genes in human and mouse. Gata4, Gata6, Sox17, Pdgfra, Col4a1 and Sparc are known to be associated with the mouse PE or its derivatives, with many functionally required for this lineage (Schrode et al., 2013). As expected, we observe abundant expression of these genes in the mouse ICM, and their human orthologues were also more highly expressed in the human PE (Fig. 5A). Hierarchical clustering revealed that human PE cells expressed FOXA2 (Fig. 2B), a gene typically associated with later endoderm development (Ang and
Rossant, 1994; Ang et al., 1993; Monaghan et al., 1993; Sasaki and Hogan, 1993). The boxplots confirmed lineage-specific expression of FOXA2 in the human PE whereas we failed to detect Foxa2 expression in the mouse samples analysed (Fig. 5A). Significantly, immunofluorescence analysis further confirmed that FOXA2 protein was specifically localized to the human PE where it was co-expressed with SOX17, indicating that it is a novel marker of this lineage (Fig. 5B, Fig. S7A). We observe co-localisation of Foxa2 with a subset of Sox17 expressing cells in the mouse late-blastocyst (Fig. S7B) but failed to detect Foxa2 in earlier stage embryos (data not shown), suggesting that Foxa2 may be a marker of the mouse late PE. This may explain the absence of Foxa2 expression in the mouse transcriptome dataset, which appears to have captured expression prior to the late-blastocyst stage, consistent with the co-expression of EPI- and PE-associated transcripts detected in these samples (Fig. 2E).

Several genes were differentially expressed between the human EPI and mouse ICM (Fig. 3F). Importantly, while known mouse pluripotency-associated factors Esrrb, Klf2 and Bmp4 (Nichols and Smith, 2012) were highly enriched in the mouse ICM, KLF2 was absent from the human blastocysts and ESRRB and BMP4 were largely restricted to PE and/or TE cells (Fig. 6A). Conversely, we observed genes that were highly enriched in the human EPI such as LEFTY1, NODAL and ACVRL1/ALK1, which were not expressed in mouse ICM at this stage (Fig. 6A). However, components of TGF-β signaling pathway including Activins, Nodal and Lefty1 are expressed in mouse preimplantation embryos as early as E3.5 (Albano et al., 1993; Paria et al., 1992; Takaoka et al., 2011; Varlet et al., 1997). Given the absence of
these factors from the dataset this further suggests that the mouse ICM samples used reflect an earlier stage of blastocyst development.

Significantly, we also identified several transcription factors that were uniquely enriched in human EPI cells including *KLF17*, which was initially expressed at the 8-cell stage in human embryos and highly enriched in human EPI cells (Fig. 6B). Despite expression in earlier stages of development, *Klf17* was absent in mouse blastocysts (Fig. 6B). By immunofluorescence analysis, we confirmed that KLF17 expression co-localised with NANOG within the EPI cells of human embryos (Fig. 6C) but was undetectable in mouse late-blastocysts (data not shown). Altogether, the single-cell RNA-seq analysis and subsequent validation allowed the confirmation of lineage-associated gene expression thereby revealing fundamental differences in the expression of factors in human and mouse blastocysts.

**TGF-β signaling is necessary for the maintenance of NANOG in human pluripotent EPI cells**

We observed robust expression of multiple components of the TGF-β signaling pathway in the human blastocyst including *SMAD2* and *SMAD4* and receptors *ACVR1, ACVR2B, BMPRI* and *BMPR2*. Interestingly, there were differences between the EPI and TE lineages (Fig. 6D). Receptors *TDGF1* and *TGFBRI*, and ligands *NODAL, GDF3* and *BMP2* were enriched in the EPI, whereas the TE showed enriched expression for *TGFB1* and the negative regulator *TGFB3*. The expression of negative regulators *LEFTY1, LEFTY2* and *BAMBI* in the EPI indicated a feedback loop regulating this pathway. Moreover, in addition to *SMAD2* and *SMAD4*, the TE
also expressed \textit{SMAD1} and \textit{SMAD5}, further suggesting that TGF-β signaling may differentially regulate these lineages.

Components of the TGF-β signaling pathway are also expressed in hESCs (Besser, 2004; James et al., 2005; Levine and Brivanlou, 2006; Sato et al., 2003; Vallier et al., 2009). TGF-β signaling contributes to the maintenance of hESCs by regulating pluripotency gene expression (Bertero et al., 2015; Brown et al., 2011; James et al., 2005; Vallier et al., 2005; Vallier et al., 2004; Xu et al., 2008). Given this role in hESCs, we sought to determine if this pathway was functional required for the EPI. We treated human embryos from E3 – E5 with a selective Activin receptor inhibitor SB-431542, at a concentration of 40 μM, which has been shown to effectively block TGF-β signaling in mouse embryos without toxicity (Granier et al., 2011) and which we confirmed downregulated NANOG expression in hESCs (Fig. S8). We performed immunofluorescence analysis of NANOG and OCT4 expression in blastocysts at E6 - E7. Significantly, most human embryos lacked detectable NANOG expression in the presence of the inhibitor (Fig. 6E, F). Moreover, SOX17 expression was also undetectable in the majority of treated embryos. While OCT4 expression was observed, there were fewer embryos with OCT4-high expressing cells, compared to controls. Altogether this suggests TGF-β signaling is required to maintain key pluripotency marker expression in human EPI cells and a PE marker in vivo.

