Unraveling the Role of MIXL1 Activation in Endoderm Differentiation of Isogenic Human Induced Pluripotent Stem Cells

Pierre Osteil,[1*ϕ] Sarah Withey,[3] Nicole Santucci,[1]
Nader Aryamanesh,[4] Ignatius Pang,[4] Nazmus Salehin,[2] Jane Sun,[1]
Annie Qin,[1] Jiayi Su,[1] Hilary Knowles,[1] Simon Cai,[4] George Craft,[5] Mark Graham,[5]
Xiucheng Bella Li,[1] Ernst Wolvetang,[3] and Patrick P. L. Tam, [1,2*]

[1]Embryology Research Unit, Children's Medical Research Institute, University of Sydney, Australia,
[2]School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Australia,
[3]Australian Institute for Bioengineering and Nanotechnology, University of Queensland, Australia
[4]Bioinformatics Group, Children's Medical Research Institute, University of Sydney, Australia,
[5], Biomedical Proteomics, Children's Medical Research Institute, University of Sydney, Australia
[*] To whom correspondence should be addressed; E-mail: pierre.osteil@uca.fr, ptam@cmri.org.au
[ϕ] Present address: Université Clermont Auvergne, CNRS, INSERM, GReD Institute, Faculté de Médecine, F-63000 Clermont-Ferrand, France.

Human induced pluripotent stem cells (hiPSC) possess the ability to differentiate into a multitude of tissue types but display heterogeneity in the propensity of differentiation into specific tissue lineages. An examination of isogenic hiPSC lines revealed variations in the endoderm propensity under directed differentiation conditions. Characterization of the transcriptome and proteome of the hiPSC lines showed that *MIXL1* activity at the early differentiation stage correlated with the efficacy of generating definitive endoderm and further differentiation into endoderm derivatives. Enforced expression of *MIXL1* in the endoderm-incompetent hiPSCs enhanced the propensity of endoderm differentiation, suggesting that modulation of key drivers of lineage differentiation can re-wire hiPSC to the desired lineage propensity for stem cell products.

Introduction

Human induced pluripotent stem cells (hiPSCs) are noted for their ability to differentiate into a multitude of cell and tissue types [1–6]. Many protocols have been developed to direct the differentiation of hiPSCs to desirable types of cells or tissues, including the endoderm precursor, the definitive endoderm[7–10] and endoderm derivatives such as intestinal cells [11–13] pancreatic cells[14] and hepatocytes[15–18]. A recent study of a bank of hiPSC lines derived from 125 individuals revealed that hiPSC lines respond differently when directed to differentiate to definitive endoderm[19], suggesting that there is innate heterogeneity in the propensity for endoderm differentiation among hiPSC lines. In this study the lineage propensity has been linked to specific quantitative trait loci (QTL). Both genetic determinants [20–22], and somatic or epigenetic memory related to the cell/tissue of origin[23–26], have been shown to underpin the variable lineage specification and differentiation propensity of hiPSCs. The impact of epigenetic memory on establishment of functional tissue nevertheless remains unresolved, even when examining isogenic cell lines from the same cellular resource and reprogrammed under the same

conditions[27]. In the present study, we investigated the propensity of endoderm differentiation of eleven pluripotent lines of four sets of isogenic hiPSCs by tracking the differentiation from pluripotent cells to definitive endoderm (DE), hepatocytes and intestinal organoids (hIO). We showed that in isogenic hiPSCs, early activation and a high level of *MIXL1* expression were associated with enhanced propensity of endoderm differentiation. In the mouse embryo, *Mixl1* is expressed in the primitive streak and the nascent mesoderm during gastrulation and expression persists in the primitive streak of the early-somite-stage embryo[28,29]. Loss of Mixl1 function is associated with deficiency of DE and under-expansion of the nascent mesoderm[30]. In the mouse embryonic stem cells, loss of *Mixl1* function leads to inefficient differentiation of lateral mesoderm tissue and hematopoietic lineages[31], whereas constitutive *Mixl1* activity promotes the differentiation of Foxa2+/ECad+ DE cells[32]. In mouse epiblast stem cells, activation of *Mixl1* at the early phase of differentiation correlates with improved endoderm differentiation[28]. Analysis of the molecular attributes of DE differentiation revealed that the activity of *MIXL1* at the early phase of hiPSC differentiation promotes the differentiation of SOX17+ DE cells when confined to defined size micropattern[33]. We further showed that enhanced expression of *MIXL1* in hiPSCs augmented endoderm propensity, advancing the understanding of how lineage propensity can be re-wired to generate fit-for-purpose pluripotent stem cells for translational application.

Results

Early onset of gastrulation is necessary for definitive endoderm formation

Eleven hiPSC lines from four isogenic groups (two males and two females)[34,35] were subjected to DE differentiation by following the manufacturer instruction (STEMDiff definitive endoderm protocol) (Figure 1A) and assessed for expression of FOXA2 and SOX17 on Day 4 of differentiation (Figure 1B and C, Supplementary Figure S1A and B). Among the hiPSCs, C32 had the lowest expression of both endodermal transcription factors, despite comparable cell morphology to other cell lines (Figure 1C), suggesting that this cell line is not amenable to definitive endoderm differentiation. To track the developmental trajectory, cells were collected in triplicate every day from Day 0 (pluripotency) to Day 4 (DE), and the expression of 96 genes involved in regulating pluripotency to gastrulation was profiled. Line C32 showed the least progression across the 4 days of differentiation in the PCA plot (Figure 1D). By taking PC1 as an endoderm differentiation efficiency proxy[19], to infer an endoderm specification pseudotime, the average of triplicate PC eigenvalue along the PC1 axis was calculated to rank the hiPSC lines (Figure 1E). The results show that C32 ranked last, indicating that low DE differentiation efficiency could account for the differentiation failure previously suggested by the absence of FOXA2 and SOX17 double positive cells.

To unveil the earliest manifestation of such discrepancies, the same PCA as Figure 1D was plotted for each day, to discern differences among cell lines that may not be visible on the full timeline of differentiation (Figure 1F). The C32 line displayed disparity on the main axis from the cohort at Day 1 (Figure 1G). The low differentiation efficiency does not appear to be linked to a slower down-regulation of pluripotency factors as there are no significant differences among the 11 cell lines for *POU5F1, NANOG, SOX2, PRDM14, ZFP42* and *FGF5* mRNA expression

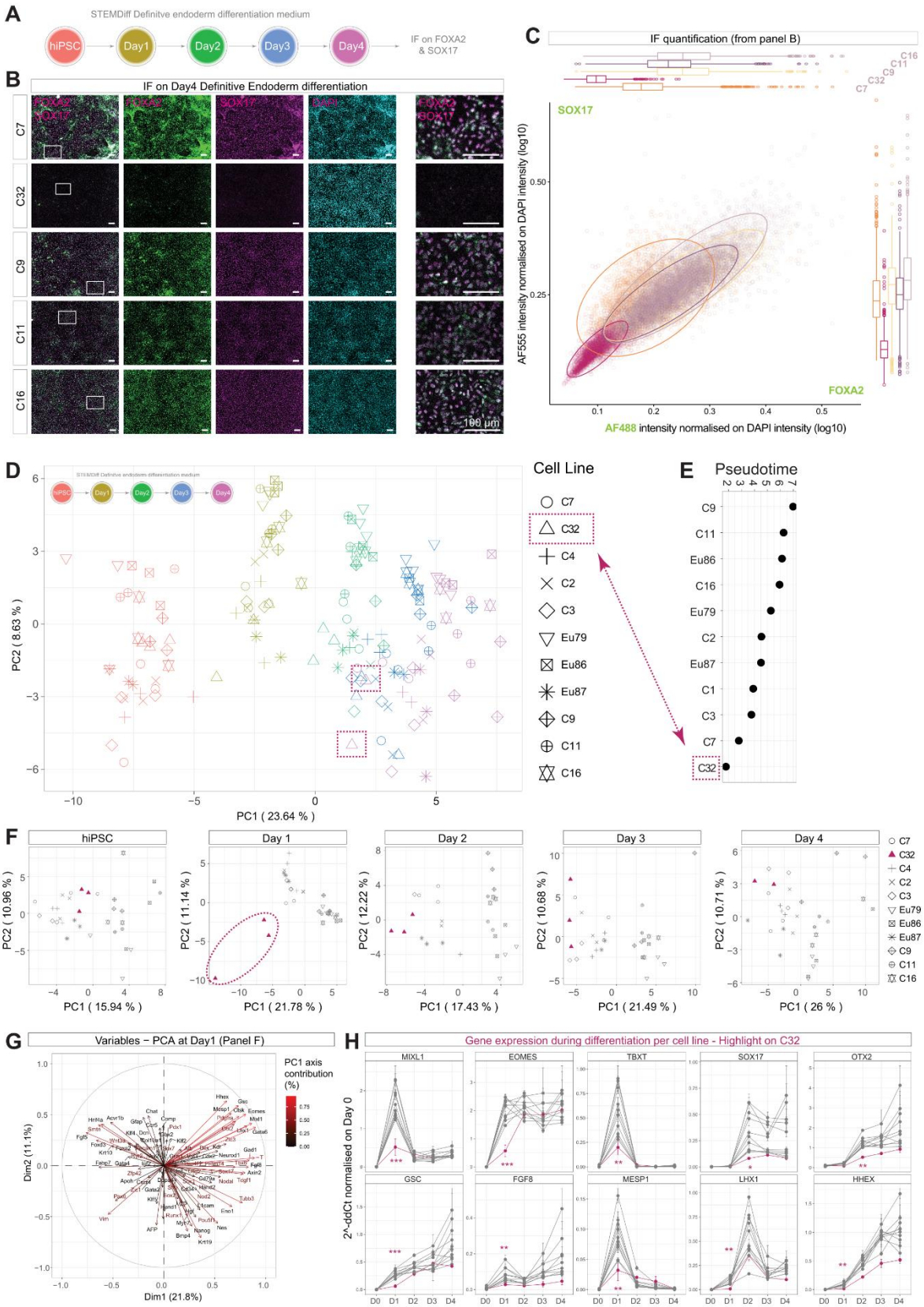


Figure 1 – Definitive endoderm differentiation heterogeneity among isogenic lines: A) Differentiation protocol used to generate definitive endoderm (DE) cells; **B)** Immunostaining images of DE cells on FOXA2 (green) and SOX17 (magenta), scale bar = 100µm; **C)** Signal intensity measurement of immunostaining of panel B, n= 3. **D)** PCA obtained from microfluidic RT-qPCR data on DE differentiation time courses. Each day is represented by the same color as Figure 1A (inserted on top left). Purple squares highlight C32 at Day 4 of DE differentiation; **E)** PC1 axis projection of hiPSC at Day 4 of DE differentiation representing efficiency of differentiation as pseudotime. C32 is highlighted in purple. **F)** PCA obtained from microfluidic RT-qPCR data on DE differentiation, identical dataset to figure 1D, but plotted for each day. C32 is highlighted in purple; **G)** Genes' contribution to PC1 and PC2 axis of the Day 1 PCA of Figure 2F. The position of the arrows correlates with the position of the samples on the PCA at Day 1. Colour scale is between 0 and 1. A score of 1 indicates maximum contribution of a particular gene to PC1. **H)** Genes' expression time course during DE differentiation. C32 is highlighted in purple. p.value: * < 0.05, ** < 0.01, *** < 0.001.

down-regulated at Day1 in C32 compared to the cohort (Figure 1H). C32 also has the lowest levels of expression of *SOX17, HHEX, OTX2* and *LHX1* at Day 4. *KRT19, AXIN2, GATA6* and *GATA4* are expressed at a lower level in the C32 line, although this did not reach significance (Supplementary Figure S2Aii). Surprisingly, *FOXA2* and *NODAL* are not dysregulated in the C32 line. In addition, mesodermal genes (*BMP4, MYH7, KLF5, KDR, PDGRFa* and *CD34*) (Supplementary Figure S2Aiii) as well as ectodermal genes (*KRT10, SOX1, NES, FOXD3, PAX6* and *DCX*) (Supplementary Figure S2Aiv) showed no significant difference between C32 and other lines.

