Unique properties of a subset of human pluripotent stem cells with high capacity for self-renewal

Kevin X. Lau1,12, Elizabeth A. Mason1,12, Joshua Kie1, David P. De Souza3, Joachim Kloehn4, Dedreia Tull3, Malcolm J. McConville3,4, Andrew Keniry5,6, Tamara Beck5, Marnie E. Blewitt5,6, Matthew E. Ritchie5, Shalin H. Naik5, Daniela Zalcenstein5, Othmar Korn7, Shian Su5, Irene Gallego Romero8, Catrina Spruce9, Christopher L. Baker9, Tracy C. McGarr9, Christine A. Wells1,2,10 & Martin F. Pera1,5,9,11

Archetypal human pluripotent stem cells (hPSC) are widely considered to be equivalent in developmental status to mouse epiblast stem cells, which correspond to pluripotent cells at a late post-implantation stage of embryogenesis. Heterogeneity within hPSC cultures complicates this interspecies comparison. Here we show that a subpopulation of archetypal hPSC enriched for high self-renewal capacity (ESR) has distinct properties relative to the bulk of the population, including a cell cycle with a very low G1 fraction and a metabolomic profile that reflects a combination of oxidative phosphorylation and glycolysis. ESR cells are pluripotent and capable of differentiation into primordial germ cell-like cells. Global DNA methylation levels in the ESR subpopulation are lower than those in mouse epiblast stem cells. Chromatin accessibility analysis revealed a unique set of open chromatin sites in ESR cells. RNA-seq at the subpopulation and single cell levels shows that, unlike mouse epiblast stem cells, the ESR subset of hPSC displays no lineage priming, and that it can be clearly distinguished from gastrulating and extraembryonic cell populations in the primate embryo. ESR hPSC correspond to an earlier stage of post-implantation development than mouse epiblast stem cells.
The successful application of human pluripotent stem cells (hPSC) in research and cell therapy relies on the ability to maintain, expand, and differentiate these cells in vitro in a tightly controlled and efficient fashion. Our understanding of the regulation of pluripotent stem cell self-renewal and lineage specification is in turn built largely on embryological paradigms, with developmental roadmaps providing critical knowledge of the key transitional stages, and pinpointing the extrinsic and internal molecular pathways drive cell fate decisions. In the mouse, we now have a fairly clear understanding of the states of pluripotency that span the developmental stages between the blastocyst and the late gastrula in vivo. The characterization of mouse naive embryonic stem cells (ESCs) and epiblast stem cells (EpiSCs) as in vitro equivalents of the preimplantation epiblast and the anterior primitive streak, respectively, has shed considerable light on the properties of the cultured cells. Recently, a stage between these two pluripotent states called formative pluripotency, corresponding to the early post-implantation epiblast, has been described. Defined by specific molecular and biological features, including an absence of lineage priming and a rapid response to induction of lineage specification (including the germline lineage), formative mouse pluripotent stem cells have yet to be successfully serially propagated in vitro.

The derivation of mouse EpiSCs, their characterization as epithelial cells dependent upon activin and FGF2 for maintenance in the pluripotent state, and their co-expression of lineage-specific and pluripotency genes as pluripotent state, and their co-expression of lineage-specific and pluripotency genes, led many researchers to the conclusion that hPSC derived and maintained under conventional culture conditions (archetypal hPSC), which share these features, equate to the primed state of mouse pluripotency. This in turn led to a search for conditions that would support long-term maintenance of hPSC in a naive state. Several culture systems have been described that support a cell with molecular features quite similar to the human preimplantation epiblast. Extended propagation of diploid hPSC in these systems remains challenging, however.

Population heterogeneity complicates the interpretation of stem cell phenotype. To dissect heterogeneous populations in archetypal hPSC cultures, we used monoclonal antibodies to cell surface antigens to define subsets of stem cells that exist in a hierarchical continuum of cell states. When we subjected these cells to transcriptional profiling, we observed co-expression of pluripotency genes with lineage specific transcription factors, particularly in subpopulations of cells with lower levels of stem cell surface marker and pluripotency gene expression. Importantly though, we also found that cells which expressed pluripotency markers at high levels were less likely to display lineage priming.

Subsequently by analyzing cells grown under defined conditions, and using a more refined sorting strategy to isolate the subpopulation enriched for high self-renewal (ESR) followed by medium throughput single-cell RT-QPCR, we were able to show that the cells at the top of the hierarchy expressed very uniform and high levels of pluripotency markers, and showed no lineage priming.

As cells traverse through stages of pluripotency in vitro and in vivo, they undergo changes in cell cycle regulation and metabolic activity, they restructure their epigenome, and their gene expression profile changes. The ability to sort highly purified populations of hPSC with high self-renewal capacity to analyse transcription in these subpopulations at the genomic level using RNA-Seq for subpopulation and single cell analysis, coupled with the availability of new single-cell gene expression data from preimplantation and post-implantation primate embryos, prompted us to re-examine the properties of the ESR subpopulation of archetypal hPSC, and to reconsider its developmental status. The results show that the ESR subpopulation resembles the primate early post-implantation epiblast, similar to the mouse formative state of pluripotency.

**Results**

**Self-renewal of subsets of hPSC.** We have previously used flow cytometry to isolate cell subpopulations, followed by assay of colony forming ability as an indicator of self-renewal. hPSC survival after dissociation to single cells and flow cytometry is poor. We therefore developed a simple methodology that would allow comparison of self-renewal of defined cell populations at reasonable levels of initial survival after flow cytometry, through sorting small aggregates of cells, to maintain cell–cell contacts and enhance post-sort survival.

Fluorescence activated cell sorting using antibodies GCTM-2 and TG30 (recognizing CD9) enabled us to recover small cell aggregates (chiefly doublets; singlets, 17%; doublets, 61%; triplets, 16%; quadruplets, 6%) (Fig. 1a). We seeded wells with equal cell numbers as aggregates or singlets. Although both aggregates and single cells attached to the plate after sorting, a much larger fraction of the aggregates had begun to spread 1 h after plating, indicative of high viability; the difference was more evident after 24 h (Fig. 1b).

Initial attachment, spreading and survival of the GCTM-2highCD9high subpopulation was similar to the GCTM-2midCD9mid subpopulation. Flow cytometry re-analysis of both subpopulations 72 h later showed that they had largely retained their cell surface phenotypes though as shown previously, the GCTM-2highCD9high population had begun to reconstruct the entire cell state continuum (Fig. 1c). By 4 days, cells plated as aggregates displayed a higher colony forming efficiency than single cells, and the GCTM-2highCD9high subpopulation had formed a much larger number of microcolonies (Fig. 1d–e). Colonies formed from GCTM-2highCD9high aggregates were larger and showed a higher proportion of cells bearing stem cell markers compared with colonies formed by GCTM-2midCD9mid cells (Fig. 1f). Time-lapse video microscopy confirmed that the initial cell numbers attaching to the dish were similar for the two subpopulations. However, subsequent monitoring showed that while both subpopulations were migratory and underwent cell division, the GCTM-2highCD9high subpopulation persisted to form colonies of 4–32 cells several days later, whereas the GCTM-2midCD9mid colonies suffered abortive expansion and, in many cases, extinction (Supplementary Movies 1 and 2).

**Differentiation potential of hPSC subpopulations.** In the mouse, the ability to differentiate efficiently into germ cells in vitro is limited to epiblast like cells at the stage of formative pluripotency. Naive pluripotent stem cells or epiblast stem cells both lack this capacity. Using the two-step protocol to generate first a mesoderm-like intermediate and subsequently convert these cells to PGC-like cells, we measured the degree of differentiation of the GCTM-2highCD9highEP CAMhigh and GCTM-2midCD9mid populations using flow cytometry to quantify EPCAM/ITGA6 double high cells which represent PGC-like cells (Fig. 2a). GCTM-2highCD9highEP CAMhigh cells showed higher expression of EPCAM and ITGA6 at the onset, as expected. After 2 days of differentiation into mesoderm-like cells, both groups showed loss of EPCAM expression but at Day 4 after PGC-like cell induction, an EPCAM/ITGA6 double positive fraction was observed in both groups. Cultures that did not receive growth factors to induce germ cell formation contained no EPCAM/ITGA6 double high population. The GCTM-2highCD9high EP CAMhigh population had largely disappeared by the end of the time course, as noted by Sasaki et al. The identity of the PGC-like cells was confirmed by staining with antibodies to PRD1
and NANOS3 (Fig. 2b). More PGC-like cells were consistently obtained from GCTM-2\textsuperscript{high}CD9\textsuperscript{high}EPCAM\textsuperscript{high} compared with GCTM-2\textsuperscript{mid}CD9\textsuperscript{mid} subpopulations in both cell lines, though there was a high degree of inter-assay variability (Fig. 2c). We and others have shown previously that the ESR fraction of hPSC is pluripotent (Figs. 4 and 5, refs. 17,19). In this study, directed differentiation in adherent culture, carried out on cell lines WA09 and WA01, confirmed that both the GCTM-2\textsuperscript{high}CD9\textsuperscript{high} subpopulation and the remaining cell population were pluripotent, as measured by expression of markers characteristic of progenitors of all three embryonic germ layers (Fig. 2d).