A Smad2/3-dependent autoregulatory loop is present in mouse preimplantation embryos, indicating a role for TGF-β signaling (Granier et al., 2011; Papanayotou and Collignon, 2014). While EPI formation is initiated, by E5.0 EPI and extraembryonic endoderm genes are misexpressed in both TGF-β signaling-mutant and SB-431542-
treated embryos and further development is compromised (Brennan et al., 2001; Camus et al., 2006; Mesnard et al., 2006; Robertson et al., 2003; Waldrip et al., 1998). Treatment of mouse embryos from the 8-cell to blastocyst stage with SB-431542 does not affect the number of Oct4 or Gata4 expressing cells prior to implantation at E4.5 (Granier et al., 2011). However, as Nanog expression had not yet been examined in SB-431542 treated mouse embryos, we sought to determine if there may be an effect on its expression. In contrast to the human, we found no effect on Nanog, Oct4 or Sox17 expression in treated mouse embryos, which robustly expressed all three markers (Fig. 6G, H), similar to controls. This further suggests that while Nodal signaling is active prior to implantation in mouse embryos, it is not required to initiate or maintain the expression of these EPI or PE markers.

**Defining human ground state pluripotency**

Existing hESCs are thought to represent a later stage of development than their mESC counterparts, despite both being derived from preimplantation blastocysts. Indeed, hESCs share several characteristics with postimplantation-derived mouse epiblast stem cells (EpiSCs), including morphological similarities, LIF-independent growth and a reliance on FGF and Activin/Nodal signalling (Brons et al., 2007; Tesar et al., 2007). Addition of Mek and Gsk3b inhibitors together with LIF (2i+LIF) allows mESCs to be propagated in defined medium thought to represent a “ground state” of pluripotency that is more similar to mouse preimplantation EPI cells, as compared to classical serum and LIF mESCs (Boroviak et al., 2014; Ying et al., 2008). Recent attempts to derive ground state hESCs have utilised combinations of ectopic transgene expression, growth factors and inhibitors to modulate signalling pathways (Chan et al., 2013; Gafni et al., 2013; Takashima et al., 2014; Theunissen et al., 2014). Mek
and Gsk3b inhibitors are often included, although 2i+LIF alone is unable to support the self-renewal of hESCs (Hanna et al., 2010). However, the benchmark against which these cells are assessed relies heavily on conclusions drawn from mouse ground state pluripotency, which our analysis suggests may not be equivalent to the human EPI.

We compared the human EPI to various hESCs using NOISeq to determine the extent to which their gene expression profiles represented the EPI programme (Chan et al., 2013; Takashima et al., 2014; Yan et al., 2014). PCA of differentially expressed genes revealed that samples largely clustered according to experimental condition and cell type (Fig. 7A). We performed unsupervised hierarchical clustering of global gene expression, which again showed that the EPI samples clustered distinctly from hESCs (Fig. 7B). Calculating the Pearson correlation coefficient between each pair of conditions indicated that the hESCs all generally remained distinct from the EPI, with correlation values ranging from 0.58 to 0.68 (Fig. 7C).

We performed GSEA to identify differentially enriched pathways between the EPI and each hESC condition (Fig. 7D). We found that the EPI was enriched for oxidative phosphorylation signalling (Fig. 7D; Fig. S5), possibly reflecting the switch to glycolytic metabolism following stem cell derivation in oxygen-rich conditions (Zhang et al., 2011). hESCs were enriched for regulation of cell proliferation (Fig. 7D; Fig. S5). This suggested that a number of the distinctions were linked to intrinsic properties required to maintain the stem cell state. hESCs were also enriched for FGF, MAPK and Wnt signaling pathways (Fig. 7D; Supplementary Fig. S5). Significantly, both the EPI and hESCs expressed a number of key pluripotency genes including
NANOG, NODAL and PRDM14 (Fig. 7E). Intriguingly, the Chan 3iL and Takashima reset hESCs cultured in alternative conditions upregulated EPI-enriched genes that were not appreciably expressed in conventional hESCs, including DPPA3, DPPA5 and DNMT3L (Fig. 7E, F), suggesting that these conditions have indeed promoted an EPI-like gene expression profile.