In conclusion, the inability of C32 to activate the molecular program of germ layer differentiation to a threshold level may be predictive of the low endoderm propensity of this hiPSC line.

Low endodermal propensity line fails to progress toward functional tissue

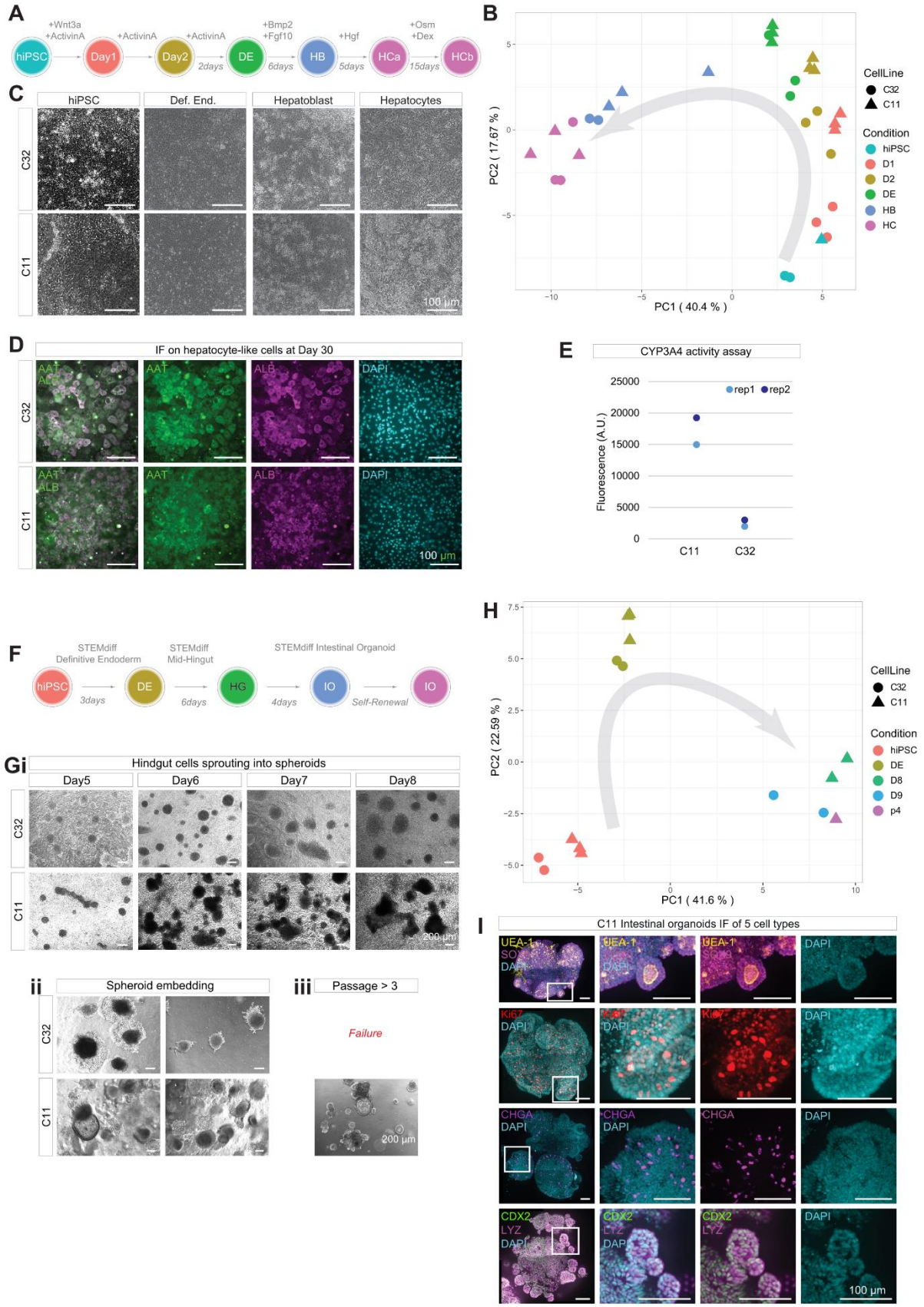
Human iPSC lineage propensity was further assessed by the outcome of differentiation into two endoderm derivatives, hepatocytes (HCm) and intestinal organoids (hIOs) (Figure 2A). C32 was compared to the higher DE propensity cell line C11[34,35]. Both cell lines were able to differentiate into hepatocytes (Figure 2C). Microfluidic RT-qPCR analysis of genes specific to hepatocyte development did not reveal any major differences in the transcriptome between these two cell lines during hepatocyte differentiation (Figure 2B). The phenotype of AAT- and ALB-expressing hepatocytes was also similar (Figure 2C, D and Supplementary Figure S1C). However, C32-derived hepatocytes showed lower Cytochrome P450 3A4 activity across replicates (Figure 2E) indicating that C32 hepatocytes might have a less efficient metabolism compared to C11.

In parallel, hIOs were generated from these two cell lines (Figure 2F). At the mid/hindgut budding spheroid stage, the C32 line generated fewer spheroids than the C11 line (Figure 2Gi). The few spheroids that were successfully generated were embedded into Matrigel but did not grow as well as C11 spheroids (Figure 2Gii). Consequently, hIO development was arrested early after passage 3 (Figure 2Giii). The hIO differentiation of C32 line was repeated with changes in seeding densities and batches of cells, but the outcome remained unsuccessful (data not shown). Microfluidic RT-qPCR analysis of genes specific to intestinal organoid differentiation did not show any discernible signature that could be indicative of inefficient differentiation prior to the hIO development failure of the C32 cell line (Figure 2H). In contrast, C11 derived hIOs harbor the representative cell types of the intestinal epithelium: enterocytes (CDX2+), intestinal stem cells (SOX9), enteroendocrine cells (CHGA), goblet cells (UEA-1) and Paneth cells (LYZ) (Figure 2I).

Together, these results indicate that C32 behaved differently from the C11 line, and displayed a lower propensity for endoderm differentiation. The C32 cell line therefore was labelled as a hiPSC line that is refractory to endoderm differentiation in our subsequent analyses.

Hippo signaling is up regulated in the refractory cell line

Bulk RNA-seq was performed at Day 1 of DE differentiation to discover genes involved in differential regulation of endodermal propensity at the early phase of germ layer differentiation



123

Figure 2 – Low endodermal propensity fails to produce functional tissue: A) Differentiation protocol used to generate hepatocytes from C32 and C11; **B)** PCA obtained from microfluidic RT-qPCR data on hepatocyte differentiation; **C)** Brightfield pictures of hepatocytes differentiation; **D)** Immunostaining images on AAT (green) and ALB (magenta) markers of hepatocytes differentiation; **E)** Results of fluorescent analysis of CYP3A4 activity; **F)** Differentiation protocol used to generate intestinal organoids; **G)** Brightfield pictures of intestinal organoid differentiation at the Hindgut stage (i), spheroid generation stage (ii) and maintenance stage (iii); **H)** PCA obtained from microfluidic RT-qPCR data on intestinal organoid differentiation; **I)** Immunostaining of intestinal organoids of C11 cell lines on different cell types of the gut epithelium: Goblet cells (UEA-1), Intestinal Stem cell (SOX9), enteroendocrine cells (CHGA), epithelium (CDX2) and Paneth cells (LYZ) as well as proliferating cells (Ki67). Nuclei are revealed by DAPI.

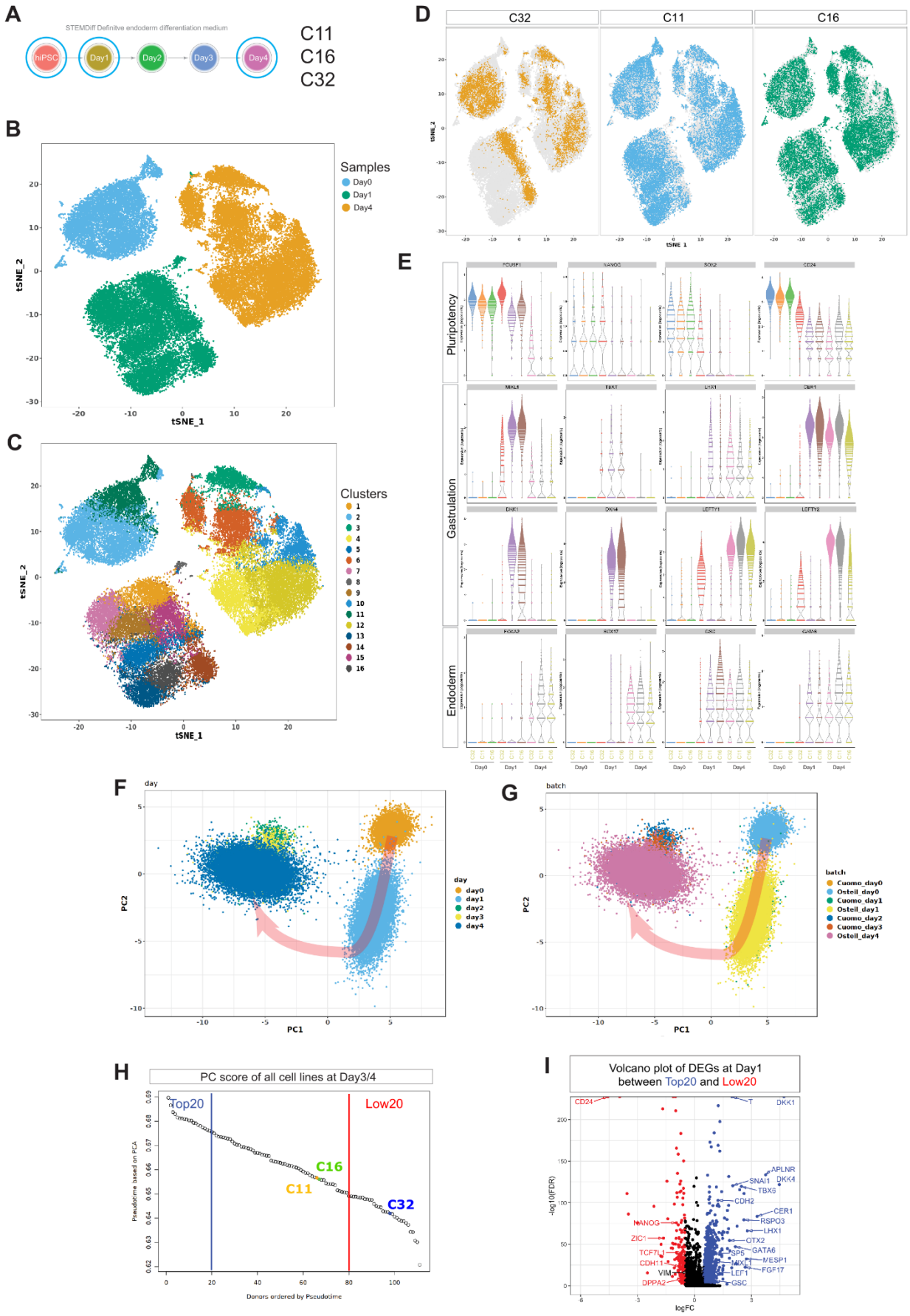
(gastrulation). Five cell lines were analyzed including two females (C32 and C7) and 3 males (C9, C11, C16). Transcriptomic differences showed that C32 differs significantly from the 4 other cell lines (Supplementary Figure S2B and D), displaying enrichment of transcriptomic signature of mesoderm (circulatory system and heart: *RUNX1, VEGFA*) and ectoderm (neural: *SOX2, POU3F2*) (Supplementary Figure S2C) derivatives. Most of the genes associated with gastrulation (e.g. *MIXL1, EOMES, MESP1, APLN, DKK1, GATA6*, etc...) (Supplementary Figure S2C) were down-regulated in C32. To eliminate the gender bias in the above analysis, C7, an isogenic clone of C32, was analyzed in parallel. The results of this comparison were similar to the global analysis. The differentially expressed genes (DEGs) analysis revealed that C7 expressed genes associated with gastrulation (*MIXL1, EOMES, TBXT, SNAI1, CER1*) at a higher level, while C32 differentially expressed genes associated with cell adhesion (*VIM, EZR, FLNA, FLNC, COL1A1, FN1, ITGA2/3/6*) and Hippo signaling (*CCN1, CCN2, AMOT, AJUBA, CDH11*) (Supplementary Figure S2E). The transcriptomic analysis thus highlights a possible ectoderm and mesoderm bias of the C32 cell line during germ layer differentiation that is in keeping with its lower endodermal propensity. It also indicated a possible role for altered Hippo signaling and cell adhesion in negatively modulating endoderm differentiation.