**ESR cells have a cell cycle with a low G1 fraction.** Initial analyses of the mouse ESC cell cycle indicated that naive mouse ESC, and cells of the epiblast, show a shortened cell cycle with a minimal
G1 component. While more recent work shows that the cell cycle state of naive mouse stem cells is dependent upon culture conditions, it remains clear that lineage commitment is coupled to cell cycle regulation in mouse and human PSC, and that the G1 phase represents a decision point for undergoing lineage specification.

Here we re-assessed our previous finding that a subset of cells expressing high levels of the stem cell antigen GCTM-2 had a reduced G1 fraction compared with the remainder of the population. Using a more refined sorting procedure and EdU incorporation to identify S-phase cells, we determined the cell cycle phase distribution of GCTM-2highCD9highEPCAMhigh, GCTM-2highCD9high, and GCTM-2lowCD9low subpopulations, compared with unsorted cells (the general population). As shown in Fig. 3, very few of the GCTM-2highCD9highEPCAMhigh or GCTM-2highCD9high cells were in G0/G1 phase, with most of this subpopulation in S or G2/M. By contrast, GCTM-2lowCD9low cell populations were predominantly in G0/G1 (~70%). The cell cycle phase distribution of the unsorted general population was consistent with these findings. Results using a different WA09 subline with a Fucci reporter confirmed these conclusions (Supplementary Fig. 1).

Active mitochondria and bivalent metabolism in ESR cells. Metabolic activity is modulated throughout early mammalian development. The preimplantation relies on a combination of aerobic glycolysis and oxidative phosphorylation, and this metabolic pattern is maintained during post-implantation development up to E7.5 in the mouse. Like the epiblast, naive state mouse ESC show this bivalent metabolism, while primed epiblast stem cells rely primarily on glycolysis. Similar transitions from bivalent metabolism to glycolysis have been reported during conversion of naive to primed hPSC.

To determine if the energy metabolism status of hPSC is dependent on their position in the pluripotency hierarchy, we first assessed the mitochondrial membrane potential of defined subpopulations of hPSC by measuring the uptake of the mitochondrial dye tetramethylrhodamine methyl ester (TMRM), or the combined uptake and the ratio of red/green fluorescence of the dye JC-1. Live cell staining of WA09 cells showed that staining with TMRM was strongest at the edge of the colonies, where cells expressing the highest levels of the stem cell marker CD9 are found (Fig. 4a). To quantitate mitochondrial activity across the cell populations, WA09 cells were incubated with TMRM or JC-1, labeled with stem cell surface makers GCTM-2, TG30 (anti-CD9), and anti-EPCAM, and the GCTM-2highCD9highEPCAMhigh and GCTM-2lowCD9low fractions were identified by flow cytometry. The GCTM-2highCD9highEPCAMhigh subpopulation stained more intensely with TMRM, and showed higher ratio of red to green fluorescence following incubation with JC-1 compared with the general (remaining) population, or the GCTM-2lowCD9low subpopulation (Fig. 4b–c), indicating increased mitochondrial activity in the GCTM-2highCD9highEPCAMhigh subpopulation.

Analysis of mitochondrial oxidative phosphorylation in live cells using the Agilent Seahorse XF apparatus confirmed that the basal oxygen consumption rate (OCR) was higher in the GCTM-2highCD9highEPCAMhigh subpopulation compared with the remaining GEN population (Fig. 5a). These cells also exhibited higher spare mitochondrial capacity, as indicated by maximum OCR achieved after addition of proton uncoupling agent FCCP (Fig. 5a). Comprehensive analysis of intracellular metabolite levels in the GCTM-2highCD9highEPCAMhigh cells and unfraccionated population using liquid chromatography and gas chromatography–mass spectrometry (LC–MS, GC–MS) provided further evidence that the GCTM-2highCD9highEPCAMhigh cells exist in a distinct metabolic state. Principal Component Analysis (PCA, Supplementary Fig. 2a, b) and hierarchical cluster analysis (Fig. 5b, LC–MS; Supplementary Fig. 2c, GC–MS) clearly separated the purified cell population from the GEN population. Consistent with the live cell staining and OCR analysis, the GCTM-2highCD9highEPCAMhigh contained elevated levels of TCA cycle metabolites (Fig. 5c) and was depleted in many amino acids and metabolites in the urea cycle (Supplementary Tables 1–2), compared with the general population. Pathway analysis further confirmed the distinct metabolic profile of the GCTM-2highCD9highEPCAMhigh subpopulation (Supplementary Fig. 2d, e).

To gain further insights into the metabolic wiring of this subpopulation of cells, GCTM-2highCD9highEPCAMhigh fractionated cells and the GEN population of hPSC were cultivated in the presence of 13C-glucose for 2 h and level of 13C-enrichment in different intermediates of central carbon metabolism monitored by GC–MS. High levels of 13C-enrichment were observed in all intermediates in glycolysis and the pentose phosphate pathway confirming that both cell populations exhibit high rates of aerobic glycolysis (Fig. 5d). Interestingly, 13C-enrichment in serine and glycine, which are synthesized from the glycolytic intermediate 3-phosphoglycerate, were higher in the GCTM-2highCD9highEPCAMhigh cells, indicating that these cells may have higher rates of amino acid and protein synthesis. Consistent with GCTM-2highCD9highEPCAMhigh cells having elevated rates of mitochondrial metabolism, 13C-enrichment in citrate and isocitrate were substantially upregulated in this subpopulation (Fig. 5d). However, 13C-enrichment in later intermediates in the oxidative cycle (succinate, fumarate, and malate) were not significantly changed between the two cell populations indicating that early intermediates in the TCA cycle may be diverted into anabolic pathways (catapleurosis). Increased catapleurosis in the GCTM-2highCD9highEPCAMhigh cells was supported by the elevated levels of 13C-labeling in glutamate, which is synthesized from the TCA cycle intermediate, α-ketoglutarate, as well as high levels of 13C-enrichment in long chain unsaturated fatty acids (i.e., oleic acid) and cholesterol, indicating high rates of membrane biogenesis or turnover. Overall, these data provide compelling evidence that the GCTM-2highCD9highEPCAMhigh cells are metabolically more active than the GEN hPSC population.
exhibiting higher rates of oxidative phosphorylation and anabolic amino acid and lipid synthesis.

**DNA methylation in hPSC cultured under defined conditions.**

Levels of DNA methylation in the mouse and human epiblast are low, but increase substantially in the mouse following embryo implantation and activation of the de novo methyltransferases Dnmt3a and Dnmt3b. Reduced representation bisulfite sequencing to assess levels of DNA methylation showed no major differences between the relatively low mean levels of overall DNA methylation or levels of DNA methylation over CpG islands between the GCTM-2<sup>high</sup>CD9<sup>high</sup>EPCAM<sup>high</sup> subpopulation and the general population (Fig. 6a, b). DNA methylation distributions were bimodal (Fig. 6c, d), and there was little evidence of differential methylation at CpG islands across any particular loci in the GCTM-2<sup>high</sup>CD9<sup>high</sup>EPCAM<sup>high</sup> subpopulation (Fig. 6e). De novo DNA methyltransferases DNMT3A and DNMT3B were both expressed...
along with TET1 across the cell populations studied (below and Supplementary Fig. 3a–j), similar to the early post-implantation epiblast in the mouse36, and suggestive of a highly dynamic state of DNA methylation in these cells.