We next integrated microarray analyses of additional alternative hESCs (Gafni et al., 2013; Theunissen et al., 2014) with the RNA-seq datasets by normalising the expression of all the samples to conventional hESC derivation conditions (MEFs plus exogenous FGF). As expected, NANOG, POU5F1 and SOX2 expression was similar in both the EPI and hESCs (Fig. 7F), and NODAL and GDF3 were also upregulated, reflecting the requirement for TGF-β signalling in maintaining NANOG expression in the human EPI (Fig. 6E). Furthermore, EPI-associated genes including NR5A2, TCFP2L1, DPPA3 and DPPA5 were expressed in several of the hESCs. However, we found inappropriate upregulation of additional signalling factors FGF2 and FGF4, and the LIF receptor LIFR in hESCs, although curiously the LIF co-receptor IL6ST (GP130) was also enriched in the EPI (Fig. 7F). Although some factors associated with the mouse ground state such as KLF4, TBX3 and DNTM3L were upregulated in both the EPI and hESCs, others, including ESRRB and KLF2, were not appreciably expressed in the human EPI. Interestingly, the novel EPI-specific transcription factor we identified KLF17 was upregulated specifically in the Takashima et al. reset and Theunissen et al. naïve cells (Fig. 7F). Altogether, this suggests that some of the alternative hESC culture conditions do indeed promote a programme closer to that of the human EPI, but extraneous signalling pathway activation may explain why these
cells remain distinct. It would be interesting to determine how best these pathways could be modulated to fully reflect the human EPI.

**DISCUSSION**

Our robust computational analyses of single-cell RNA-seq datasets revealed a number of novel temporal-, lineage- and species-specific factors in human and mouse embryos. Our findings have significance for stem cell biology as the gene networks and signaling pathways regulating human pluripotency during development have yet to be elucidated and this work provides a molecular blueprint to uncover these mechanisms.

Using two independent data mining approaches, our analysis suggests a single wave of genome activation between the 4-cell and 8-cell stage in human embryos, thereby supporting conclusions from uracil radiolabeling and alpha-amanitin transcriptional inhibition experiments (Braude et al., 1988; Tesarík et al., 1987). This is in contrast to findings that suggest a minor wave of genome activation and transcript up-regulation before the 4-cell stage in human (Dobson et al., 2004; Xue et al., 2013). The apparent early detection of transcripts could be due to a subset of preferentially stable transcripts, or alternatively, the delayed polyadenylation of maternal mRNAs (Aanes et al., 2011). To resolve this discrepancy, it may be possible to combine new advances in single-cell transcriptomics together with techniques to enrich for nascent RNA production (Jao and Salic, 2008) to distinguish embryonically transcribed mRNAs.

In the mouse, *Id2* and *Cdx2* are among the earliest transcription factors expressed in TE cells, followed by the expression of *Eomes* and *Elf5* (Guo et al., 2010; Ng et al.,
Our surprising discovery that most of these factors are absent in the human TE suggests that there are fundamental species differences in TE specification, consistent with the temporal differences in CDX2 expression we reported previously (Niakan and Eggan 2013). In the mouse, Tcfap2c is required for the maintenance of the TE lineage, and induced expression of Tcfap2c in mESCs is sufficient to derive mouse trophoblast stem cells (Auman et al., 2002; Kuckenberg et al., 2010; Werling and Schorle, 2002). In the human placenta, TFAP2C is expressed in all trophoblast lineages (Biadasiewicz et al., 2011). Our finding that TFAP2C is more broadly expressed in the human blastocyst is a significant cautionary note against using this gene to assess TE identity. We propose CLDN10, PLAC8 and TRIML1 along with others identified in our analysis as candidates to more appropriately distinguish TE cells.

We have identified human-specific EPI enriched genes, such as KLF17. As alternative members of the KLF family are involved in pluripotency, it would be interesting to investigate if KLF17 might replace known reprogramming factors, such as Klf4, and to determine its function in alternative hESCs. Furthermore, additional gene networks were enriched in both conventional and alternative hESCs compared to the human EPI, including the FGF and Wnt signalling pathway. Given differences in the signaling environment in the human EPI compared to hESCs that were noted previously (Kuijk et al., 2012; Kunath et al., 2014; Roode et al., 2012) and highlighted in this study it will be intriguing to further investigate the possibility of a distinct human pluripotent state.
Recent work has suggested that distinct genetic programs and signaling pathways involved in lineage specification may exist in human and mouse blastocysts, for example the differential requirement for FGF signaling in EPI and PE lineage specification (Kuijk et al., 2012; Kunath et al., 2014; Lanner and Rossant, 2010; Roode et al., 2012). Significantly, we found several key TGF-β signaling pathway components were highly enriched and differentially expressed in the human EPI and TE, and that inhibiting this pathway lead to downregulation of NANOG expression in human but not mouse EPI cells. It was previously suggested that TGF-β signaling inhibition increases EPI proliferation and enhances the outgrowth of cells during hESC derivation (Van der Jeught et al., 2014). The discrepancy with our results may be due to the 4-fold lower concentration of SB-431542 used in the previous study as well as presence of mouse embryonic fibroblasts, known to secrete factors promoting TGF-β signaling, during the hESC derivations described, suggesting that this pathway may not have been completely abolished. Altogether, this suggests that TGF-β signaling may be required for the development of the pluripotent EPI in human blastocysts and further supports the requirement of this signaling pathway in pluripotent hESCs. It would therefore be interesting to determine if stimulating TGF-β signaling in the absence of FGFs during hESC derivation may better recapitulate the embryo signaling environment. Finally, additional alternative signaling pathways may be required for the development of the human EPI and subsequent stem cell derivation. Our dataset provides a resource to discover these developmental cues.
METHODS AND MATERIALS

Human embryo culture and manipulation

Human embryos were donated to the research project by informed consent under the UK Human Fertilisation and Authority Licence number R0162. Embryos were thawed according to recommendations from Bourn Hall Clinic, the in vitro fertilization clinic coordinating donations. Single cells were isolated with the assistance of a Saturn 5 laser (Research Instruments). Further details of the protocols can be found in supplementary methods.