Refractory cell line harbors a unique molecular signature during differentiation of primitive-streak like cells

We further investigated three cell lines (i.e., C32, C11 and C16) by single-cell RNAseq (scRNA-seq) at 3 time points during differentiation: Day 0 (pluripotency), 1 (peri-gastrulation) and 4 (definitive endoderm) (Figure 3A). After filtering, 41336 cells were retained for analysis. The tSNE plot (Figure 3B) showed that cells were segregated by time into three major clusters corresponding to each day. Each cluster were further divided into smaller clusters based on their transcriptomic differences (Figure 3C). While C32 cells displayed a unique transcriptomic profile at Day 1, the individual sub-clusters of C32 cells were found within the cell clusters of each of the other cell lines at Day 4 (Figure 3D), albeit present at different abundance, suggesting that the molecular signatures were shared by cells of the three hiPSC lines.

To better appreciate the discrepancy of the cells at Day 1, the identity of single cells was annotated based on a human post-implantation embryo dataset[36], as Epiblast, Primitive Streak, Nascent Mesoderm, Emergent Mesoderm, Advanced mesoderm and Endoderm (Supplementary Figure S3A and B). Cells of the C32 line retained an Epiblast signature at Day 1 (55% of cells) and did not display a primitive-streak like cell state like the C11 and C16 cell lines (Supplementary Figure S3C and D).

Finally, individual gene comparison of cell states across the 4-day differentiation showed C32 cells maintained a robust expression of pluripotency-related factors, *(SOX2, NANOG, POU5F1)*. Interestingly, Nodal targets and antagonists, *LEFTY1* and *LEFTY2*, were up-regulated at Day 1 (Figure 3E). Genes associated with germ layer differentiation, *MIXL1, LHX1, DKK1, DKK4*, and endoderm related genes (*GSC, GATA6*) were down-regulated. While *SOX17* and *FOXA2* were not down-regulated, the proportion of endoderm cells in the C32 line was reduced relative to other cell lines (Figure 3E and Supplementary Figure S3D), in agreement with the IHC data. Collectively, scRNA-seq, bulk RNA-seq and microfluidic RT-qPCR all pointed at a failure or a delay of the C32 line in differentiating to primitive-streak like cells.



167

Figure 3 – scRNA-seq demonstrates that C32 is part of low propensity cell lines: A) Definitive Endoderm Differentiation protocol used for C32, C16 and C11. Blue circles highlight the samples selected for scRNA-seq. **B)** tSNE plot obtained from scRNA-seq data. **C)** SC3 clustering obtained from scRNA-seq data. **D)** tSNE plot colored for each cell line. **E)** Gene expression for each sample. **F and G)** PCA showing integration of our scRNA-seq data with data from Cuomo et al, **F)** represents cells grouped for each day; **G)** grouped for each sample. The pink arrow indicates time progression during differentiation; **H)** PC1 axis projection of hiPSC at Day 3/4 of DE differentiation representing efficiency of differentiation as pseudotime. **I)** Differentially Expressed Genes (DEGs) between the Top 20 cell lines (better) (blue) versus the Low 20 (less efficient) cell line (red) represented on a Volcano plot.

Single-cell transcriptomic analysis as a tool to rank lineage propensity

To test if the information learned from the scRNA-seq analysis from our cell line cohort may be extrapolated to hiPSC lines in general, the scRNA dataset was combined with that of 125 hiPSC lines previously profiled for endoderm differentiation *in vitro*[19] (Figure 3F and G). Each cell line was ranked for endoderm propensity based on their PC1 eigenvalue at Day 4 (Figure 3H) as described before. The data show that C32 ranks among the lowest 20% propensity cell lines, while C16 and C11 ranked higher.

Using this classification, the Top 20 cell lines (Top20) and the lowest 20 cell lines (Low20), were analyzed for DEGs. This analysis revealed that genes associated with primitive streak formation were up-regulated in Top20 cell lines (*TBXT, DKK1, SNAI1, MIXL1, LHX1, etc...*) while Low20 cell lines retained a pluripotent signature (*NANOG, DPPA2, ZIC1*) (Figure 3I). The data infer that the Low20 lines are less competent for germ layer differentiation and that the C32 characteristics apply to a number of other hiPSCs of low endoderm differentiation propensity.

We conclude that irrespective of the sequencing technology used (SMART-seq2 vs 10X) and the differentiation protocol (homemade versus commercial), single cell data sets can be used to rank hiPSC cell lines in terms of endodermal differentiation propensity based on PC1 Eigen value score. This in turn provides a useful tool that will enable the identification of cell lines with superior differentiation propensity prior to the use of these cells to productively generate the endodermal derivatives.

Refractory hiPSC showed unique proteomic signature

Although C32 differed from other cell lines at Day 1, the single cell transcriptomic data showed little difference between C32 and the two other cell lines (C11 and C16) at Day 0, while cells were pluripotent and undifferentiated. To explore other signatures that may reflect inter-line differences in endoderm differentiation propensity, the proteome of eleven cell lines was analyzed at Day 0 (Supplementary Figure S4). This discovery proteome screen revealed that C32 clustered separately from other groups of hiPSC lines and its isogenic counterpart, C7 clustered with a different group of female cell lines (Supplementary Figure S4A). Compared to C7, C32 displayed a higher level of expression of pluripotency-related factors, PODXL and FZD7, and downregulation of FN1 (associated with pharyngeal endoderm)[37] (Supplementary Figure S4B and C). Of note, PODXL expression is maintained at Day1 of DE differentiation (Figure S2E) and is also associated with kidney differentiation, a mesoderm derivative[6], correlating with our previous observation that C32 might be poised for mesoderm differentiation. The proteomic signature supports our previous findings that the retention of pluripotency in the C32 line contributes to the failure of differentiation towards the endoderm lineage.

MIXL1 is required for promoting endoderm differentiation

Since our data of the refractory line C32 showed that dysregulation of *MIXL1* expression early in lineage differentiation may underpin the low endoderm propensity, we next tested the requirement of *MIXL1* for endoderm differentiation. To this end, frameshift mutations of *MIXL1* were engineered in C32 and C16 lines by CRISPR editing to generate *MIXL1* loss of function cell lines: C32-MKO line and C16-MKO line respectively. We validated the introduction of the

MIXL1 gene edits by sequencing the targeted regions and confirmed this did not impact pluripotency of these lines (Supplementary Figure S5).

Separately, the C32 line was engineered by constitutive expression of a dead Cas9, with no nuclease activity, linked to VP64, a potent transcriptional activator (dCas9-PVP64)[38,39]. This line was genetically modified further to express two sgRNAs that targeted the dCas9-VP64 to the promoter of *MIXL1*, in a doxycycline controllable fashion (C32-Dox)[40]. These two guides (sgRNA4 and 7 – Figure S6F) displayed the strongest activation of the gene compared to 7 other guides when tested in HEK cells (Figure S6G).

We next quantified *MIXL1* expression in the KO lines and the C32-Dox line with and without induction at Day 1 of DE differentiation. No *MIXL1* expression can be detected in C32-MKO and C16-MKO (Figure 4A and B) relative to C32-Dox cells treated with DMSO (C32-Dox0) (Figure 4A and B). Maximal induction of *MIXL1* expression was achieved, at a saturating concentration of 2μg/mL Dox. Beyond which, at 16μg/mL Dox, MIXL1 expression was reduced possibly due to the toxicity that impacts negatively on mitochondrial gene activity. The levels of induced expression of MIXL1 with 1μg/mL of Doxycycline, quantified by immunofluorescence, were within the physiological range (comparable to C16 at Day 1). Further studies of C32-Dox cells were therefore performed at 1μg/mL Doxycycline (C32-Dox1) (Figure 4A and B).

To assess how outcomes of endoderm differentiation were modulated by different levels of *MIXL1* expression, FOXA2 and SOX17 expression was quantified after 4 days of differentiation. Increasing MIXL1 expression in C32-DOX resulted in higher expression of both markers (Figure 4C and D), further reinforcing a role of *MIXL1* in promoting endoderm differentiation. Surprisingly, however, C32-MKO cells displayed FOXA2 and SOX17 expression levels similar to C16-MKO and C32-Dox0. This finding suggests that MIXL1 dysregulation may not be the sole cause of inefficient endoderm specification and differentiation. However, these results indicate that modulation of MIXL1 expression can have an effect on DE formation.