Differential chromatin accessibility in hPSC subpopulations. We next identified critical differences in the chromatin landscape between GCTM-2highCD9highEPCAMhigh and GCTM-2midCD9mid populations using the assay for transposable accessible chromatin37. In total, we identified 118,442 regions as accessible across both populations. Of these, 3144 were more accessible in the GCTM-2highCD9highEPCAMhigh cells, while 4730 were more accessible in the GCTM-2midCD9mid population (Fig. 7a, FDR < 0.01; quality assessment data Supplementary Fig. 4, Supplementary Data 1). Generally, peaks with increased accessibility in the high population were distant from transcription start sites (TSS), more often found in intergenic and introns typical of enhancers, whereas peaks with increased accessibility in the middle populations were closer to TSS (Fig. 7b) identified as promoters (Supplementary Fig. 5a). To annotate open chromatin sites in each population we compared their location to the entire set of human cell types identified by the ENCODE project38 using a locus overlap enrichment analysis39 (Supplementary Data 2). Genomic locations more accessible in the GCTM-2highCD9highEPCAMhigh population were highly enriched for TF binding sites identified in hPSC compared with the GCTM-2midCD9mid population (Fig. 7c). Furthermore, the most enriched TFs binding at the GCTM-2highCD9highEPCAMhigh population sites included known pluripotency factors NANOG, POU5F1 (OCT4), and TCF12 and BCL11A (Fig. 7d, upper panel), the latter two TF having been previously identified as highly expressed in the GCTM-2highCD9highEPCAMhigh population16. While these same factors were still identified as enriched in the regions in the GCTM-2midCD9mid populations, other general chromatin factors showed the highest overlap (Fig. 7d, lower panel). Finally, we compared regions of increased accessibility in both populations to the tissue and cell-type specific DNAse hypersensitivity clusters identified in human samples40. Again, this analysis identified that the regions with greatest enrichment in the GCTM-2highCD9highEPCAMhigh population were annotated as being unique to stem cells (Supplementary Fig. 5b, c). In summary, these data indicate that the GCTM-2highCD9highEPCAMhigh population has greater open chromatin at putative enhancers bound by canonical pluripotent factors at sites unique to stem cells, whereas the low population has greater DNA accessibility at promoters bound by general chromatin and transcription factors. Together these data support the hypothesis that the chromatin of cells in the GCTM-2highCD9highEPCAMhigh population exists in a distinct state compared with the low population.

Comparison of ESR1 cell transcriptome with primate epiblast. RNA-seq analysis comparing gene expression in the GCTM-2highCD9highEPCAMhigh subpopulation to the general (total unfractionated) population (Supplementary Data 3) identified 515 genes differentially expressed between the GCTM-2highCD9highEPCAMhigh subset and the total population (132 upregulated in the GCTM-2highCD9highEPCAMhigh cells and 383 downregulated relative to the unfractionated cells, at >1.5-fold change in expression level with an adjusted p value <0.05; Fig. 8a). Genes upregulated in the GCTM-2highCD9highEPCAMhigh high fraction included NODAL and its antagonists LEFTY1 and LEFTY2, in agreement with our previous study16. POUSF1, a marker of the naive to formative transition in the mouse, was also upregulated. Notably, a number of small nuclear and small nucleolar RNAs were expressed at high levels in the GCTM-2highCD9highEPCAMhigh population. Negative regulators of MAPK signaling, including DUSP5 (inactivator of ERK1), DUSP6 (inactivator of ERK2), and SPRY2, were upregulated in the self-renewing fraction, as was DACT1 (an antagonist of canonical WNT signaling). Negative regulators of MAPK signaling including Dusp4 and Spry were recently shown to be upregulated at an early stage during dissolution of the mouse naive state41. Amongst the genes expressed at lower levels in the GCTM-2highCD9highEPCAMhigh cellular subset relative to the unfractionated population were members of the WNT signaling pathway, including WNT4, FRZB, FZD3, FZD5, and FZD8. Of the top 100 genes upregulated in the general population, 47 genes (all upregulated at twofold change or higher) were previously reported to be expressed at peak levels at the onset of neural differentiation in the CORTECON study of Temple and colleagues42 (global analysis, Supplementary Fig. 6). Examination of previously published microarray data43 for a subset of these neural induction genes confirmed a pattern of continuous upregulation in cell subsets with decreasing levels of pluripotency associated cell surface markers in multiple cell lines (Supplementary Fig. 7, data visualized in the Stemformatics platform https://www.stemformatics.org). BMP2, BMP4, and FST were also upregulated in the general population, consistent with our previous results16.

Some genes characteristic of the primate preimplantation epiblast and naïve hPSC were expressed in the GCTM-2highCD9highEPCAMhigh fraction (PRDM14, TCFCP2L1, ZFP42, DPPA2, and TFAP2C), but others were not (ARGFX, KLF17, TXB3, NLRP7). Genes expressed in primitive endoderm (SOX17, GATA4, GATA6, FOXA2, and APOA2) were not found in either the GCTM-2highCD9highEPCAMhigh or the general population, nor were genes activated during early gastrulation (T, MIAxl, GSC, EOMES, FOXA2, LHX1).

scRNA-seq on hPSC fractionated into four separate subpopulations enabled us to compare gene expression in the fractionated subpopulations with the single-cell data of Nakamura et al.14 for Macaca fascicularis pre- and post-implantation embryos (quality
We analyzed 300 cells, and we detected expression of 8403 genes. All cells uniformly expressed the general pluripotency associated transcription factors POU5F1, SOX2, and NANOG (Supplementary Fig. 10a–d). ZFP42, a marker of the naive state in mouse, was expressed throughout the population, but another naive state marker, TFCPL1 was expressed primarily in the GCTM-2highCD9highEPCAMhigh subsets; POU3F1, characteristic of post-implantation epiblast, was expressed throughout (Supplementary Fig. 10e–g). PCA of the human cell subpopulations alone indicated that they could be clearly separated along a continuum of cell states (Fig. 8b). Ontology analysis of differential gene expression highlighted a number of pathways involving ribonucleoprotein complexes, ribosomes, and metabolic processes, in concurrence with the expression of small nuclear and nucleolar RNAs noted above in the subpopulation analysis (Supplementary Table 3), and a number of pathways related to oxidative metabolism, including metabolic processes, mitochondrion, mitochondrion organization, generation of subpopulations.
grown in proprietary serum replacement with mouse embryo of green to red emission, indicative of high mitochondrial membrane potential. Results in a display representative outcomes from three experiments.

similarly de microarray analyses: an independent study that examined these findings, we compare these results with two previous microarray analyses: an independent study that examined similarly defined subpopulations of cell lines MEL1 and WA09 grown in proprietary serum replacement with mouse embryo precursors metabolites and energy, mitochondrial inner membrane, and hydrogen ion transmembrane transporter activity.

The GCTM-2_highCD9_highEPCAM_high subset and general populations both expressed markers of the post-implantation mouse epiblast, including POU3F1, OTX2, DNMT3A, DNMT3B, SOX4, SOX11, LIN28A, and ZNF281 (Supplementary Data 3). In a PCA of the human single-cell data and the cynomolgus data of Nakamura et al.44 (Fig. 8c), the first principal component resolved the two experiments with cynomolgus and human cells. The second principal component resolved the inner cell mass and preimplantation epiblast, along with extraembryonic cells. In this dimension, the human cells aligned with preimplantation epiblast stage cells. In the third principal component, which separated post-implantation stages of cynomolgus development, human cells aligned with the post-implantation epiblast, with cells in the GCTM-2_highCD9_highEPCAM_high fraction between early and late post-implantation stages. The human cells were clearly distinguished from inner cell mass and preimplantation epiblast stages, from gastrulating cells, and from extraembryonic tissues. PCA of the cynomolgus cells alone revealed that it was difficult to separate early from late post-implantation epiblast (Supplementary Fig. 11).

The cells used in this study were WA09 cells grown on mTeSR in feeder-free, serum-free conditions. To assess the generality of these findings, we compare these results with two previous microarray analyses: an independent study that examined similarly defined subpopulations of cell lines MEL1 and WA09 grown in proprietary serum replacement with mouse embryo fibroblast feeders or mTeSR1 (WA09)43, and our previous study using cell line ES02 grown in serum-supplemented medium in the presence of mouse fibroblast feeder cells17. We identified a panel of stage-specific genes (Supplementary Table 4) on the basis of their differential expression in the data of Nakamura et al.44 and a recent scRNA-seq study of the human preimplantation embryo45 (Fig. 9). Genes specific to the inner cell mass, or mainly expressed in the inner cell mass and preimplantation epiblast, were very weakly expressed in the GCTM-2_highCD9_high population in the previous works and in the GCTM-2_highCD9_highEPCAM_high cells in the current RNA-seq study. Expression levels of the gene panel characteristic of the inner cell mass, preimplantation epiblast and early post-implantation epiblast were found at appreciable levels in self-renewing hPSC in all studies. Those genes with highest expression levels in all three cynomolgus epiblast populations, or in early and late post-implantation epiblast, were expressed robustly in all hPSC subpopulations in all studies. The gene panel specific to late post-implantation epiblast and gastrulating epiblast stages was expressed at somewhat lower levels in all of our populations relative to pan-epiblast specific genes. Last, genes characteristic of gastrulation stages 2A and 2B (nomenclature, ref. 44) were expressed at low levels in the GCTM-2_highCD9_highEPCAM_high and GCTM-2_highCD9_high and GCTM-2_midCD9_mid subpopulations, with levels rising in the GCTM-2_lowCD9_low and GCTM-2~CD9~ fractions. Thus, these data support the same conclusion as PCA of the single cell data: hPSC cell subpopulations enriched for self-renewal capacity show a pattern of gene expression that is strongly similar to early

---

**Fig. 4 Mitochondrial activity in subpopulations of hPSC.** a Double label staining of live cells with TMRM and stem cell surface marker CD9. Cells at the edges of colonies stain most strongly with antibody and dye. Scale bar = 100 micron. b Flow cytometry analysis of TMRM staining in GCTM-2_highCD9_highEPCAM_high (HHH), GCTM-2_lowCD9_low (LOW), and unsorted (GEN) population. c Flow cytometry analysis of JC-1 dye emission in GCTM-2_highCD9_highEPCAM_high (HHH), GCTM-2_lowCD9_low (LOW), and unsorted (GEN) population. GCTM-2_highCD9_highEPCAM_high cells display highest ratio of green to red emission, indicative of high mitochondrial membrane potential. Results in a display representative outcomes from three experiments.
post-implantation epiblast stages in the primate embryo, but clearly distinguished from inner cell mass or gastrulation stages.