Immunofluorescence analysis

Samples were fixed in 4% paraformaldehyde at 4°C for 1 hour and immunofluorescently analysed as described previously (Niakan and Eggan, 2013). The primary antibodies used include: Oct4 (SC-5279, SC-8628 or SC-9081, Santa Cruz Biotech), Nanog (AF1997 R&D, REC-RCAB0001P 2B Scientific or ab21624 Abcam), Cdx2 (MU392A-UC, Biogenex), Klf17 (HPA024629, Atlas), Ap2γ (AF5059, R&D) Sox17 (AF1924 R&D) and Foxa2 (3143, Cell Signaling). Embryos were imaged on a Leica SP5 inverted confocal microscope (Leica Microsystems (UK) Ltd).

cDNA synthesis, shearing and library preparation

RNA was extracted from single cells and processed for cDNA synthesis using the SMARTer Ultra Low RNA Kit for Illumina Sequencing-HV (Clontech Laboratories, Inc.). Libraries were prepared using Clontech Low Input Library Prep Kit according to manufacturer’s instructions. An extended protocol can be found in the supplementary methods.
Data acquisition and processing

Human and mouse single-cell RNA-seq data normalised using the RPKM method were taken from two previous publications (Deng et al., 2014; Yan et al., 2013) and integrated with our own blastocyst sequencing data. We filtered these datasets, retaining only genes having RPKM > 5 in at least one sample. Extended methods can be found in supplementary information.

Data has been deposited into Gene Expression Omnibus (GEO66507).
ACKNOWLEDGMENTS

We are grateful to the embryo donors for their support of this research. We thank the High Throughput Sequencing, Genome Manipulation Services, Human Embryonic Stem Cell and Confocal Microscopy facilities at The Francis Crick Institute. This work was supported by The Francis Crick Institute, which receives its core funding from Cancer Research UK, the UK Medical Research Council (MC_UP_1202/9) and the Wellcome Trust. In addition, this research was supported by the March of Dimes Foundation (FY11-436) to KKN. THX and PR were supported by the Agency for Science Technology and Research (A-STAR), Singapore.

AUTHOR CONTRIBUTION

KKN conceived and coordinated the project, cultured human embryos, performed microdissections and single-cell collection. NMEF cultured human embryos and performed cDNA synthesis and library preparation of single-cell samples. PB performed bioinformatics analysis of all the human and mouse embryo and stem cell datasets, IDV performed bioinformatics analysis of mouse datasets and PR and THX performed independent analysis of Yan et al. dataset. SEW performed analysis of the stem cell dataset. KKN, NMEF and IDV performed immunofluorescence and additional data analysis. KE, PS and LC obtained informed consent for the research and collected embryos. KKN, PB, NMEF and SEW prepared the manuscript with help from all the authors.

Albano, R. M., Groome, N. and Smith, J. C. (1993). Activins are expressed in preimplantation mouse embryos and in ES and EC cells and are regulated on their differentiation. Development 117, 711-723.