MIXL1 plays a role in chromatin organization

To elucidate the impact of MIXL1 on chromatin accessibility, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) was performed on C16 and C16-MKO cell lines at Day 1 of DE differentiation, when MIXL1 expression is maximal. Comparing the accessible regions (or reads pile-up called as peaks) showed that MIXL1 deletion led to multiple changes in chromatin accessibility (Supplementary Figure S6A), and in particular, less closing regions (Supplementary Figure S6B), than opening (Supplementary Figure S6C). This observation suggests that MIXL1 may be responsible for opening and closing regions during germ layer differentiation. To understand the role of these regions, motif discovery was performed on differentially accessible chromatin regions. This analysis revealed that peaks with less accessibility in MKO lines contain the motifs TAATNNNATTA (PROP1, PHOXA2), which is the dual homeobox motif recognized by MIXL1 (Supplementary Figure S6D). In the absence of *MIXL1*, more accessible peaks are associated with TEAD1 and FOXH1 motifs (Supplementary Figure S6E). FOXH1 is known as a cofactor of GSC which negatively regulates MIXL1 in the mouse[41]. TEAD1 is the transcription factor bound by YAP/TAZ when Hippo signaling is inhibited[42]. This is consistent with the observation that the C32 cell line expresses low *MIXL1*

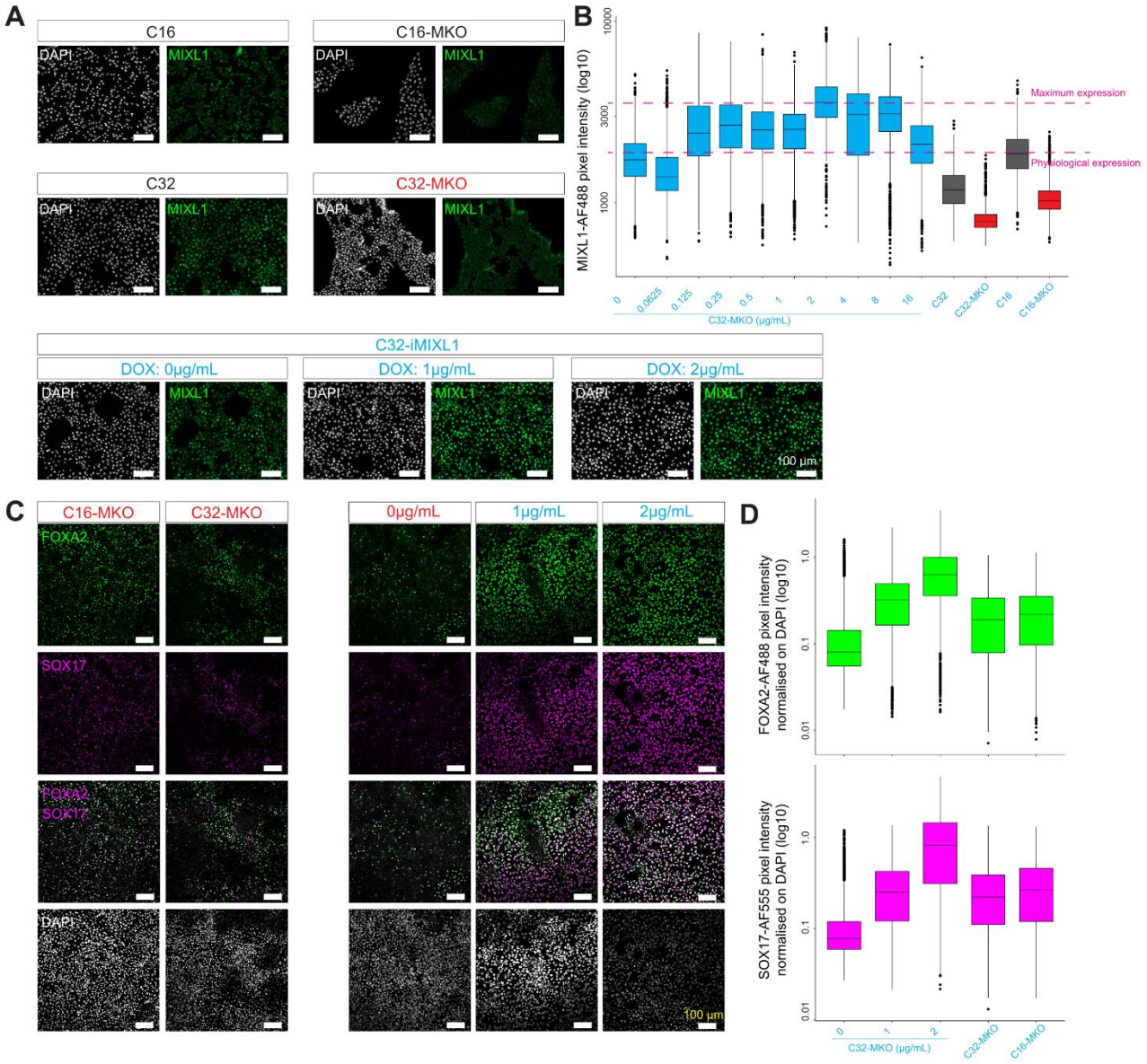


Figure 4 – MIXL1 functional genomic study reveals its role in endodermal differentiation:

A) Immunostaining images on MIXL1 (green) for C16 and C32 either WT or MIXL1-KO (MKO) and C32 with inducible MIXL1 expression under three different concentrations of doxycycline (0, 1 and 2µg/mL) called C32-iMIXL1. Nuclei are revealed by DAPI (white). **B)** MIXL1 signal intensity normalized on DAPI signal for concentration of doxycycline spanning from 0 to 16µg/mL, WT and MKO conditions. 2 levels are highlighted a physiological level, corresponding to the level of C16 and 1ug/mL of doxycycline and an overexpression level corresponding to 2µg/ml. (n= 3). **C)** Immunostaining images of FOXA2 (green) and SOX17 (magenta) on Day 4 of DE differentiation in MKO and iMIXL1 cell lines. **D)** Quantification of FOXA2 and SOX17 immunostaining.

expression and exhibits up-regulation of YAP/TAZ targets (*CCN1*, *CCN2*) (Supplementary Figure S2E).

Collectively the data reveal that MIXL1 may be a pioneer transcription factor involved in modulating chromatin accessibility of its targets and possibly influences signaling pathways such as Hippo and WNT, as inferred from the TCF3 motifs found in open regions in the C16-MKO lines.

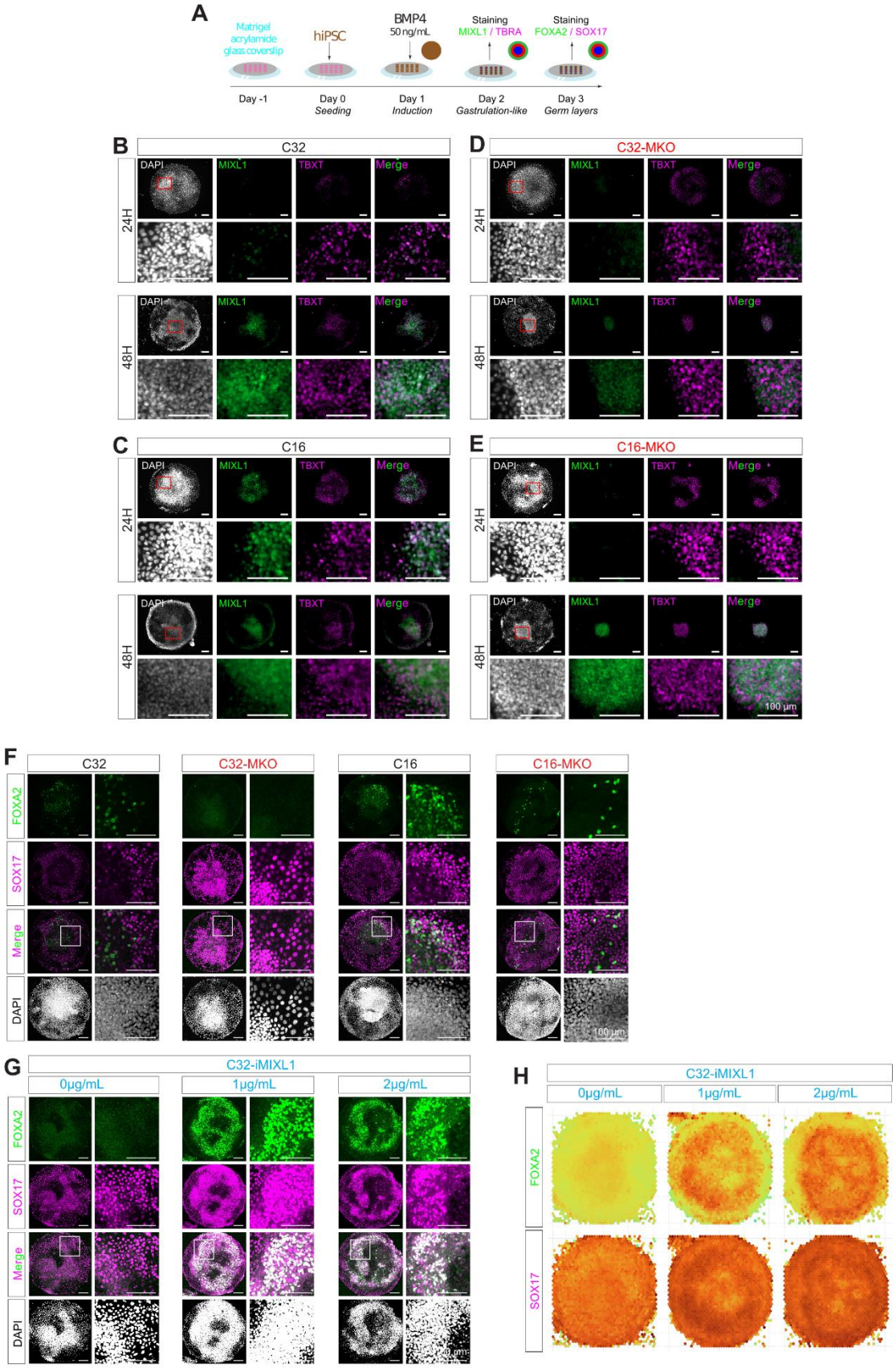
Physiological levels of MIXL1 activity can rescue the endoderm propensity of refractory cells in germ layer differentiation.

To assess the function of *MIXL1*, we used the 2D stem cell micropattern model to elucidate its functional attribute in germ layer differentiation (Figure 5A, Supplementary Figure S7 and S8). Five cell lines, C32, C32-MKO, C32-Dox, C16 and C16-MKO, were used to generate micropatterned cultures that recapitulate germ layer formation in response to BMP4[33,43,44].

The emergence of primitive streak-like cells was assessed via the immunostaining of TBXT and MIXL1 proteins 24h after BMP4 supplementation. In the C32 cell line, similar expression kinetics were observed, but was temporally delayed from 48h onwards. At 24h MIXL1 signal was almost undetectable (Figure 5B). In the C16 line TBXT and MIXL1 were expressed at peak level at 24h followed by decreased expression at 48h. Some cells were co-expressing both proteins (Figure 5C). At 48h, DE differentiation was assessed by FOXA2 and SOX17 expression. FOXA2+/SOX17+ cells were sparse in the C16 line (Figure 5F). This low propensity for DE cells in micropattern has already been documented[33,43,45]. However, no double positive cells were detected in C32 line (Figure 5F). Instead, separate domains of FOXA2+ cells (inner ring) and SOX17+ cells (outer ring) were identified. These results from the C16 and C32 lines confirm that a change in culture format could not rescue the phenotype of refractory endoderm differentiation of the C32 line.

MKO lines were analyzed similarly and MIXL1 could not be detected in both KO lines, and TBXT did not appear to be affected in its spatiotemporal pattern (Figure 5D and E). Regarding DE differentiation, the main effect of the loss of MIXL1 activity was the reduced population of FOXA2+ cells (Figure 5F).