**Discussion**

We showed previously that the minority ESR subpopulation of hPSC can be isolated with cell surface markers and identified by colony formation assay. Colony formation measures both survival and self-renewal, and dissociation to single cells and flow cytometry compromises survival. We have used several assay strategies to avoid conflation of survival and self-renewal\textsuperscript{16,17}, but the approach described in this study of isolating aggregates is simple, and yields defined subpopulations with high initial survival for
subsequent analysis. Although the initial survival of aggregates of the hPSC subpopulations was similar, further development of viable stem cell colonies was observed predominantly in the fraction bearing the highest level of stem cell markers. Time-lapse video microscopy revealed that cells in the GCTM-2.hiCD9.hiEPCAM.hi fraction formed microcolonies that persisted during extended propagation, whereas microcolonies of cells in the lower fraction underwent gradual extinction. This is similar to the observations of Barbaric et al.46, who found that only a subset of SSEA3-positive hPSC formed microcolonies that persisted and grew. Their findings and ours suggest that self-renewal might be a function of the formation of a critical mass of cells expressing...
survival or growth factors at a sufficiently high local level. We have shown that cells in the GCTM-2\textsuperscript{high}CD9\textsuperscript{high} or GCTM-2\textsuperscript{high}CD9\textsuperscript{high}EPCAM\textsuperscript{high} fractions express the highest levels of components of the NODAL signaling pathway\textsuperscript{16,17}. Recent results in zebrafish indicate that cell–cell contacts are key to Nodal signaling\textsuperscript{47}, suggesting the possibility that hPSC might similarly depend on a positive feedback loop of NODAL signaling and cell–cell adhesion to drive self-renewal.

In the mouse, under appropriate culture conditions, naive cells pass through an intermediate state between naive and primed pluripotency, and in this transient state, are competent to undergo germline differentiation\textsuperscript{21}. It was previously reported
that archetypal hPSC can form primordial germ cell-like cells\(^22\). We confirm this finding and show that ESR hPSC have the capacity for germline differentiation. This distinguishes these cells from the naive and primed states in the mouse. Cells in the GCTM-2\(^{\text{high}}\)CD9\(^{\text{high}}\)EPCAM\(^{\text{high}}\) fraction displayed a cell cycle with a very limited G1 fraction relative to other cells in the population. It has been shown that hPSC pause in G1 when preparing to embark on differentiation\(^{25,26,28}\). It is possible that ESR stem cells do not execute such a differentiation checkpoint and continue in a self-renewing loop, until a pause in the cell cycle is activated, possibly through an RB-dependent mechanism\(^{48}\).

Cells in the GCTM-2\(^{\text{high}}\)CD9\(^{\text{high}}\)EPCAM\(^{\text{high}}\) fraction show higher mitochondrial membrane potential and increased oxidative phosphorylation compared with the general population. Comprehensive metabolite profiling and \(^{13}\)C-glucose labeling studies confirmed increased rates of catabolism of pyruvate in the mitochondrial TCA cycle. These studies also showed that increased TCA cycle flux in the GCTM-2\(^{\text{high}}\)CD9\(^{\text{high}}\)EPCAM\(^{\text{high}}\) cells is used to generate key intermediates including citrate and \(\alpha\)-ketoglutarate which are subsequently exported from the mitochondria and used for synthesis of lipids and amino acids/proteins. Elevated rates of amino acid synthesis in the GCTM-2\(^{\text{high}}\)CD9\(^{\text{high}}\)EPCAM\(^{\text{high}}\) cells

---

**Fig. 8 Global gene expression analysis of hPSC subpopulations by RNA-seq.**

- **a** Volcano plot illustrating differentially expressed genes in GCTM-2\(^{\text{high}}\)CD9\(^{\text{high}}\)EPCAM\(^{\text{high}}\) versus general (unsorted) populations.
- **b** Principal component analysis of single-cell RNA-seq data on GCTM-2\(^{\text{high}}\)CD9\(^{\text{high}}\)EPCAM\(^{\text{high}}\), GCTM-2\(^{\text{mid}}\)CD9\(^{\text{mid}}\), and GCTM-2\(^{\text{low}}\)CD9\(^{\text{low}}\) subpopulations.
- **c** Joint species principal component analysis of single-cell RNA-seq data on GCTM-2\(^{\text{high}}\)CD9\(^{\text{high}}\)EPCAM\(^{\text{high}}\) (HHH), GCTM-2\(^{\text{high}}\)CD9\(^{\text{high}}\) (HH), GCTM-2\(^{\text{mid}}\)CD9\(^{\text{mid}}\) (MID), and GCTM-2\(^{\text{low}}\)CD9\(^{\text{low}}\) (LOW) subpopulations alongside cynomolgus embryo data from ref.\(^{30}\); single embryo cells classified according to Houghton et al.\(^{30}\). Top left: screeplot demonstrating the amount of variability in the data accounted for by each component; top right: graph displaying data distribution along first and second components; bottom left: graph displaying data distribution along the second and third components. Color and shape of point indicate sample phenotype, each point representing a single cell.
...described therein; embryonic stages in which particular gene sets (Supplementary Table 4, each row in the expressed are listed to the left of the plots. The second column shows microarray data from ref. 27 for the same sets of genes in GCTM-2highCD9high (H), GCTM-2midCD9mid (M), GCTM-2lowCD9low (L), and GCTM-2negCD9neg (N) subpopulations of MEL1 cells grown in the presence of medium containing hPSC. Colors of violins correspond to embryonic stages or cell subpopulations identified in labels at left or top of figure (last three columns), respectively.

Fig. 9 String section plot comparison of expression of embryonic stage-specific genes by scRNA-seq in cynomolgus embryo with expression in subpopulations of hPSC. The violin plots in the left hand column show expression from single-cell RNA-seq data in ref. 30, with classification of cells as described therein; embryonic stages in which particular gene sets (Supplementary Table 4, each row in the figure displays violin plots for one gene set) are expressed are listed to the left of the plots. The second column shows microarray data from ref. 27 for the same sets of genes in GCTM-2highCD9high (H), GCTM-2midCD9mid (M), GCTM-2lowCD9low (L), and GCTM-2negCD9neg (N) subpopulations of MEL1 cells grown in the presence of medium containing Knockout Serum Replacer on mouse embryo fibroblast feeder cell layers (MEF/KOSR), or WA09 cells grown in MEF/KOSR or mTeSR defined medium. The third column shows microarray data from ref. 16 for the same sets of genes on GCTM-2highCD9high (H), GCTM-2midCD9mid (M), GCTM-2lowCD9low (L), and GCTM-2negCD9neg (N) subpopulations of ES02 cells grown in serum-containing medium on mouse embryo fibroblast feeder cell layers (MEF/FCS). The last column shows RNA-seq bulk data on the same sets of genes from the current study for GCTM-2highCD9highEPCAMhigh (HHH) and unsorted (GEN) cells. Colors of violins correspond to embryonic stages or cell subpopulations identified in labels at left or top of figure (last three columns), respectively.

was also supported by increased flux of glycolytic intermediates into serine and glycine synthesis. The high rate of de novo synthesis of cholesterol in hPSC is unusual and further highlights the importance of TCA cycle in generating precursors (i.e., citrate) for this pathway. Overall, these data indicate that the GCTM-2highCD9highEPCAMhigh cells retain a highly active anabolic metabolism.

DNA methylation levels in mouse naïve ES are generally low, similar to the pattern in the preimplantation epiblast, and rise in vivo as development progresses toward gastrulation. In the mouse or primate post-implantation embryo, DNMT3A and DNMT3B, and TET1 are activated shortly after implantation. Levels of methylation in the GCTM-2highCD9highEPCAMhigh and general population were similar in this study, and considerably lower than levels reported for mouse epiblast stem cells. In human hPSC, co-expression of DNA methyltransferases and TET enzymes could account for the dynamic nature of DNA methylation, exemplified by the dramatic response of these cells to the presence of ascorbic acid, an activator of TET enzymes, in the cell culture medium.