Figures

[Images of figures A, B, and C showing data plots and gene expression patterns for human and mouse.]
**Fig. 1:** Global gene expression dynamics in human and mouse preimplantation development. (A) Principal component analysis of human (Yan et al., 2013) or mouse (Deng et al., 2014) single-cell RNA-seq transcriptomes. Each point represents a single cell and labelled according to developmental stage. Data were plotted along the first and second principal components and the second and third principal components. (B) K-means clusters showing selected genes co-expressed with *Pou5f1/POU5F1*, Sox2/SOX2 or Nanog/NANOG in mouse or human pre-implantation embryos. Grey line corresponds to scaled RPKM values for genes and black line to median expression within the cluster. (C) Boxplots of RPKM values for selected genes showing the range of single-cell gene expression at each of the selected development stages. Boxes correspond to the first and third quartiles, horizontal line to the median, whiskers extend to 1.5 times the interquartile range and dots were outliers.
Fig. 2: Lineage specific gene expression in human and mouse blastocysts. (A, B, D) PCA at the late-blastocyst stage. Each point represents the gene expression profile of a single cell from blastocysts and labelled according to both lineage identity and experiment. Data were plotted along the first and second principal components and
the second and third principal components. Data are from (A) Yan et al. 2013; (B) a combined dataset including our additional dataset together with data from Yan et al. 2013; (D) Deng et al. 2014. (C, E) Unsupervised hierarchical clustering of samples and heatmaps of differentially expressed genes. Normalized expression was plotted on a high-to-low scale (purple-white-green) and genes grouped according to lineage-associated expression. (C) A combined human late-blastocyst dataset including samples generated in our lab together with data from Yan et al. 2013. (E) Mouse late-blastocyst dataset from Deng et al. 2014.
Fig. 3: Genes showing similar lineage-associated expression in human and mouse blastocysts. (A) NOISeq was used to calculate the probability of differential expression between (A) human TE versus EPI, or mouse TE versus ICM. The log2
fold change (FC) difference in expression is noted. (B) Cytoscape enrichment map of GSEA results comparing human TE (blue) versus EPI (red), and mouse TE (blue) versus ICM (red) (p-value < 0.01). (C-E) Boxplots of RPKM values for selected genes in human or mouse (C) TE; (D) EPI; or (E) PE. The range of expression in human EPI (green), PE (red) or TE (blue) and in mouse ICM (orange) or TE (blue). Boxes correspond to the first and third quartiles, horizontal line to the median, whiskers extend to 1.5 times the interquartile range and dots were outliers. (F) VENN diagram of overlapping orthologous gene expression in human EPI and mouse ICM.
Fig. 4: Differences in TE-associated gene expression in human versus mouse blastocysts. (A) Boxplots of RPKM values for selected genes. The range of expression in human EPI (green), PE (red) or TE (blue) and in mouse ICM (orange) or TE (blue). Boxes correspond to the first and third quartiles, horizontal line to the median, whiskers extend to 1.5 times the interquartile range and dots were outliers. (B) Boxplots of RPKM values for Tcfap2c/TFAP2C in human or mouse late-blastocysts and at each of the selected development stages. (C) Immunofluorescence
analysis of human or mouse blastocysts for Ap2γ/AP2γ (green), Nanog/NANOG (purple), Cdx2/CDX2 (red) or DAPI (blue) with merged and projection images. Scale bar: 25 μm.
Fig. 5: Similarities in the expression of PE-associated genes in human and mouse blastocysts. (A) Boxplots of RPKM values for selected genes. The range of expression in human EPI (green), PE (red) or TE (blue) and in mouse ICM (orange) or TE (blue). Boxes correspond to the first and third quartiles, horizontal line to the median, whiskers extend to 1.5 times the interquartile range and dots were outliers. (B) Immunofluorescence analysis of human or mouse blastocysts for Foxa2/FOXA2 (green), Sox17/SOX17 (red), Oct4/OCT4 (purple) or DAPI (blue) with merged images. Scale bar: 25 μm.
**Fig. 6:** Differences in the expression of EPI-associated genes in human versus mouse blastocysts. (A) Boxplots of RPKM values for selected genes. The range of expression in human EPI (green), PE (red) or TE (blue) and in mouse ICM (orange) or TE (blue). Boxes correspond to the first and third quartiles, horizontal line to the median, whiskers extend to 1.5 times the interquartile range and dots were outliers. (B) Boxplots of RPKM values for *Klf17/KLF17* in human or mouse at each of the selected development stages. (C) Immunofluorescence analysis of human blastocysts for KLF17 (green), NANOG (purple), CDX2 (red) or DAPI (blue) with merged image. Scale bar: 25 µm. (D) Summary of TGF-β signaling components expressed at an RPKM value >5 in human EPI or TE. Bold denotes differentially expressed genes. *Indicates genes whose expression falls just below the RPKM threshold. (E) Immunofluorescence analysis of SB-431542 treated or DMSO control human embryos for NANOG (green), OCT4 (purple), SOX17 (red) or DAPI (blue) with merged images. Scale bar: 25 µm. (F) Fluorescence intensity of NANOG, OCT4 or SOX17 in individual cells in each control or SB-431542 (SB) treated embryo. (G) Immunofluorescence analysis of SB-431542 treated mouse embryos for Nanog (green), Oct4 (purple), Sox17 (red) or DAPI (blue) with merged image. Scale bar: 25 µm. (H) Fluorescence intensity of Nanog, Oct4 or Sox17 in individual cells in each control or SB-431542 (SB) treated embryo.
Fig. 7: Defining human ground state pluripotency. (A) PCA of human EPI and hESCs grown in distinct culture conditions. Each point represents the gene expression profile of a single cell from the human EPI, single cell from Yan et al. late or early hESCs,
clumps of hESCs from either Chan et al. (3iL or mTeSR) or Takashima et al. (reset or primed). (B) Unsupervised hierarchical clustering of global gene expression of human EPI or hESCs. (C) Pearson correlation coefficient between each pair of conditions indicated. (D) Cytoscape enrichment map of GSEA results comparing human EPI (red) versus 3iL or reset hESCs (blue) (p-value < 0.01). (E) Heatmaps of selected differentially expressed genes in human EPI and hESCs. Expression levels were plotted on a high-to-low scale (purple-white-green). (F) The log2 fold change for selected genes in each condition relative to the expression of hESCs maintained on MEFs.
Supplementary Materials and Methods