To elucidate the effect of induced *MIXL1* activity on DE differentiation of the incompetent iPSC line, the C32Dox lines were cultured in micropatterns under BMP4 condition with induction by DOX at 2 doxycycline concentration: 1µg/mL, corresponding to physiological condition and 2µg/mL, a condition where *MIXL1* is overexpressed. Endoderm differentiation at Day 2 of differentiation of induced C32Dox line (C32-iMIXL1), compared with C32 line (parental, low propensity line) and C16 line (high propensity line), was assessed by the presence of SOX17+ and FOXA2+ cells in the micropatterns. C32-induced cells displayed increased number and density of SOX17+ and FOXA2+ cells (Figure 5G, H) compared to C32 line and C16 line (Figure 5F) and the control condition without doxycycline. Reconstitution of physiological levels of MIXL1 activity therefore restored the endoderm propensity of iPSCs that are inherently incompetent for endoderm differentiation. In addition, increasing the doxycycline concentration increased the level of double positive cells in the micropatterned cultures indicating a correlation between *MIXL1* level and DE cells formation.



291

Figure 5 – MIXL1 induction rescues DE phenotype in a pseudo-embryo model. A) Protocol to establish stem cell-micropattern model of germ layer differentiation; **B to E)** Immunostaining images of MIXL1 and TBXT at 24h and 48h after BMP4 induction on C32 **(B)**, C16 **(C)**, C32-MKO **(D)** and C16-MKO **(E)**. **F** and **G)** Immunostaining images of FOXA2 (green) and SOX17 (magenta) at 48h after BMP4 induction on C32, C32-MKO, C16 and C16-MKO **(G)** and C32-iMIXL1 under three different concentrations of doxycycline (0, 1 and 2µg/mL). **H)** Average signaling for FOXA2 and SOX17 measured on micropatterns (n= 10).

Discussion

In this study, isogenic human induced pluripotent stem cell (hiPSC) lines were analyzed and compared for their propensity in generating definitive endoderm that is capable of progressing to functional endoderm derivatives, here tested by the formation of intestinal organoids and hepatocytes (See Figure 2). One cell line of the cohort (C32) was found to be inefficient in differentiation towards the endoderm lineage. The low propensity for endoderm derivatives is accompanied by a bias toward the mesoderm lineage. Our results suggest that during germ layer differentiation, C32 activates the genetic program of vascularization and heart formation more efficiently than the rest of the cohort (See Supplementary Figure S2C). In addition, the C32 hiPSC line has been used in a previous study to produce kidney organoid[6] supporting the potential mesoderm bias of the C32 cell line.

We sought to determine molecular markers of endoderm propensity of hiPSC lines, at the pluripotency and early exit stages. Transcriptomic and proteomic analyses did not reveal evident bias in lineage propensity of the refractory line C32, except for a higher level of expression of some pluripotency-related factors at the proteomic level (Supplementary Figure S4). The higher pluripotency level of C32 may underpin the poor performance in endoderm differentiation of this cell line. To gain a holistic view of the differentiation potential of isogenic hiPSC lines, 4 groups of isogenic hiPSC lines (2 males and 2 females) were subjected to deep transcriptomic analysis. This again highlighted the failure of C32 in activating the gastrulation genetic program properly. In other words, C32 differentiation was inefficient and delayed in initiating germ layer formation compared to its isogenic clone C7 and the rest of the cohort.

The single cell transcriptomic data was analyzed in conjunction with the transcriptomic data of a previous large scale study[19] surveying 125 hiPSC lines during definitive endoderm differentiation. Despite the use of different protocols (homemade medium versus commercial kit for the present study) and sequencing chemistry (SMART-seq2 versus 10X Chromium for the present study), the data were remarkably comparable. We found that the cell lines with the lowest differentiation score at the end of the DE differentiation, have a significantly low expression of genes involved in gastrulation including *MIXL1*. These results further validated our microfluidic RT-qPCR findings and enabled the identification of a gene panel and novel tool for ranking hiPSC lines for the propensity of endoderm differentiation and ability to generate mature endoderm tissues.

The transcriptomic survey of the cohort revealed that the gene *MIXL1* is expressed at a low level in endoderm-incompetent cell lines, and this is corroborated via cross comparison of the aforementioned study[19]. Our endoderm differentiation data of stem cell-derived micropattern further points to a causal relationship between the expression of *MIXL1* and efficiency of DE differentiation. We proposed that *MIXL1* is a useful biomarker for screening human pluripotent stem cells for competency of endoderm differentiation by quantification of *MIXL1* expression in cells at 24h of directed differentiation.

The ATAC-seq data pointed to a possible role for MIXL1 in regulating chromatin accessibility. Of note, the target regions opened in the MIXL1-KO cell line are strongly enriched for TEAD1 binding sites. TEAD1 is a transcription factor involved in Hippo signaling[42]. This data, when combined with our observation that Hippo target genes (e.g., *CCN1, CCN2*) are more strongly activated in the refractory C32 line (Supplementary Figure S2E), suggests an important

relationship between chromatin status and the Hippo pathway that is regulated by *MIXL1*. Interestingly, the closed chromatin region in the MIXL1-KO line mainly contains the dual homeobox binding sites TAATNNNATTA, recognized by MIXL1. This suggests that MIXL1 may be involved in regulating accessibility to its own transcriptional targets.

Our study has provided a comprehensive survey of endoderm differentiation in an isogenic cohort and provides a framework for future study of the molecular mechanisms that underpin endoderm specification during germ layer differentiation of pluripotent stem cells and in embryonic development.

Acknowledgments

We would like to acknowledge the support of Advanced Imaging for all imaging and Vector and Genome Engineering facilities at CMRI for the help in generating the KO lines. The Advanced Microscopy Facility at CMRI for the image generation. Prof. Kristopher Kilian and his team for sharing the micropattern protocols and training. This work was supported by the National Health and Medical Research Council (NHMRC) (Project Grant ID1127976). PPLT was supported by Senior Principal Research Fellowship (NHMRC Grant ID110751)

Author contributions

PO, EW and PPLT conceived of the study. PO, AW, Ni.S, PPLT designed experimentation. PO and Ni.S were the primary scientist and completed most of the cell and molecular biology, live and fixed imaging, and biochemical analysis. AQ, Jy.S, XBL assisted with molecular biology, live and fixed imaging, and biochemical analysis. HK and Ni.S completed the RNA-seq and scRNA-seq and NA performed the associated analysis. SC helped with preliminary analysis. Ni.S completed ATAC-seq and Na.S the associated analysis. Ja.S, IP, GC and MG completed all mass spectrometry and associated analysis. PO, EW and PPLT secured funding. PO completed experimental analysis and interpretation and created the figures with assistance from NA and PPLT. PO, Ni.S and PPLT wrote the manuscript with assistance from all authors.

Conflict of interest

We the authors declare no conflict of interest

Figures Legend

Figure 1 – Definitive endoderm differentiation heterogeneity among isogenic lines:

A) Differentiation protocol used to generate definitive endoderm (DE) cells; **B)** Immunostaining images of DE cells on FOXA2 (green) and SOX17 (magenta), scale bar = 100µm; **C)** Signal intensity measurement of immunostaining of panel B, n= 3. **D)** PCA obtained from microfluidic RT-qPCR data on DE differentiation time courses. Each day is represented by the same color as Figure 1A (inserted on top left). Purple squares highlight C32 at Day 4 of DE differentiation;

E) PC1 axis projection of hiPSC at Day 4 of DE differentiation representing efficiency of differentiation as pseudotime. C32 is highlighted in purple. **F)** PCA obtained from microfluidic RT-qPCR data on DE differentiation, identical dataset to figure 1D, but plotted for each day. C32 is highlighted in purple; **G)** Genes' contribution to PC1 and PC2 axis of the Day 1 PCA of Figure 2F. The position of the arrows correlates with the position of the samples on the PCA at Day 1. Colour scale is between 0 and 1. A score of 1 indicates maximum contribution of a particular gene to PC1. **H)** Genes' expression time course during DE differentiation. C32 is highlighted in purple. p.value: * < 0.05, ** < 0.01, *** < 0.001.

Figure 2 – Low endodermal propensity fails to produce functional tissue:

A) Differentiation protocol used to generate hepatocytes from C32 and C11; **B)** PCA obtained from microfluidic RT-qPCR data on hepatocyte differentiation; **C)** Brightfield pictures of hepatocytes differentiation; **D)** Immunostaining images on AAT (green) and ALB (magenta) markers of hepatocytes differentiation; **E)** Results of fluorescent analysis of CYP3A4 activity; **F)** Differentiation protocol used to generate intestinal organoids; **G)** Brightfield pictures of intestinal organoid differentiation at the Hindgut stage (i), spheroid generation stage (ii) and maintenance stage (iii); **H)** PCA obtained from microfluidic RT-qPCR data on intestinal organoid differentiation; **I)** Immunostaining of intestinal organoids of C11 cell lines on different cell types of the gut epithelium: Goblet cells (UEA-1), Intestinal Stem cell (SOX9), enteroendocrine cells (CHGA), epithelium (CDX2) and Paneth cells (LYZ) as well as proliferating cells (Ki67). Nuclei are revealed by DAPI.

Figure 3 – scRNA-seq demonstrates that C32 is part of low propensity cell lines:

A) Definitive Endoderm Differentiation protocol used for C32, C16 and C11. Blue circles highlight the samples selected for scRNA-seq. **B)** tSNE plot obtained from scRNA-seq data. **C)** SC3 clustering obtained from scRNA-seq data. **D)** tSNE plot colored for each cell line. **E)** Gene expression for each sample. **F and G)** PCA showing integration of our scRNA-seq data with data from Cuomo et al, **F)** represents cells grouped for each day; **G)** grouped for each sample. The pink arrow indicates time progression during differentiation; **H)** PC1 axis projection of hiPSC at Day 3/4 of DE differentiation representing efficiency of differentiation as pseudotime. **I)** Differentially Expressed Genes (DEGs) between the Top 20 cell lines (better) (blue) versus the Low 20 (less efficient) cell line (red) represented on a Volcano plot.

Figure 4 – MIXL1 functional genomic study reveals its role in endodermal differentiation:

A) Immunostaining images on MIXL1 (green) for C16 and C32 either WT or MIXL1-KO (MKO) and C32 with inducible MIXL1 expression under three different concentrations of doxycycline (0, 1 and 2µg/mL) called C32-iMIXL1. Nuclei are revealed by DAPI (white). **B)** MIXL1 signal intensity normalized on DAPI signal for concentration of doxycycline spanning from 0 to 16µg/mL, WT and MKO conditions. 2 levels are highlighted a physiological level, corresponding to the level of C16 and 1ug/mL of doxycycline and an overexpression level corresponding to 2µg/ml. (n= 3). **C)** Immunostaining images of FOXA2 (green) and SOX17 (magenta) on Day 4 of DE differentiation in MKO and iMIXL1 cell lines. **D)** Quantification of FOXA2 and SOX17 immunostaining.

Figure 5 – MIXL1 induction rescues DE phenotype in a pseudo-embryo model.

A) Protocol to establish stem cell-micropattern model of germ layer differentiation; **B to E)** Immunostaining images of MIXL1 and TBXT at 24h and 48h after BMP4 induction on C32 **(B)**, C16 **(C)**, C32-MKO **(D)** and C16-MKO **(E)**. **F** and **G)** Immunostaining images of FOXA2 (green) and SOX17 (magenta) at 48h after BMP4 induction on C32, C32-MKO, C16 and C16-MKO **(G)** and C32-iMIXL1 under three different concentrations of doxycycline (0, 1 and 2µg/mL). **H)** Average signaling for FOXA2 and SOX17 measured on micropatterns (n= 10).