In contrast to DNA methylation levels, chromatin accessibility varied strikingly in the cell subpopulations that we studied. Chromatin regions that showed high accessibility in the GCTM-2highCD9highEPCAMhigh mapped to sites of previously identified pluripotency transcription factor binding in hPSC, and to sites of DNase hypersensitivity that were found to be unique to stem cells. This pattern changed markedly in the GCTM-2midCD9mid population. These findings suggest that targets of pluripotency transcription factors in the GCTM-2highCD9highEPCAMhigh cell fraction might be involved in the regulation of self-renewal. We found no evidence for enrichment of binding sites for TFAP2C in open chromatin in our GCTM-2highCD9highEPCAMhigh cells,
thus distinguishing this population from the naive state of hPSC described earlier.

The properties of the ESR subpopulation of hPSC discussed above are consistent with those of the mouse early post-implantation epiblast. The ability of archetypal hPSC to undergo differentiation into amnion51-53 or the germline22,54 is also consistent with a phenotype closer to an earlier post-implantation state rather than to the EpiSC. We have confirmed the capacity of archetypal hPSC for germline differentiation, a capacity which is found in the majority of the population of archetypal hPSC grown in defined conditions; further study with refinements to this assay will be required to confirm a higher capacity for germline differentiation in GCTM-2 hiPSC-CD9+CD24+CD45+CAM5.2hi cells. We previously used embryoid body assays6 to show that the ESR subpopulation is pluripotent, as did another study using the teratoma assay19. These findings are not surprising, since flow cytometry analyses of replated populations of the ESR subpopulation reported here (Fig. 1a vs. 1c) and in a more extensive previous work19 indicate that this subpopulation can eventually regenerate the entire hierarchy of archetypal hPSC. Here quantitative assessment of directed differentiation into three germ layer lineages in adherent cultures confirmed that both the ESR and the remaining population are pluripotent. Together, our current data and the previous work are consistent with a developmental equivalence of the ESR with early post-implantation epiblast, and not naive or primed pluripotent states.

Previous studies have indicated that gene expression in archetypal hPSC aligns with the preimplantation but not preimplantation epiblast44,45. Our studies on the gene expression in the subpopulation of hPSC with a high capacity for self-renewal are consistent with these findings. The post-implantation epiblast persists longer in the primate than in the mouse, and it is difficult to resolve changes in epiblast gene expression during the post-implantation period until gastrulation, when pluripotency genes turn off and the activation of lineage-specific programs occurs. Our results here and elsewhere demonstrate that the cell cycle and metabolic profile, and the lack of lineage priming in the archetypal hPSC subpopulation showing a high capacity for self-renewal, are all consistent with early post-implantation epiblast, but not an mouse EpiSC-like state. Whether a subpopulation within mouse EpiSC cultures with high self-renewal capacity exists is unknown.

Recently Nakashima et al.55 also described a subpopulation of hPSC with high self-renewal capacity. These cells, which reside at the periphery of hPSC colonies, were isolated on the basis of cell surface expression of NCAD, and display priming toward the primitive endoderm lineage. We previously identified a subset of cells at the periphery of hPSC colonies which co-isolated with the self-renewing subpopulation and similarly co-expressed pluripotency and primitive endoderm genes (GATA4/GATA6/HNF4A/BMP2/FN1)16. We found that these cells were more abundant in cultures grown in medium supplemented with serum replacer on a monolayer cell culture rather to defined conditions; they were quite rare in the HSR subpopulation isolated from cultures grown in defined medium16. We did not observe cells expressing markers of primitive endoderm in the current study. It is possible that these lineage primed cells are descendants of the subpopulation we describe here. Determination of the relationship between the NCAD+ cells and the HSR fraction we have identified will require further analysis, but it is clear that the majority of HSR cells in hPSC cultures grown under defined conditions resemble the early post-implantation epiblast, and not primitive endoderm. Cornacchia et al.56 recently reported that hPSC cultured in E8 media display features that are intermediate between naive and primed hPSC. We have used E8 and mTeSR interchangeably in these studies and have not noted differences between the two with respect to the parameters we have studied. Several features of the E8 cultures described by Cornacchia et al.56 including increased capacity for self-renewal, pluripotency associated marker and gene expression, and the metabolome, are similar to the properties of the ESR population we describe. A simple interpretation of both sets of data and consistent with prior observations16,43 is that culture in defined media shifts the overall population toward the ESR state.

The self-renewing subpopulations of archetypal hPSC that we have identified here and elsewhere represent a minority of the culture with distinct properties. The majority of the archetypal hPSC population shows features closer to mouse EpiSC, as noted by others. Even under the defined culture conditions used in this study that enhance self-renewal, transcriptional evidence of neural lineage priming is evident in the general population. The development of methods to propagate pure populations of self-renewing archetypal hPSC populations in a state similar to the formative stage in mouse might enhance the efficiency of cloning and differentiation protocols, and reduce the variability in differentiation efficiency often observed between hPSC lines from different genetic backgrounds. Further elucidation of the molecular regulatory pathways that maintain cells in this state will enhance our understanding of human development and guide efforts to model embryology in a dish.

**Methods**

**hPSC culture, differentiation, and marker expression.** Experimental procedures for culture and differentiation of human embryonic stem cells, indirect immunofluorescence microscopy and flow cytometry followed minor modifications to established protocols.

**hPSC culture.** Human embryonic stem cell stocks (WA09 and FUCGI-G1) were maintained as described previously57. The FUCGI cells were a gift from Prof Jonathan S. Draper58. Routine maintenance of hPSC was carried out in serum-supplemented medium with fibroblast feeder cell support and subculture was performed using Dispase or collagenase to harvest fragments of colonies dissected manually.

For experiments, cells were transferred to defined feeder-free conditions (mTeSR conditions) using either mTeSR1 medium or Essential 8 medium on Matrigel or recombinant vitronectin and subculture performed with Dispase or Tryple Express according to the manufacturers’ protocols.

For dissociation to single cells, cultures were treated with 10 μM of InSolution®Blebbistatin (Merck Millipore cat. no. 203389) for 1 h prior to dissociation to single cells using Trypel™ (Life Tech, cat. no. 12605). Cultures were incubated in media supplement with 10 μM of InSolution®Blebbistatin overnight after which Blebbistatin was removed.

Routine tests confirmed the absence of mycoplasma contamination and a diploid karyotype (20/20 cells on G-banding) in the cell lines used.

**Fluorescence activated cell sorting for colony formation assays.** Cells grown for 24 h in the presence of Y-27632 Rho kinase inhibitor were dissociated using Accutase, harvested, and examined under the microscope to ensure that most of the cells had not completely dissociated into single cells. Immunolabeling was carried out by incubation in a primary antibody cocktail containing TG30 anti CD9 antibody (mouse IgG2a) and MCTM-2 (mouse IgM) (neat supernatants, this laboratory) for 20 min at 4 °C, followed by incubation in a secondary antibody cocktail containing goat anti-mouse IgG Alexa Fluor 488 antibody and goat anti-mouse IgM Alexa Fluor 647 antibody (Jackson). Cells were then fixed in 4% Formaldehyde (from Thermofisher), diluted in 2% BSA, both from Thermolab. Incubated in DMEM-F12 flow cytometry buffer, again at 4 °C for 20 min. The cells were resuspended the cells in 1X DAPI in cold mTeSR1 before filtering the suspension through a 35-μm cell strainer to remove any larger aggregates and debris. Commercial antibodies were used at the dilutions recommended by the manufacturer.