Human embryo culture

Vitrified embryos frozen in straws were thawed by quickly transferring the contents of the straw from liquid nitrogen directly into thaw solution (Irvine Scientific Vitrification Thaw Kit). Embryos frozen in cryopets were first thawed for 3 seconds in a 37°C waterbath and transferred into thaw solution (Irvine Scientific Vitrification Thaw Kit). After 1 min the embryo was transferred from thaw solution into dilution solution for 4 min followed by two washes in wash solution for 4 min each (Irvine Scientific Vitrification Thaw Kit). Embryos frozen in a glass ampoules were thawed completely in a 37°C waterbath after the top of the vial was removed under liquid nitrogen. The contents were emptied onto a petri dish and the embryo transferred through a 0.5 M sucrose solution for 5 minutes, 0.2 M sucrose solution for 10 min and diluent for 10 min (Quinn’s Advantage Thaw Kit, Origio). The embryos were cultured in Global Media (LifeGlobal) supplemented with 5 mg/mL LifeGlobal Protein Supplement pre-equilibrated overnight in an incubator at 37°C and 5% CO2. Embryos cultured in
Micromanipulation

Single cells were isolated from blastocyst stage embryos (6-7 days post fertilisation) for subsequent analysis by micromanipulation, with a duration of less than 20 minutes. Embryos were placed in drops of G-MOPS solution (Vitrolife) on a petri dish overlaid with mineral oil. The plate was placed on a microscope stage (Olympus IX70) and the embryos were held with an opposing holding pipette and blastomere biopsy pipette (Research Instruments) using Narishige micromanipulators (Narishige, Japan). The biopsy mode of a Saturn 5 laser (Research Instruments) was used to separate the majority of the mural TE from the ICM and polar TE. The ICM and polar TE were washed quickly in PBS without Ca\(^{2+}\) and Mg\(^{2+}\) (Invitrogen) then placed in 0.05% trypsin/EDTA (Invitrogen) for 5 minutes at room temperature. Trypsin was quenched using Global Media supplemented with 5 mg/mL LifeGlobal Protein Supplement. After quenching, the cell clump was placed back on the stage in a drop of G-MOPS solution and pipetted up and down several times using the blastomere biopsy pipette.

cDNA synthesis and amplification

cDNA was generated from single cells using the SMARTer Ultra Low Input RNA kit for Illumina Sequencing–HV (Clontech Laboratories, Inc.) according to manufacturers’ guidelines. Single cells were picked using 100 μm inner diameter Stripper pipette (Origio) and transferred to individual low bind RNAse-free tube containing 0.25 μl RNase inhibitor, 4.75 μl Dilution buffer and 5 μl nuclease-free water on a -80°C pre-chilled CoolRack (Biocision, CA). Samples were stored at -80°C until ready to be processed. 1 μl of 3’ SMART CDS Primer II A was added to
the sample, mixed well and incubated at 72°C for 3 min. First strand cDNA was synthesised by adding 4 μl 5X First-Strand Buffer, 0.5 μl 100 mM DTT, 1 μl 20 mM dNTP mix, 1 μl SMARTer IIA Oligonucleotide, 0.5 μl RNase Inhibitor and 2 μl SMARTScribe Reverse Transcriptase (100 U/μl) directly to a tube containing the sample and incubating at 42°C for 2 hours followed by 10 min at 70°C.

First strand cDNA was purified by adding 36 μl of room temperature SPRI Ampure XP beads (Beckman Coulter Genomics), mixing well and incubating at room temperature for 8 min. Tubes were placed on a MagnaBot II Magnetic Separation device (Promega) and allowed to stand until all beads were immobilised into a pellet. The supernatant was removed and discarded. Tubes were briefly spun and any residual liquid was removed.

Double stranded cDNA was amplified from the template bound to the beads using Advantage 2 PCR kit (Clontech Laboratories, Inc.). 5 μl 10X Advantage 2 PCR Buffer, 2 μl 10 mM dNTP Mix, 2 μl IS PCR Primer, 2 μl 50X Advantage 2 Polymerase Mix and 39 μl Nuclease-Free water were added to the tube containing the sample to give a total volume of 50 μl. PCR amplification was performed at 95°C for 1 min, followed by 18 cycles of 15 sec at 95°C, 30 sec at 65°C and 6 min at 68°C followed by final extension step of 10 min at 72°C. Amplified cDNA was purified by adding 90 μl SPRI Ampure XP beads, mixing well and incubating at room temperature for 8 min to allow amplified cDNA bind to the beads. Sample tubes were placed on the magnet and allowed to stand until all beads had been immobilised. Supernatant was removed and discarded and beads were washed twice by adding 200 μl freshly prepared 80% ethanol and leaving for 30 sec before discarding the
supernatant. Tubes were spun briefly to collect residual liquid. The bead pellet was allowed to air dry. 12 μl of purification buffer was added to rehydrate the pellet and incubated for 2 min at room temperature. cDNA was eluted by pipetting up and down 10 times before returning the tube to the magnet. The clear supernatant containing the cDNA was removed from the immobilised beads and transferred to a new low-bind tube. cDNA was stored at -80°C until library preparation.

cDNA quality was assessed by High Sensitivity DNA assay on an Agilent 2100 Bioanalyser with good quality cDNA showing a broad peak from 300 to 9000 bp. cDNA concentration was measured using QuBit dsDNA HS kit (Life Technologies UK Ltd.) Typical yields from a single cell ranged from 1 ng to 9 ng.