Resource availability

Lead contacts

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contacts, Patrick Tam (p.tam@cmri.org.au) and Pierre Osteil (pierre.osteil@uca.fr).

Material availability

The materials used in this study are listed in the key resources table. Materials generated by our laboratory in this study are available on request, however, there are restrictions to the availability of human iPSC lines due to a Material Transfer Agreement.

Data and Code availability

- All raw sequencing data can be found on Gene Expression Omnibus under the accession number GSE260552, GSE260553, GSE260554 and are available as of the date of publication.
- The raw mass spectrometry datasets generated in this study are available via PRIDE: PXD048788, http://www.ebi.ac.uk/pride/archive/projects/PXD048788
- All microfluidic RT-qPCR data can be found on the GitHub page (https://github.com/PierreOsteil/ScriptsForOsteilEtAl2024) and are available as of the date of publication.
- All original codes are available as of the date of publication and can be found on the following GitHub page: https://github.com/PierreOsteil/ScriptsForOsteilEtAl2024. Bioinformatic source codes and their corresponding DOIs are listed in the key resources table
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.t

Experimental Model and Subject Details

Cell Lines

A cohort of 11 human iPSC lines composed of 2 to 3 isogenic cell lines from 4 patients (2 males and 2 females) was provided by the Australian Institute for Bioengineering and Nanotechnology

(AIBN), University of Queensland. Briefly, hiPSC lines derived from fibroblast cells or foreskin tissue were generated using a non-integrating episomal reprogramming system (oriP/EBNA1-based pCEP4 episomal vectors pE-P4EO2SCK2MEN2L and pEP4EO2SET2K from Addgene) carrying *OCT4, SOX2, KLF4* and *CMYC*. All lines maintained a normal karyotype and were capable of forming teratomas that contained derivatives of the three germ layers[34,35]. For routine maintenance, hiPSCs were cultured in mTeSR1 (STEMCELL Technologies) on six-well plates precoated with hESC-qualified Matrigel (Corning). The culture plates were incubated at 37°C and 5% CO2. The medium was changed daily. The colonies morphology was evaluated under an inverted microscope. Cells were passaged at 70-80% confluency with ReLeSR(STEMCELL Technologies) at a split ratio of 1/5 to 1/30 depending on the cell line, into a new well of a 6-well plate. Experiments were approved by the Sydney Children's Hospitals Network Human Research Ethics Committee under the reference: HREC/17/SCHN/167.

Method details

CRISPR-KO engineering

gRNA design and cloning

Single *S. pyogens* Cas9 gRNA (GCGCCGCGTTTCCAGCGTACCGG) targeting *MIXL1* exon 1 was designed using Geneious software, (http://www.geneious.com[46]) based on the presence of a canonical NGG PAM (underlined in gRNA sequence) at the target site. Potential off-target sites were identified using Geneious software, (http://www.geneious.com[46]). gRNA was cloned in Addgene plasmid 62988 following adopted protocol from Ran et al[47]. Oligos used for cloning were:

Forward: 5' CACCGCGCCGCGTTTCCAGCGTAC
Reverse: 5' AAACGTACGCTGGAAACGCGGCGC.

Nucleofection, clone selection and sequencing

Cells were transfected using a plasmid expressing Cas9 protein and gRNA targeting MIXL1 exon 1 following Amaxa™ 4D Transfection protocol for 20 μl Nucleocuvette® Strip using P3 Primary Cell 4D-Nucleofector® X Kit with program CA-137. After transfection cells were plated into 10 cm dish, coated with hESC-qualified Matrigel (BD Biosciences), prefilled with mTESR medium (Stem Cell Technologies) mixed with 100% CloneR (Stem Cell Technologies). Twenty-four hours post transfection cells were puromycin (Thermo Fisher Scientific) selected with concentration of 1μg/ml for the next 48 hours. Following puromycin selection media was changed every day and the percentage of CloneR (Stem Cell Technologies) in media was reduced during the next days as single cells were dividing and started forming individual colonies. Single colonies were picked and transferred individually in a single well of a 96 well plate where they were grown to be split and frozen for further sequencing analysis. Cells were detached using ReLeSR (Stem Cell Technologies) and clones were frozen as cell aggregates in CryoStor® CS10 (Stem Cell Technologies). Clone selection, screening of the CRISPR/Cas9 clones for editing events and validation of allelic deletions of individual clones was done following protocol from Bauer et al. [48] for genomic deletions in mammalian cells. The PCR was designed to amplify the sequence flanking the gRNA on exon 1 targeting location with the expected

amplicon of 800bp. PCR analysis for the presence of indels were done with primers: Forward: 5'GGAGGGTATAAGTGCGGCC Reverse: 5'CCTCATCTGTGTCTTCTTCCCG

All PCR reactions were done in 50μl volume using Q5 high fidelity polymerase (NEB) following NEB Q5 high fidelity PCR protocol. In short, PCR reaction mix was made by mixing 100ng of genomic DNA sample from each clone with 10μl of 5xQ5 reaction Buffer, 1μl of 10mM dNTPs, 2.5μl of each (forward and reverse) 10 μM primer, 10μl of 5xQ5 High GC Enhancer, 0.5μl of Q5 Polymerase and topped up to 50μl with H_2O. PCR reaction started with initial denaturation with temperature of 98°C for 30s followed by 34 cycles of 10s denaturation at 98°C, annealing at 60°C for 20s and extension at 72°C for 20s ending with final extension at 72°C for 5min. PCR reaction was run on 1.5% agarose gel where expected amplicon of 800bp for each analyze clone was detected. In total, 43 samples were separately amplified by PCR and analyzed by sequencing for the presence of indels at the exon 1 targeted site. Next, sequenced clones were analyzed for genome editing and indel percentages were calculated via TIDE[49] using a control chromatogram for comparison. Decomposition windows, left boundaries, and indel ranges were optimized to have the highest alignment possible. After TIDE analysis 11 clones were selected for validation of biallelic deletion clones for targeted genomic region of exon 1, which was done following standard protocol from Bauer et al.[48]

Differentiation protocols

Definitive endoderm differentiation and characterization

For direct differentiation into Definitive Endoderm, the cells were subject to induction using STEMdiff Definitive Endoderm kit (STEMCELL Technologies) for 4 days, following manufacturer's protocol. Briefly, cells were passaged into single cells with StemProAccutaseCell Dissociation Reagent (Life Technologies) and seeded with mTeSR1 containing Y-27632 dihydrochloride Rock Inhibitor (Tocris, Cat. No. 1245). After 24 hours, the cells are washed with PBS and then cultured for 4 days in STEMdiff Definitive Endoderm Basal medium with Supplements A and B for the first day and then Supplement B only for the subsequent 3 days, with daily medium changes. Samples were harvested daily for RNA and Protein extraction (n=3).

To characterise the definitive endoderm, cells were seeded onto glass coverslips coated with hESC-qualified Matrigel (Corning) before treating with STEMdiff™ Definitive Endoderm kit (STEMCELL) as described above. Cells were fixed in 4% paraformaldehyde in PBS at RT for 20 min. They were washed with PBS twice and then permeabilized with 0.1% Triton X-100 (Merck) in dPBS (Gibco) (PBST) at RT for 5min. The cells were blocked with 3% bovine serum albumin (Merck Aldrich) in PBST at room temperature for 1 hour. They were incubated with primary antibody at 4°C overnight (FOXA2 (Abcam) 1:300, SOX17 (R&D Systems) 1:20). Cells were washed with dPBS three times, then incubated with corresponding secondary antibodies at RT for 1 hour. The cell nuclei were stained with DAPI (1μg/ml) (Thermo Fisher Scientific) in dPBS for 10 min at RT, and then washed three times with dPBS. Cells were mounted with Fluoromount-G (Thermo Fisher Scientific) and imaged on Ziess Axio Imager Z2 widefield microscope.

Human Intestinal Organoids differentiation and characterization

hiPSC-derived intestinal organoids were formed using the STEMdiffIntestinal Organoid Kit (StemCell Technologies), following the manufacturer's protocol. Briefly, cells were passaged as clumps using ReLeSR (StemCell Technologies). Once 80-90% confluency was reached, differentiation was initiated with DE Medium (STEMdiffEndoderm Basal Medium plus STEMdiffDefinitive Endoderm Supplement CJ) for 3 days, with daily medium changes. Subsequent mid-hindgut differentiation was induced with MH Medium (STEMdiffEndoderm Basal Medium plus STEMdiffGastrointestinal Supplement PK and STEMdiffGastrointestinal Supplement UB) for 6 days, with daily medium changes. Free-floating mid-/hindgut spheroids, collected at 24-hour intervals within the MH Medium treatment, were embedded in Matrigel (Corning) in wells of NunclonDelta surface 24-well plate (Thermo Fisher Scientific), overlaid with STEMdiffIntestinal Organoid Growth Medium (STEMdiffIntestinal Organoid Basal Medium plus STEMdiffIntestinal Organoid Supplement (StemCell Technologies) and GlutaMAX (Gibco)), performing medium change every 3 - 4 days, incubating at 37°C with 5% CO_2. After 7 - 10 days of incubation, cultures were passaged. Briefly, all plasticware were pre-wetted with Anti-Adherence Rinsing Solution (StemCell Technologies). Matrigel domes containing organoids were broken manually by pipetting up and down with cold DMEM/F-12 (Gibco), seeding 40-80 organoid fragments per 50μl Matrigel dome.

Organoids were removed from Matrigel similar to that described for splitting organoids above and were fixed in 4% paraformaldehyde in PBS at RT for 30 min. They were washed with PBS twice and then permeabilised with 0.1% Triton X-100 (Merck) in dPBS (Gibco) (PBST) at RT for 1 hour. The organoids were blocked with CAS-Block (Thermo Fisher Scientific) for 90min and then incubated with primary antibody (Sox9 (Merck) 1:500, Ki67 (Abcam) 1:250, CHGA (Novus Biologicals) 1:200, CDX2 (Biogenex) 1:250, Lysozyme (Dako) 1:200) overnight at 4°C. Organoids were washed with PBST four times, then incubated with corresponding secondary antibodies and stain (DAPI (1 μg/ml) (Thermo Fisher Scientific) for nuclei in all samples and UEA-1 (Vector Laboratories) 1:200, for select organoids) in CAS-Block at RT for 3 hours. The organoids were then washed four times with PBST, followed by clearing in FUNGI solution (50% (v/v) glycerol, 9.4% (v/v) dH_2O, 10.6M Tris base, 1.1mM EDTA, 2.5M fructose and 2.5M urea) for 40min. Organoids were imaged using a μ-slide (Ibidi) on Zeiss Cell Observer Spinning Disc confocal microscope.

Hepatocytes differentiation and characterization

hiPSC lines were differentiated toward hepatocytes following the protocol from Baxter et al.[50] Briefly, cells were directed into Stage 1a/definitive endoderm-like cells by culturing in RPMI media containing 1mM L-glutamine, 0.5% FBS, 100ng/mL Activin-A and 25ng/mL Wnt3a for 2 days, followed by Stage 1b/definitive endoderm-like cells by culturing in RPMI media containing 1mM l-Glutamine, 0.5% FBS and 100ng/mL Activin-A for a further 2 days. Stage 2/hepatoblast-like cells was initiated by incubating for a further 6 days with Hepatocyte culture medium (HCM) containing 20ng/mL BMP2 and 30 ng/mL Fgf4. Hepatocyte-like cells were made by further incubating the cultures in Stage 3a media (HCM containing 20ng/mL Hepatocyte Growth Factor (Peprotech)) for 5 days followed by a further 15 days in Stage 2b media (HCM containing 10 μg/mL Oncostatin M (R&D Systems Cat No. 295-OM) and 10nM dexamethasone).