Diffusion sorting was performed on a BD FACSAria III, using the 100-μm nozzle at 17 psi. A custom cytometer configuration was created that used the 17 psi pressure as opposed to the typical 20 psi pressure. This was performed to minimize aggregates from dissociating into single cells as they were deposited into the 96-well plates. To isolate the GCTM-2 hiPSC-CD9hi and GCTM-2 hiPSC-CD9hi fractions of aggregates and single cells, the gating strategy in Fig. 1a was used (forward and side scatter were employed to gate out debris and to identify single cells and aggregates; propidium iodide or AAD-7 was used to identify dead cells). The various fractions were deposited into 96-well tissue culture plates coated with Matrigel via the automated cell deposition unit on the Aria III. Subsequent to sorting, tissue culture plates were centrifuged at 200 × g for 2 min at room temperature. The sheath solution and media containing mixture was discarded and replaced with 150 μL of mTeSR1 medium.
Single cells from each of two FACs populations (GCTM-2highCD9high and GCTM-2lowCD9low) and GCTM-2lowEPCAM were plated at a density between 100 and 400 cells per well of a 96-well Matrigel-coated plate. Aggregates from each of two FACs populations were placed between 50 and 2000 aggregates per well of a 96-well Matrigel-coated plate (Corning, Costar 3603). Cultures were maintained in mTeSR1 with medium changed once after 48 h post-plating. After 72 h, wells were fixed and stained for colony counting as described. The culture media was removed from cells and each well rinsed once with PBS prior to fixing with 100% ethanol for 10 min at room temperature. Ethanol was removed and wells allowed to air dry for 30 min prior to staining with haematoxylin for 10 min. Wells were then washed four times with Milli-Q water. In all, 0.08% aqueous ammonia solution was added into each well and incubated at 70 °C for 1 h for decolorization. Two anti-mouse IgG Alexa Fluor 488 and two anti-rabbit IgG Alexa Fluor 594 were harvested and re-plated as aggregates into Glasgow Minimal Essential Medium supplemented with LIF (10 ng/mL) and B27 (200 ng/mL) and KITLG (100 ng/mL), and EGf (50 ng/mL), or without these supplements in low attachment 96-well plates. Samples were taken at Day 0, 4, 6, and 8, and analyzed by flow cytometry for expression of cell surface markers CD9, GCTM-2, EPCAM, and TG30. For differentiation assays, we used the StemCell Technologies StemDiff Trilineage differentiation kit (A21131) according to manufacturer instructions. Then cultures were fixed and re-plated in P30A for 30 min, incubated in 0.3% Triton X-100 in PBS, and blocked with 1% IgG-free BSA, incubated in antibody GCTM-2 overnight at 4 °C. Cells were washed three times with Milli-Q water and allowed to dry overnight. For time-lapse video microscopy, cells were imaged during 1–24 h post-plating in an environmental chamber (Clear State Solutions TCH 885-9G) to maintain a humid atmosphere with 5% CO2 in air while phase contrast exposures were recorded every 15 min with a Leica SP8 Confocal microscope.

**Differential potential of hPSC subpopulations.** Differentiation into primordial germ cell-like cells was carried out essentially as described21. Cultures grown in defined, feeder-free conditions (mTeSR1 medium, above) were sorted into GCTM-2highCD9high and GCTM-2lowCD9low fractions and plated onto fibronectin-coated dishes in Glasgow Minimal Essential Medium supplemented with Activin A (50 ng/mL), CHIR99021 (3 μM), and Y-27632. Rho kinase inhibitor (1 μM) and SGK inhibitor (1 μM) was added to the medium. BJ-5ta anti-mouse IgG2a (BD Cat. #563180) (triple stain). Primary antibodies against GCTM-2 and TG30 were used to detect TG30 in experiments that used FUCCI cell lines, mitochondrial dye MitoTracker Red CMXROS (5 μM, 100 μm nozzle) and low-pressure conditions. Cells were harvested with PBS prior to fixation with 100% ethanol for 10 min at room temperature. Ethanol was removed and cells allowed to air dry for 30 min prior to staining with haematoxylin for 10 min. Cells were then washed four times with Milli-Q water. In all, 0.08% aqueous ammonia solution was added into each well and incubated at 70 °C for 1 h for decolorization. Two anti-mouse IgG Alexa Fluor 488 and two anti-rabbit IgG Alexa Fluor 594 were harvested and re-plated as aggregates into Glasgow Minimal Essential Medium supplemented with LIF (10 ng/mL), BMP4 (200 ng/mL), KITLG (100 ng/mL), and EGf (50 ng/mL), or without these supplements in low attachment 96-well plates. Samples were taken at Day 0, 4, 6, and 8, and analyzed by flow cytometry for expression of cell surface markers CD9, GCTM-2, EPCAM, and TG30. For differentiation assays, we used the StemCell Technologies StemDiff Trilineage differentiation kit (A21131) according to manufacturer instructions. Then cultures were fixed and re-plated in P30A for 30 min, incubated in 0.3% Triton X-100 in PBS, and blocked with 1% IgG-free BSA, incubated in antibody GCTM-2 overnight at 4 °C. Cells were washed three times with Milli-Q water and allowed to dry overnight. For time-lapse video microscopy, cells were imaged during 1–24 h post-plating in an environmental chamber (Clear State Solutions TCH 885-9G) to maintain a humid atmosphere with 5% CO2 in air while phase contrast exposures were recorded every 15 min with a Leica SP8 Confocal microscope.

**Immunofluorescence microscopy.** For staining of cells in colony formation assays, hPSC that were cultured in 96-well matrigel-coated plates were washed twice with PBS prior to fixation with 2% paraformaldehyde (PFA) for 30 min at room temperature. Cells were permeabilized with 0.3% Triton X-100 in PBS and blocked with 1% IgG-free BSA, incubated in antibody GCTM-2 overnight at 4 °C, followed by incubation in secondary antibodies goat anti-rabbit IgG2a Secondary Antibody, PE/Cy7 (Biolegend 407017, RMG2a-62) was used to detect TG30. Stained single cells were subsequently labeled with Click-it® EDU Alexa Fluor® 488 Flow Cytometry Assay Kit (Thermo Fisher, Cat. #D10425) according to manufacturer’s method. All FACs plots were created using FCS express (De Novo Software) and the coefficient of variation was also calculated using FCS express or using the Flojo software package.

**Metabolomics studies.** The GCTM-2highCD9highEPCAMhigh population or unsorted cells were plated in 6-well format and allowed to expand for 24 h. Spent media was aspirated, and cells were washed once in Milli-Q water to remove extraneous media, then sufficient liquid nitrogen was added to cover the base of the culture surface and enable metabolic arrest. Next, cells were incubated with 600 μL (per 10 cm² surface area) of ice cold 9:1 MeOH/CHCl₃ containing 0.83 μM ¹³C₆-sorbitol and 8.3 μM ¹³C⁵,¹⁵N-valine for 10 min on ice. The cell lysate was scraped and then transferred to a clean tube, incubated on ice for another 5 min, and centrifuged at 16,100 × g for 5 min at 4 °C following which the supernatant was collected for mass spectrometric analysis. For stable isotope labeling studies, cells were prepared as described above, but the culture media was supplemented with U-¹³C₆-glucose for 24 h prior to cell harvest.

**LC-MS intracellular metabolite profiling analysis.** Metabolite analysis was performed using LC-MS, using hydrophilic interaction (HILIC) LC and high resolution Orbitrap QTOF mass spectrometry. Standards (10 μL) were injected onto an Agilent 1209 LC fitted with a ZIC-hILIC column (5 μm, 2.1×150 mm; Merck), and 20 mM ammonium carbonate (A) and acetonitrile (B) as the mobile phases. A 14 min gradient starting from 90% B to 40% B over 12 min, held for 2 min followed by washing at 5% B for 3 min and re-equilibration at 90% B, was used. Mass spectrometry utilized an Agilent 6545 QTOF with heated electrospray source operating in negative ionization mode and scan range m/z 50–1700. Conditioning was performed before each batch using 2–3 blanks and 5 mixtures of authentic standards (234 metabolites), which were analyzed in data-dependent MS/MS mode to facilitate downstream metabolite identification where possible. PBQC samples were analyzed periodically throughout the analysis.

**GC-MS intracellular metabolite profiling analysis.** Five hundred microliters of the cell extract was evaporated to complete dryness under vacuum (Christ RVC 2-33). Polar metabolites were derivatized online using a Gerstel MPS XL auto-sampler robot (Gerstel, Germany). Samples were first methoxamidated by the addition of 20 μL methanamine (30 mg/mL in pyridine, 2 h, 37 °C, 750 rpm), followed by trimethylsilylation with 20 μL BSTFA + 1% TMCS (1 h, 37 °C, 750 rpm). Metabolite profiles were acquired on an Agilent 7890A Gas Chromatograph coupled to a 5975C Mass Selective Detector, where 1 μL of derivatised sample was injected into a split/spitless inlet set at 250 °C. Chromatographic separation was achieved using an Agilent VF-5 ms capillary column (30 μm x 0.25 mm x 0.25 μm film thickness) held for 2 min, then ramped to 25 °C in 3 min and held for 5 min. Helium was used as the carrier gas at a flow rate of 1 mL/min. Compounds were fragmented by electron impact (EI) ionization and detected across a m/z range of 50–600 amu, with a scan speed of 9.2 scans/s. Chromatograms were processed using PyMS to align metabolite peaks and quantify ions. Sub-micrometer size particles subsequently generate a data matrix. Metabolites were annotated using the in-house Metabolomics Australia (MA, 25C) metabolite library and NIST11 database.
GC-MS intracellular stable isotope incorporation analysis. Stable isotope labeled samples were prepared for GC-MS and analyzed as described above. 13C-labeled metabolic precursors were fragmented to 0.01% of GC-MS was carried out using an Agilent 7890 GC system, VF-5 capillary column with 10 m inert guard (J&W Scientific, 30 m, 250 μm inner diameter, 0.25 μm film thickness) and an Agilent 5975 MSD (Agilent Technologies, Santa Clara, USA) in electron ionization (EI) mode. We used mass isomoter peak shift analysis to monitor 13C-glucose derived labelling in key metabolites of the glycolytic and TCA cycle. Elution times and the fragmentation can be found on the NIST database.