**cDNA shearing and library preparation**

In preparation for library generation cDNA was sheared using Covaris S2 to achieve cDNA in 200-500 bp range. 10 μl of cDNA sample and 65 μl purification buffer was added to Covaris AFA Fiber Pre-Slit Snap Cap microTUBE. cDNA was sheared using the settings 10% Duty, Intensity 5, Burst Cycle 200 for 2 min. Sheared cDNA was transferred to a new 0.2 ml low-bind tube.

Libraries were prepared using Low Input Library Prep Kit (Clontech Laboratories, Inc.) according to manufacturer’s instructions. The amount of input cDNA was calculated from the concentration measured by the Bioanalyser assay prior to shearing, taking into account the dilution involved in the shearing step. The appropriate amplification cycle number was selected according to manufacturer’s
guidelines. Library quality was assessed by Bioanalyser and the concentration was measured by QuBit assay. The molar concentration of library was calculated thus:

Library molecular weight = average size in bp (from Bioanalyser) x 650 g/mol per bp
Molar concentration = library concentration from QuBit/library molecular weight
Libraries with a molar concentration greater than 2nM were submitted for 50-bp paired-end sequencing on Illumina HiSeq 2000.

Data acquisition and processing

We integrated previously published datasets with our own blastocyst sequencing data using a consistent read alignment method. SRA files were obtained via ftp from the Gene Expression Omnibus, under the accession numbers GSE36552 and GSE45719. The SRA files were converted into FASTQ format using the fastq-dump program from the SRA toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/). The reference human genome sequence was obtained from Ensembl, along with the gene annotation (GTF) file. The reference sequence was indexed using the bowtie2-build command.

Read mapping and counting

Reads were aligned to the reference human genome sequence using Tophat2 (Kim et al., 2013), with gene annotations to obtain BAM files for each of the single-cell samples. BAM files were then sorted by read coordinates and converted into SAM files using SAMtools. The process of mapping and processing BAM files was automated using a custom Perl script. The number of reads mapping to each gene were counted using the program htseq-count (Anders et al., 2015). The resulting count files for each sample were used as input for differential expression analysis using DESeq using the hg19 human or mm9 mouse genome reference sequence.
**Expression analysis**

To investigate differences in global gene expression, a PCA of the top 8000 genes with the most variable expression was performed on the human and mouse RPKM data separately. The R package prcomp was used to generate the PCA, using both the scaling and centering options. The R package NOISeq (Tarazona et al., 2011) was used to identify genes differentially expressed between the TE and EPI cells in human, and between TE and ICM cells in mouse. To increase sensitivity, genes with an RPKM > 5 in four or more samples were retained for NOISeq analysis. Differentially expressed genes were identified after applying a 95% probability threshold. Ensembl Biomart was used to find human-mouse orthologous pairs within the list of differentially expressed human and mouse gene.

We used a second independent method to detect differentially expressed genes between human EPI and TE using DESeq (Anders and Huber, 2010). Firstly, the function `estimateSizeFactors` and `estimateDispersions` were used to estimate biological variability and calculate normalised relative expression values across the different blastocyst samples. Initially, this was performed without sample labels (option: `method='blind'`) to allow unsupervised clustering of the blastocyst samples using principal components analysis and hierarchical clustering. The dispersion estimates were recalculated with the sample labels included and with the option: `method='pooled'`. The function `nbinomTest` was then used to calculate p-values to identify genes that show significant differences in expression between different cell types.
A k-means clustering analysis was performed to find clusters of genes co-expressed during pre-implantation development. The mean RPKM value for each developmental time point was calculated for subsequent k-means clustering analysis (Figs S1, S2). Genes with a fold change of greater than two between any two stages were retained. The R package ‘MFuzz’ was used to generate the k-means clusters using the kmeans2 function, with the number of clusters set to 50. A custom R script was used to generate plots for the k-means clusters and trendlines were drawn based on the k-means centroids. The k-means clusters were clustered further using the R function ‘hclust’ and heatmaps were generated using R package ‘pheatmap’.

**PCA comparison of EPI versus hESC gene expression**

We compared the EPI single cell RNA-seq dataset to distinct hESC lines (Yan et al., Takashima et al. and Chan et al.) (GSE36552, E-MTAB-2857, E-MTAB-2031). These data were processed using our computational pipeline to generate read counts and RPKM values for each gene. NOISeq was used to perform differential expression analysis of the EPI versus hESC samples. Samples from Yan et al. were grouped into early or late hESCs, and NOISeq analysis was performed independently on these two groups. Samples from Chan et al. and Takashima et al. were grouped into primed or reset hESCs, and NOISeq analysis was performed independently on each of the groups. RPKM values for genes showing differential expression in least one of pairwise test were used to generate a PCA plot, showing the relationship of gene expression between the hESC lines and the EPI. In addition, the Pearson correlation coefficient was calculated using the median RPKM between each pair of conditions. The R function ‘pairs’ was used to generate scatterplots comparing the median RPKMs between each pair of conditions.
Pathway enrichment analysis