11 and C32, by incubation with P450-Glo™ CYP3A4 assay reagent (Promega). The analysis was performed according to manufacturer recommendation.

Micropatterns preparation

Micropattern chip fabrication

Micropattern chip fabrication was conducted using the protocol of Lee et al.[51], with specific modification and optimization. In brief, coverslips were sonicated in 70% ethanol for 15min and in deionized water for 15min. The clean coverslips were sequentially incubated in 0.5% (3-aminopropyl)triethoxysilane (APTS) (Merck) for 3min, 0.5% glutaraldehyde (Merck, Cat. No. G6257) for 30min. After air drying, the coverslips were deposited on a 20μL drop made of 10% acrylamide (Merck), 0.87% bisacrylamide (Merck), 0.1% ammonium persulfate (Merck, Cat. No. A3678) and 0.1% N,N,N',N'-Tetramethylethylenediamine (Merck), to make the gel at a stiffness of 100KPa. After the stiffness droplet was semi-solidified, the whole system was submerged into 70% ethanol, resulting in a smooth polyacrylamide gel forming. Gelled coverslips were sequentially coated with 64% hydrazine hydrate (Fisher Scientific) for 1h and 2% glacial acetic acid (Merck) for 1h. To generate polydimethylsiloxane (PDMS) stamp, SYLGARD™ 184 Silicone Elastomer Curing Agent and Base (Dow) were mixed at a 1:10 ratio before loading to the stamp mold, provided by the Kilian Lab at the University of New South Wales. Next, the solidified PDMS stamp was coated with 25μg/mL vitronectin (Life Technologies) and 3.5 mg/mL sodium periodate (Merck) for 1h. After air-drying the stamp, patterned vitronectin was stamped onto the gelled coverslip at 0.343N for 1min. Stamped gels were stored overnight in PBS + 1% Penicillin-Streptomycin at 4°C.

Germ layer differentiation on micropatterns and analysis

Differentiation protocol was adapted from Warmflash et al.[44] Since the micropatterned chip generation required many hands-on manipulations, all culture media were supplemented with 1% Penicillin-Streptomycin. hiPSCs were seeded as single cells to micropattern chip at a density of 2.5×10^5 cells/cm^2 with 10μM Y-27632 ROCK inhibitor in mTeSR Plus supplemented 1% Penicillin-Streptomycin into a total volume of 1mL per well of a 24-well plate. At about 80% confluency, germ layer differentiation was induced by adding 50 ng/mL BMP4 (R&D) in mTeSR1. The cells grown on micropatterns were washed with PBS, fixed in 4% paraformaldehyde in PBS at RT for 20 min. They were washed with PBS twice and then permeabilised with 0.1% Triton X-100 (Merck) in dPBS (Gibco) (PBST) at RT for 1 hour. The cells were blocked with 3% bovine serum albumin (Merck) in PBST at room temperature for 1 hour. They were incubated with primary antibody at 4°C overnight (MIXL1 (Abcam) 1:50, T/Brachyury (Santa Cruz) 1:50, FOXA2 (Abcam) 1:300, SOX17(R&D Systems) 1:20). Cells were washed with PBST three times, then incubated with corresponding secondary antibodies at RT for 1 hour. The cell nuclei were stained with DAPI (1 μg/ml) (Thermo Fisher Scientific) in dPBS for 10 min at RT, and then washed twice more with PBS. Cells were mounted with Fluoromount-G (Thermo Fisher Scientific). Micropatterned coverslips were imaged on Zeiss AiryScan LSM880 confocal microscope. All image analysis was performed using a custom macro. The nuclei from micropatterned images taken were segmented using the StarDist method[52] via Fiji software[53] using default parameters (except probability/score threshold = 0.7) and the versatile (fluorescent nuclei).pb model. R

software was used with a custom script where target protein immunofluorescence was normalized to the DAPI intensity of the same cell.

Microfluidic RT-qPCR preparation and analysis

RNA extraction

Snap frozen cell pellets had total RNA extracted using ISOLATE II RNA mini kit (Bioline) following manufacturer's instructions. Briefly, samples were lysed, homogenized and passed through a spin column containing a silica membrane to which the RNA binds. DNase 1 digestion eliminated potential genomic DNA contamination and the preparation was washed to remove impurities such as cellular debris and salts. The purified RNA was eluted with RNase free water and total RNA concentration was determined using Nanodrop ND-1000 Spectrophotometer (ThermoFisher Scientific). RNA was used either for Microfluidic RT-qPCR or RNA-sequencing.

cDNA synthesis and preparation

Total RNA was adjusted to a concentration of 200ng/μl. A 5μl reaction mix was prepared composing of 1μl Reverse Transcription Master Mix (Fluidigm), 3μl of RNase free water and 1μl of RNA and incubated in a thermocycler using the following conditions: 25°C for 5min, 42°C for 30min and 85°C for 5min.

cDNA preamplification

3.75μl of preamplification mix (comprising 105.6μl of Preamp MasterMix (Fluidigm), 52μl of 100 μM pooled primer and 237.6μl DNAse/RNAse free water) and 1.25μl of cDNA sample was added into each well of a 96 well plate and incubated as follows: 95°C for 2min then 10 cycles of 95°C for 15secand 60°C for 4min.

cDNA clean-up

cDNA clean-up was performed by adding 2μl of the following mix: 168μl DNase free water, 24μl 10x Exo1 reaction buffer and 48μl Exonuclease I (New England Biolabs), into each well and incubating in a thermocycler for 30min at 37°C followed by 15min at 80°C. Samples were diluted 10x with low EDTA TE buffer.

Primer and sample set-up

A sample mix was prepared as follows (per 96-well plate): 495μl of 2X SsoFast EvaGreen SuperMix with low ROX (Biorad) and 49.5μl 25X DNA Binding Dye (Fluidigm). 4.95μl sample mix was added with 4.05μl of diluted sample. Primers were prepared in the following mix (per 96-well plate): 450μl 2X Assay Loading Reagent (Fluidigm) and 405μl low EDTA TE buffer. 105μl of primer mix was added with 0.45μl combined forward and reverse primers. Samples and primer mixes were loaded onto a 96.96 Dynamic Array IFC plate (Fluidigm) and run on the Biomark System.

ATAC-seq samples preparation and analysis

Cell Preparation

hiPSC were collected at Day1 of the DE diff protocol (STEMDiff) and were processed following [54]. Briefly, 5 x 10^4 cells were collected at Day 1 of DE differentiation and lysed in cold lysis buffer (10 mM Tris·Cl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% (v/v) Igepal CA-630). Intact nuclei were separated by centrifugation at 500xg and immediately digested in transposase mix containing 25μl 2x Tagment DNA buffer, 2.5μl Tagment DNA enzyme I (Illumina) and 22.5μl nuclease-free water for 30min at 37°C. Digested chromatin fragments were then purified using the MinElute PCR Purification Kit (Qiagen), according to manufacturer's instructions. The fragments of DNA were then pre-amplified by adding 10μL purified DNA sample, 10μL RNase-free water, 2.5μL of each primer (Where each reaction had non-barcoded primer "Ad1_noMix" and one barcoded primer 'Ad2.1' - 'Ad2.9' added) and 25μL NEBNext High-Fidelity 2x PCR Master Mix (NEB) and was run under the following conditions: 72°C for 5min, 98°C for 30secand then 5 cycles of 98°C for 10sec, 63°C for 30sec and 72°C for 1min. The number of additional cycles to run was calculated by running a RT-qPCR side reaction - a reaction mixture containing 5μL of the pre-amplified PCR product, 3.9μL nuclease-free water, 0.25μL of each primer, 0.6μl 25x SYBR Green and 5μL NEBNext High-Fidelity 2x PCR Master Mix was run under the following conditions: 98°C for 30secand then 20 cycles of 98°C for 10sec, 63°C for 30sec and 72°C for 1min. The linear fluorescence versus cycle number was plotted and the cycle number (N) required to reach one-third of the maximum relative fluorescence was determined. The final amplification reaction (the remaining 45μl pre-amplified PCR product) was run under the following conditions: 98°C for 30sec and then N cycles of 98°C for 10sec, 63°C for 30sec and 72°C for 1min. Amplified samples were then purified using AMPure XP magnetic beads (Beckman Coulter) to remove small fragments and primer-dimers less than 100 bp long (1.3x beads) and large fragments (0.5x beads) using a Dynamag-2 magnet (Thermo Fisher Scientific).

To determine the integrity, fragment size and concentration, the DNA library was analyzed using the Agilent HSD5000 ScreenTape System (Agilent). Libraries were then 101 bp paired-end sequenced on an Illumina HiSeq 4000 (Illumina)

Data analysis

ATAC-seq reads were processed using the alignment and filtering functions of the PreNet pipeline[54]. Paired-end reads were mapped to the hg38 genome using bowtie2[55], allowing for local mapping, a maximum insert size of 2000 bp and a maximum of 4 multimapping hits (–local -X 2000 -k 4). Multimapping reads were allocated using 'assignmultimappers.py' from the ENCODE ATAC-seq pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline/tree/master/src). Reads with MAPQ < 30 were excluded and only unique, paired reads that aligned outside blacklisted regions[56] were used for subsequent analyses. Filtering steps were performed using samtools[57] and sambamba[58]. Qualifying reads were then converted to pseudo-single end reads and peaks were detected using MACS2[59], with BED input files and reads shifted by -100 bp and extended to 200bp to capture Tn5 transposase events: -f BED –shift -100 –extend 200 -q 0.05. Biological replicates were analyzed individually and then consensus peak list was created to include only peaks appearing in at least two of the three replicates. Accessibility within a pooled consensus peak list was estimated by quantifying Tn5 events in each of the biological replicates using featureCounts[60]. The DESeq2[61] package within

R was used to identify differentially accessible regions. The regions were filtered for adjusted P¬-value < 0.05 and an absolute Log2(Fold change) > 1. Coverage tracks were created using the bamCoverage function of deepTools[62] (–normalization RPKM -bs 10) and visualized within Integrative Genomics Viewer.[63] findMotifsGenome.pl from HOMER[64] was used to identify over enriched motifs, between 6 bp and 12 bp in size, within regions of differential accessibility using a repeat masked version of the hg38 sequence (-mask -len 6,7,8,9,10,11,12). Coverage tracks summarizing and combining biological replicates were created using WiggleTools[65] to quantify the mean coverage per 10 bp bin. These tracks were used for heatmap visualizations created using plotHeatmap from deeptools.

Bulk RNA sequencing analysis

Cell preparation and library prep

Illumina RNA Library prep was performed by GENEWIZ (https://www.genewiz.com/Public/Services/Next-Generation-Sequencing). Samples at Day 1 of DE differentiation (1 - 20μg RNA) were run on HiSeq4000 with a read depth of 20M paired end reads (2x 150PE).