Seahorse flux analysis. The oxygen consumption rate and extracellular acidification rate were determined using an extracellular flux analyzer (Seahorse XF96 Analyzer, Agilent). GCTM uptake was determined per the manufacturer’s instructions. DNA was treated with RNase A then purified through the DNA Clean and Concentrator column (Zymo). RRBS libraries were made from 100 ng of purified DNA using the Ovation RRBS Methyl-Seq System (NuGEN), according to the manufacturers recommendations, which includes the Quagen Epitect kit for bisulfite conversion. Sequencing was performed on a NextSeq 500 with a 75 bp single-end sequencing protocol. Reduced representation bisulphite sequencing was performed using the FAST-ATAC protocol with 50,000 sorted cells for each of three biological replicates of both high and low samples to form a universal set of peaks (peakome), and regions high percentages of reads mapping to mitochondria (Supplementary Fig. 4a).

Single-cell RNA sequencing. Single cells were flow sorted into a chilled 384-well PCR plate containing 1.2 μl of primer/lysis mix (20 μm indexed polydT-1 primer, 1X DEPC water, a dilution of EBCR, 45°C). 100 ng of each dNTPs, 1.2 units SUPERaseIN RNase Inhibitor (Thermo Fisher, AM2692), DEPC water (Thermo Fisher, AM9920) were added to a BD FACSaria III flow cytometer (BD Biosciences, San Jose, CA, USA) and the protocol described above. Sorted plates were sealed, centrifuged for 1 min at 3000 rpm and immediately frozen upside down at −80 °C until further processing using an adapted CellSeq2 protocol17.

Sequencing was performed on a NextSeq 500 with a 75 bp single-end sequencing protocol60. The fragmented RNA was transcribed into cDNA using 5′-tagged random hexamer primers (GCCTTTGCACCCGGAATTCCANNNNNNN) introducing a double-stranded adapter as also described in ref.72 in reverse transcription reaction mix 1 (Ambion, AM13344) was added and incubated at 37 °C for 30 min and then chilled at 4 °C to add the left over primers, 6 μl ExoSAP-IT for PCR Product Clean-Up (Affymetrix, 78200) was added and the sample was incubated at 37 °C for 15 min and then chilled at 4 °C. The sample was then purified by adding 5.5 μl of 10X fragmentation buffer (RNA fragmentation reagents, AM8740) to the sample and incubation in pre-heated thermal cycler at 94 °C for 2.5 min followed by immediately chill on ice and addition of 2.75 μl of Fragmentation Stop buffer (RNA fragmentation reagents, AM8740). The fragmented amplified RNA was purified using 1.8X RNA Clean XP beads (Beckman Coulter, A63987) according to manufactures instruction and eluted in 6 μl DEPC water of which 5 μl (no beads) were transferred to a fresh tube for library preparation.

RNA-seq data analysis. The fragment RNA was transcribed into cDNA using 5′-tagged random hexamer primers (GGCCTTTGCACCCGGAATTCCANNNNNNN) introducing a double-stranded adapter as also described in ref.72 in reverse transcription reaction mix 1 (Ambion, AM13344) was added and incubated at 37 °C for 30 min and then chilled at 4 °C to add the left over primers, 6 μl ExoSAP-IT for PCR Product Clean-Up (Affymetrix, 78200) was added and the sample was incubated at 37 °C for 15 min and then chilled at 4 °C. The sample was then purified by adding 5.5 μl of 10X fragmentation buffer (RNA fragmentation reagents, AM8740) to the sample and incubation in pre-heated thermal cycler at 94 °C for 2.5 min followed by immediately chill on ice and addition of 2.75 μl of Fragmentation Stop buffer (RNA fragmentation reagents, AM8740). The fragmented amplified RNA was purified using 1.8X RNA Clean XP beads (Beckman Coulter, A63987) according to manufactures instruction and eluted in 6 μl DEPC water of which 5 μl (no beads) were transferred to a fresh tube for library preparation.

The fragment RNA was transcribed into cDNA using 5′-tagged random hexamer primers (GGCCTTTGCACCCGGAATTCCANNNNNNN) introducing a double-stranded adapter as also described in ref.72 in reverse transcription reaction mix 1 (Ambion, AM13344) was added and incubated at 37 °C for 30 min and then chilled at 4 °C to add the left over primers, 6 μl ExoSAP-IT for PCR Product Clean-Up (Affymetrix, 78200) was added and the sample was incubated at 37 °C for 15 min and then chilled at 4 °C. The sample was then purified by adding 5.5 μl of 10X fragmentation buffer (RNA fragmentation reagents, AM8740) to the sample and incubation in pre-heated thermal cycler at 94 °C for 2.5 min followed by immediately chill on ice and addition of 2.75 μl of Fragmentation Stop buffer (RNA fragmentation reagents, AM8740). The fragmented amplified RNA was purified using 1.8X RNA Clean XP beads (Beckman Coulter, A63987) according to manufactures instruction and eluted in 6 μl DEPC water of which 5 μl (no beads) were transferred to a fresh tube for library preparation.

Reduced representational bisulphite sequencing. FACS sorted cells were snap frozen in buffer IBT plus (Quagen) before DNA extraction using the AllPrep DNA/RNA Mini Kit (Qiagen) as per the manufacturer’s instructions. DNA was treated with RNAse A then purified through the DNA Clean and Concentrator column (Zymo). RRBS libraries were made from 100 ng of purified DNA using the Ovation RRBS Methyl-Seq System (NuGEN), according to the manufacturers recommendations, which includes the Quagen Epitect kit for bisulfite conversion. Sequencing was performed on a NextSeq 500 with a 75bp single-end sequencing protocol. Sequencing was performed on a NextSeq 500 with a 75 bp single-end sequencing protocol. Reduced representation bisulphite sequencing was performed using the FAST-ATAC protocol with 50,000 sorted cells for each of three biological replicates of both high and low samples to form a universal set of peaks (peakome), and regions high percentages of reads mapping to mitochondria (Supplementary Fig. 4a).
compare between human and macaque expression levels, we defined a set of high confidence orthologous metaexons between the two species as in ref. 20. Briefly, we used BLAST to compare every annotated human exon in ENSEMBL release 86 (737,982 unique exons across 63,305 genes) against the human (hg38) and Macaca fascicularis (macf5as) genomes, retaining those that matched the macaque genome with at least 92% sequence identity and mapped back to their annotated location in humans. We then excluded all exons that had a second match with >90% sequence similarity in either genome, to control for interspecies differences in mappability. Overlapping exons from the same gene (associated with different isoforms) were collapsed into a single “metaexon”.

We discarded overlapping exons associated with more than one ENSEMBL gene ID, or exons associated with any gene annotated to two or more chromosomes in either species, and exons where the difference in intron size between the two species is ≥10,000 bp, suggestive of poor genome assembly or annotation. After applying these quality control criteria, we ultimately retained 198,172 unique metaexons in humans and macaques across 34,142 annotated ENSEMBL human genes. The final table, as well as code and additional documentation for metaexon identification is available at http://www.bitbucket.org/ec_reh_neb/orthoxon.

We processed all 2526 files from ref. 44 (from 390 cells with sickle [https://github.com/najoshi/sickle]) to remove bad reads and trim low-quality bases from the 3’ end. We then mapped all reads to macf5as using Rsubread 1.2.6 and R 3.2.2, allowing up to 2 mismatches and 2 indels per 50 bp read, which is proportional to our setting of 5 mismatches or indels for a 100 bp read. Mapped reads were assigned to the orthologous metaexon list using featureCounts at both the single metaexon and whole-gene level, and summed within individuals in R 3.2.2.

**Human scRNA-seq data processing.** Sequence analysis was performed on the Illumina Next-Seq 500 platform.

Quality control of raw data was assessed using FASTQC and visualized using MultiQC. The scPipe package v1.0 for R was used to count genes based on UMI profile. Gene expression was normalized using scaterv1.6.1 and calculated on the merged data using prcomp function in the stats base package for R statistical software version 3.3.2. Downstream analysis of the principal components was performed using the miMixons package version 6.3.1 for R version 3.3.2. Panther.db was used to perform Fisher’s exact test for over-representation of ontological terms in gene sets of interest.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability.** RNA-seq, scRNA-seq, and ATAC-seq data are available at Gene Expression Omnibus under accession GSE147338 (RNA-seq, GSE119324, scRNA-seq GSE119323, ATAC-seq GSE147338).