The GSEA method (Subramanian et al. 2005) was used to identify pathways and GO terms enriched in each hESC line versus the human EPI, the human trophectoderm versus the human EPI or the mouse ICM versus the TE. We used a compendium of multiple pathway interaction databases downloaded from the Bader Lab: http://baderlab.org

Genes were ranked according to log2fold change between the each pairwise comparison. The resulting .rnx files were then used as input for the GSEAPreranked module and enrichment analysis was performed. The GSEA output files were loaded into the Cytoscape module EnrichmentMap (Shannon, P. 2003; Merico, D. 2010) and the relationship between the expressed signaling pathways was displayed in an interaction network map.

DESeq Analysis

For DESeq analysis, firstly, the function ‘estimateSizeFactors’ and ‘estimateDispersions’ were used to estimate biological variability and calculate normalised relative expression values across the different blastocyst samples. Initially, this was performed without sample labels (option: method='blind') to allow unsupervised clustering of the blastocyst samples using principal components analysis and hierarchical clustering. The dispersion estimates were recalculated with the sample labels included and with the option: method='pooled'. The function ‘nbinomTest’ was then used to calculate p-values to identify genes, which show significant differences in expression between different cell types.
Quantification of Immunofluorescence

MINS 1.3 software was used to detect and segment nuclei and generate tables of fluorescence intensity for each channel (http://katlab-tools.org/) (Lou et al Stem Cell Reports 2014). Embryos were imaged at a z-section thickness of 3μm. Confocal stacks in .tif format were loaded into the MINS pipeline for automated nuclear segmentation. The MINS segmentation output was manually checked for appropriate segmentation and tables were amended accordingly. Mitotic nuclei were removed from the analysis and the background adjusted using a method described previously (Schrode et al., Dev Cell 2014). Data were subsequently plotted using GraphPad Prism version 6 (GraphPad Software, La Jolla, CA).
Figure S1. Time course plots for 50 k-means clusters generated from the human single-cell RNA-seq dataset. RPKM values were averaged for each developmental stage and invariable genes filtered prior to clustering. Black lines indicate median expression profiles.
Figure S2. Time course plots for 50 k-means clusters generated from the mouse single-cell RNA-seq dataset. RPKM values were averaged for each developmental stage and invariable genes filtered prior to clustering. Black lines indicate median expression profiles.
**Figure S3.** (A) Estimating the optimum number of k-means clusters based on the Bayesian Information Criterion (BIC). RPKM values from the human dataset were averaged within each developmental stage and invariable genes filtered prior to BIC estimation. The BIC score was plotted for up to 150 clusters and shows the inflection point lies at approximately 40-50 clusters. (B-C) Heatmap of the hierarchical clustering of (B) human and (C) mouse k-means clusters reveals global gene expression patterns across time. Expression levels were plotted on a high-to-low scale (purple-white-green).
Figure S4. Lineage-specific gene expression from the single-cell RNA-seq data. (A) Dendrogram representing hierarchical clustering of human blastocyst single-cell samples from Yan et al., 2013. (B) PCA incorporating our additional single-cell RNA-seq samples (red) together with the samples (zygote to blastocyst) from Yan et al., 2013. (C) Differentially expressed genes identified by DESeq. The log2 fold change (FC) difference in expression is noted. (D) Loading scores for first two principal components based on the human RNA-seq data. Each point represents the loading score for a single gene. (E) PCA of single-cell RNA-seq data from blastocyst samples, coloured according to lineage assignments made by Yan et al., 2013.
Figure S5. Cytoscape enrichment map of GSEA results comparing epiblast (EPI, red) versus human embryonic stem cells (blue) (p-value < 0.01) or EPI versus TE (p-value < 0.05). Differentially expressed genes were identified using gene set enrichment analysis using a combination of pathways compiled by the Bader lab (http://baderlab.org).
Figure S6. Immunofluorescence analysis of (A) human and (B) mouse blastocysts. The expression of Ap2γ/AP2γ, Nanog/NANOG, Cdx2/CDX2 or DAPI are indicated in green, purple, red or blue, respectively. The merged image and projection of expression is shown. The scale bar is 25 µm.
Figure S7. Immunofluorescence analysis of (A) human and (B) mouse blastocysts. The expression of Foxa2/FOXA2 (green), Oct4/OCT4 (purple), Sox17/SOX17 (red) or DAPI (blue) with merged images. Scale bar: 25 μm.
Figure S8. Human embryonic stem cells (H9) were cultured on Matrigel coated glass bottom microwell dishes (MatTek) in mTeSR for 3 days in the presence or absence of 40 μM SB-431542 then immunofluorescently analysed for the expression of NANOG (red) and DAPI (blue). Scale bar: 100 μm.
Table S1

Click here to Download Table S1

Table S2

Click here to Download Table S2

Table S3

Click here to Download Table S3

Table S4

Click here to Download Table S4

Table S5

Click here to Download Table S5