Received: 13 June 2019; Accepted: 16 April 2020; Published online: 15 May 2020

**References.**

1. Ramos-Beas, P., Nichols, I. & Alberio, R. States and origins of mammalian embryonic pluripotency in vivo and in a dish. Curr. Top. Dev. Biol. **128**, 151–179 (2018).

2. Ying, Q. L. et al. The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).

3. Brons, I. G. et al. Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191–195 (2007).

4. Tesar, P. J. et al. New cell lines from mouse epiblast share defining features with mammalian embryonic stem cells. *Nature* **448**, 196–199 (2007).

5. Kojima, Y. et al. The transcriptional and functional properties of mouse epiblast stem cells resemble the anterior primitive streak. *Cell Stem Cell* **14**, 107–120 (2014).

6. Smith, A. Formative pluripotency: the executive phase in a developmental continuum. *Development* **144**, 365–373 (2017).

7. Bernemann, C. et al. Distinct developmental ground states of epiblast stem cell lines determine different pluripotency features. *Stem Cells* **29**, 1496–1503 (2011).
of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).
38. Consortium, E. P. A. Integrated Encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
39. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics 32, 587–589 (2016).
40. Shefﬁeld, N. C. et al. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. 23, 777–788 (2013).
41. Kalkan, T. et al. Tracking the embryonic stem cell transition from ground state pluripotency. Development 144, 1221–1234 (2017).
42. van de Leemput, J. et al. CORTECON: a temporal transcriptome analysis of in vitro human pluripotent cortical development from human embryonic stem cells. Neuron 83, 51–68 (2014).
43. Kolle, G. et al. Identiﬁcation of human embryonic stem cell surface markers by combined membrane-polysome translation state array analysis and immunotranscriptional proﬁling. Stem Cells 27, 2446–2456 (2009).
44. Nakamura, T. et al. A developmental coordinate of pluripotency among mice, monkeys and humans. Nature 537, 57–62 (2016).
45. Stirparo, G. G. et al. Integrated analysis of single-cell embryo data yields a uniﬁed transcriptomic signature for human the pre-implantation embryo. Development 145, https://doi.org/10.1242/dev.158501 (2018).
46. Barbaric, I. et al. Time-lapse analysis of human embryonic stem cells reveals multiple bottlenecks restricting colony formation and their relief upon culture adaptation. Stem Cell Rep. 3, 142–155 (2014).
47. Barone, V. et al. An effective feedback loop between cell-cell contact duration and morphogen signaling determines cell fate. Dev. Cell 43, 198–211 (2017).
48. Boward, B., Wu, T. & Dalton, S. Concise Review: Control of cell fate through alternative splicing in primates. Cell 185, 1848–1855 (2010).
49. Chung, T. L. et al. Vitamin C promotes widespread yet speciﬁc DNA demethylation of the epigenome in human embryonic stem cells. Nat. Biotechnol. 33, 2446–2456 (2015).
50. Pastor, W. A. et al. TFAP2C regulates transcription in human naive pluripotent stem cells. Dev. Cell 43, 198–211 (2017).
51. Deglincerti, A. et al. Self-organization of the in vitro attached human embryo. Nature 533, 251–254 (2016).
52. Shahbazi, M. N. et al. Self-organization of the human embryo in the absence of fetal membrane and morphogen signaling determines cell fate. Nat. Cell Biol. 18, 700–708 (2016).
53. Shao, Y. et al. Self-organized amniogenesis by human pluripotent stem cells in a biomimetic implantation-like niche. Nat. Mater. 16, 419–425 (2017).
54. Kojima, Y. et al. Evolutionarily distinctive transcriptional and signaling programs drive human germ cell lineage speciﬁcation from pluripotent stem cells. Cell Stem Cell 21, 517–532 (2017).
55. Nakanishi, M. et al. Human pluripotency is initiated and preserved by a unique subset of founder cells. Cell https://doi.org/10.1016/j.cell.2019.03.013 (2019).
56. Cernacchia, D. et al. Lipid deprivation induces a stable, naive-to-prime intermediate state of pluripotency in human PSCs. Cell Stem Cell 25, 120–136 (2019).
57. Reubinoff, B. E., Pera, M. F., Fong, C. Y., Trounson, A. & Bongso, A. Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. Nat. Biotechnol. 18, 399–404 (2000).
58. O’Callaghan, S. et al. PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools. BMC Bioinform. 13, 115 (2012).
59. Kowalski, G. M. et al. Application of dynamic metabolomics to examine in vivo skeletal muscle glucose metabolism in the chronically high-fat fed mouse. Biochem Biophys. Res. Commun. 462, 27–32 (2015).
60. Yin, D. et al. High concordance between Illumina HiSeq2500 and NextSeq500 for reduced representation bisulﬁte sequencing (RRBS). Genom. Data 10, 97–100 (2016).
61. Andrews, S. FastQC (Babraham Bioinformatics, 2010).
62. Krueger, F. Trim Galore v0.6.5 (Babraham Bioinformatics, 2019).
63. Krueger, F. & Andrews, S. Bismark: a ﬂexible aligner and methylation caller for Bisulﬁte-Seq applications. Bioinformatics 27, 1571–1572 (2011).
64. Andrews, S. SeqMonk (Babraham Bioinformatics, 2010).
65. Corces, M. R. et al. Lineage-speciﬁc and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. Genet. 48, 1193–1203 (2016).
66. Bolget, A. M., Leduc, M. & Usadel, B. Trimomatic: a ﬂexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
68. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137 (2008).
69. Quinlan, A. R. & Hall, J. M. BEDTools: a ﬂexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
70. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).
71. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 17, 77 (2016).
72. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-extend. Nucleic Acids Res. 41, e108 (2013).
73. Tian, L. et al. scPipe: a ﬂexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. PLoS Comput Biol. 14, e1006361 (2018).
74. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 17, 75 (2016).
75. Nakamura, T. et al. SC3-seq: a method for highly parallel and quantitative measurement of single-cell gene expression. Nucleic Acids Res. 43, 860 (2015).
76. Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-speciﬁc and lineage-speciﬁc alternative splicing in primates. Genome Res. 20, 180–189 (2010).
77. Kent, W. J. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664 (2002).

Acknowledgements
This study was supported by the Australian Research Council Special Research Initiative in StemCell Sciences (SR1101002), and Bioplatforms Australia, through funding from the Australian Government National Collaborative Research Infrastructure Strategy. M.J.M. is a Principal Research Fellow of the National Health and Medical Research Council. M.E.B. was supported by a Bellberry-Viertel Senior Medical Research Fellowship.

Author contributions
Contributions of authors to drafting of speciﬁc sections of the paper are identiﬁed by name of section titles. K.X.L.: design of experiments, collection of data, and writing of the paper (Methods); E.A.M.: analysis of data and writing of the paper (Methods); J.K.: design of experiments, collection of data, and writing of the paper (Methods); D.P.D. and M.M.: design of experiments, analysis of data, and writing of the paper (Methods, Results); J.K.: collection and analysis of data; D.T., T.B., C.S., and T.C.M.: collection of data; A.K.: design of experiments, analysis and collection of data, and writing of the paper (Methods); M.E.B.: design of experiments, analysis of data; M.E.R.: analysis of data; S.H.N.: design of experiments, analysis of results, writing of the paper (Methods); D.Z.: design of experiments and collection of data; O.K. and S.S.: analysis of results; I.G.R., M.M.: design of experiments, analysis of results, writing of the paper; C.A.W.: design of experiments, analysis of results, and writing of the paper (Methods, Results); M.F.P.: conception of study, design of experiments, analysis of results, and writing of the paper.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-16214-8.

Correspondence and requests for materials should be addressed to M.F.P.

Peer review information Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional afﬁliations.
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Lau, KX; Mason, EA; Kie, J; De Souza, DP; Kloehn, J; Tull, D; McConville, MJ; Keniry, A; Beck, T; Blewitt, ME; Ritchie, ME; Naik, SH; Zalcenstein, D; Korn, O; Su, S; Romero, IG; Spruce, C; Baker, CL; McGarr, TC; Wells, CA; Pera, MF

Title:
Unique properties of a subset of human pluripotent stem cells with high capacity for self-renewal

Date:
2020-05-15

Citation:
Lau, K. X., Mason, E. A., Kie, J., De Souza, D. P., Kloehn, J., Tull, D., McConville, M. J., Keniry, A., Beck, T., Blewitt, M. E., Ritchie, M. E., Naik, S. H., Zalcenstein, D., Korn, O., Su, S., Romero, I. G., Spruce, C., Baker, C. L., McGarr, T. C., ... Pera, M. F. (2020). Unique properties of a subset of human pluripotent stem cells with high capacity for self-renewal. NATURE COMMUNICATIONS, 11 (1), https://doi.org/10.1038/s41467-020-16214-8.

Persistent Link:
http://hdl.handle.net/11343/240710

File Description:
Published version

License:
CC BY