Data pre-processing

Details of the procedure can be found in Aryamanesh and colleagues [66]

Statistical Analysis

Details of the procedure can be found in Aryamanesh and colleagues [66]

Single-cell RNA sequencing

Cell preparation

hiPSCs were dissociated using Accutase and counted to load 10,000 cells into on channel of a 10X Chromium chip. One channel per sample was used. After emulsion cell lysis and RNA was extracted followed by library preparation. Libraries were sent to Novogene for sequencing.

Single cell suspensions were passed through 40μm cell strainer (Corning) and concentration was adjusted to 1000 cells/μL. Suspensions were loaded in single-cell-G Chip (10X Genomics) for target output of 10,000 cells per sample. Single-cell droplet capture was performed on the Chromium Controller (10X Genomics). cDNA library preparation was performed in accordance with the Single-Cell 3' v 3.0 or v3.1 protocol. Libraries were evaluated for fragment size and concentration using Agilent HSD5000 ScreenTape System (Agilent).

Samples were sequenced on an Illumina HiSeq4000 instrument according to manufacturer's instructions (Illumina). Sequencing was carried out using 2x150 paired-end (PE) configuration with a sequencing depth of 20,000 reads per cell. Sequencing was performed by GENEWIZ.

Data pre-processing

Details of the procedure can be found in Aryamanesh and colleagues [66]

Statistical Analysis

Details of the procedure can be found in Aryamanesh and colleagues [66]

Proteomics

Sample were prepared and data were obtained by the Proteomic facility at CMRI.

Proteomics sample preparation and Mass spectrometry

Eleven hiPSC lines from four isogenic groups were prepared for proteomics analyses using the Accelerated Barocycler lysis and extraction digestion sample preparation method[67]. The tryptic peptides were desalted using Waters Oasis C18 HLB 30mg SPE cartridges. The amount of peptide in each sample was measured using the absorption of 280 nm light (Implen Nanophotometer, Labgear, Australia).

A reference sample was prepared by pooling equal amount of peptide from 25 of the 45 hiPSC samples. Aliquots containing 10 µg of peptide from each sample were labeled with tandem mass tag (TMT) 16-plex reagents (TMTpro, Thermo Fisher Scientific), according to the manufacturer's instructions. Three separate TMT16plex sets were prepared with eleven samples per set and the reference sample included in each set.

#	hiPSC cell line_Name	TMT_ Set-1	TMT_ Set-2	TMT_ Set-3	TMTpro 16-plex label
1	Eu79	+	+	+	126
2	Eu86	+	+	+	127N
3	Eu87	+	+	+	127C
4	C9	+			128N
5	C11			+	128C
6	C16	+			129N
7	C7	+			129C
8	C32	+			130C
9	C4	+	+	+	131N
10	C2	+	+	+	131C
11	C3		+	+	132N
12	Universal / reference sample				134

+, samples were combined to produce the reference sample used as the common sample in the three TMT 16plex experiments.

High pH fractionation was performed using the Pierce High pH Reversed-Phase Peptide Fractionation Kit according to the manufacturer's instructions. A total of 80 µg of peptide for each of the TMT16-plex sets were loaded, washed with water then 5% Acetonitrile/0.1% Triethylamine solution. A total of fifteen High-pH step elution's were collected from 8% to 50% Acetonitrile / 0.1% Triethylamine. The high-pH elution's were dried to completeness and resuspended in 0.1% Formic acid and the peptide concentration was determined using the absorption of 280 nm light.

The peptides from each High pH elution were resolved by reversed phase chromatography on a 300 x 0.075 mm column packed with ReproSil Pur C18 AQ 1.9 mm resin (Dr Maisch, Germany) using an Ultimate 3000 RSLC nano system (Thermo Fisher Scientific). The column was heated to 50 °C using a column oven (PRSO-V1, Sonation lab solutions).

The chromatography buffer A was 0.1% formic acid in water and buffer B was 0.1% formic acid, 90% acetonitrile and 9.99% water and the flow rate was 250 nL/min. For each high pH fraction, 1 to 2 µg of peptide was directly loaded onto the column in 99% buffer A and 1% buffer B for 30 min. The gradient started from 1% to 7% buffer B in 6 min, then to 30% buffer B in 51 min, then to 35% buffer B in 10 min and to 99% buffer B in 3 min and held at 99% buffer B for 8 min. MS acquisition was performed for the entire 120 min. The fifteen high pH elution steps collected for each TMT16-plex set were individually analysed using a data-dependent acquisition LC-MS/MS method. Between each TMT16plex set, one blank was run.

Peptides were detected by tandem mass spectrometry using a Q Exactive Plus hybrid quadrupole-orbitrap mass spectrometer (Thermo Fisher Scientific). The Nanospray Flex ion source (Thermo Fisher Scientific) spray operated at 2.3 kV. The capillary temperature was 250°C and the S lens radio frequency level was 50. The MS scan was from m/z 375 to 1500 at a resolution of 70,000 full width at half maximum with an automatic gain control target of 3×10^6 counts for a maximum ion time of 100 ms. For each MS scan, up to 12 of the most intense ions above a threshold of 5.2×10^4 counts were selected for an MS/MS scan. MS/MS scans were at a resolution of 35,000 full width at half maximum for a maximum ion time of 115 ms and automatic gain control target of 2×10^5 counts. The isolation window was 1.1 units of the m/z scale, the fixed first mass was set at m/z 120 and the normalized collision energy was 30. Peptides with charge state <2 + or >8 + or with unassigned charge were excluded. Dynamic exclusion of previously scanned peptides was for 35 s.

The raw LC-MS/MS data were processed with MaxQuant v1.6.7.0 [68] using the following settings: The fasta file was the Human reference proteome downloaded from UniProtKB on January 12, 2022 and containing 101,017 entries including protein isoforms and Retention time standards. Protease specificity was Trypsin/P with up to 3 missed cleavages. Carbamidomethyl (C) was a fixed modification and the TMTpro 16plex reagents were designated isobaric labels. Deamidation (N and Q), oxidation (M) and acetylation (protein N-terminus) were variable modifications. A maximum of 5 modifications per peptide was allowed. The minimum score for modified peptides was 40. The minimum peptide length was 6 and maximum peptide mass was 6,000 Da. The peptide spectrum match, protein, and modification site false discovery rate was 1%. A dependent peptide search was performed with a 1% false discovery rate. Modified peptides and their counterpart non-modified peptides were excluded from protein quantification. A second peptide search was enabled. The tolerance for MS and MS/MS spectra was 4.5 ppm

and 20 ppm, respectively. All other settings were left as the default within MaxQuant v1.6.7.0. The three TMTpro 16plex sets were searched together (Set1, Set2, Set3); fractions 1 to 15 for each TMT set representing the High pH elution fractions.

Data cleaning, normalization and hypothesis testing

The data cleaning, normalization and hypothesis testing were performed using the ProteomeRiver pipeline[69] and the implementation is briefly described here. To extract the protein abundance data from the MaxQuant output, the 'proteinGroups.txt' output file from MaxQuant[70] were processed. Each protein group must have at least one unique peptide. Any proteins were removed from further analysis if they match any entries in the contaminants or the reverse sequence decoy databases, in which the protein accession starts with CON_ or REV_ prefixes respectively. The 'reporter intensity corrected' column were used for further analysis. Proteins with one or more missing values in any samples were removed from further analysis.

The following rules from Engholm-Keller et al.[71] were used to identify a representative UniProt accession for each protein group. 1. For proteins that mapped to multiple UniProtKB protein accessions, the accession with the highest 'protein existence (PE)' value was kept as the best evidence. Where the protein accession was an isoform (therefore lacking PE information), the PE value was taken from the parent protein. 2. When the PE value was equal, a Swiss-Prot (sp) entry was taken over a TrEMBL (tr) entry. 3. If both entries were Swiss-Prot, the non-isoform was selected. 4. If both entries were isoforms, the longest isoform was selected.

To perform data normalization, samples were log (base 2) transformed and between sample normalization were performed using scaled normalization from the 'limma' R package. The remove unwanted variation 'ruv' R package[72] was used to remove batch effects. The method relies on having a set of endogenous negative control proteins, which are proteins with little changes in protein abundances between different cell types or experimental treatments. For this study, a set of 500 empirical negative control proteins with high q-values indicating little or no change in protein expression across sample were identified from an initial ANOVA test. The RUVIII method[73] was used to remove the unwanted variations across the samples and six unwanted components (k = 7) were removed by the tool. The RUVIII method requires the experiment design matrix, a matrix describing the replicates for each treatment condition, and the list of negative control proteins.

Differential abundance analysis of proteins was performed using the adjusted abundance matrix. Differential abundance analysis of proteins involved pairwise hypothesis testing of samples from FA3 cell line with samples of another type of cell line and all possible pairs were analyzed. Linear model for comparing each pair of time points was fitted using the 'lmFit' function and the p-values calculated using the empirical Bayes method 'eBayes' function. Trended and robust analysis were enabled. The false discovery rate correction was applied to the moderated p-values by calculating the q-values[74]. Significant differentially expressed proteins were defined as those with q-values less than 0.05.

Clustering Analysis

For a protein to be included in the clustering, it must be statistically down-regulated (q-value <0.05) in six or more FA3 versus another cell line, or statistically up-regulated (q-value <0.05) in six or more FA3 versus another cell line. The protein must also be statistically significant (q-value <0.05) in at least one of the comparisons of FA3 with any of the KO or DOX cell lines. The z-standardized log2 abundance of the sample, excluding samples of the KO or DOX cell lines, were used in the clustering analysis (e.g. 10 values were used in clustering). Consensus clustering were performed 'diceR' R library[75]. All statistically significant differentially abundant proteins. After assessment with the consensus clustering tools, using multiple clustering algorithms, including 'pam', 'km', 'som', 'hc', and 'diana', and different distance metrics 'euclidean', 'canberra', 'minkowski', and 'spearman', self-organizing maps (som) with seven clusters (k=7) was identified to be a reliable method to use. The diceR tool automatically identifies robust consensus clusters by merging the results from 100 runs of self-organizing maps. The clusters identified were used for functional enrichment analysis (see below for details).

Functional Enrichment

Functional enrichments were performed using the Fisher's exact test implemented in the 'clusterProfiler' R library[76]. The background list of proteins included all the proteins in the dataset after the data cleaning step. The query list of proteins includes the following: 1) significantly differentially abundant proteins with positive log fold-change, 2) significantly differentially abundant proteins with negative log fold-change, 3) the list of proteins from each cluster from self-organizing maps. Gene ontology annotations from UniProt[77] and the KEGG[78] and Reactome[79] pathway databases were used for the enrichment analyses.

Quantification and Statistical Analysis

Statistical analysis was performed using R software. The type of statistical test performed, the meaning of dispersion and precision measurements as well as the significance of each experiment is indicated in the corresponding figure, figure legends and/or in the method details. Outliers have been omitted to facilitate visualization. For micropattern quantification, images were taken from 3 to 10 micropatterns within a coverslip. For microfluidic PCR, three biological replicates were collected for each cell line and timepoint.

Supplementary Figures

Figure S1: A) Brightfield pictures of definitive endoderm differentiation for the 5 cell lines. Day4 images were zoomed in (inlet) to show the cell morphology of the endodermal cells at the end of the differentiation protocol. Scale bar are indicated on the lowest right image of each panel; **B)** Immunostaining quantification of FOXA2 (green) and SOX17 (magenta); related to Figure 1; **C)** Immunostaining images on AAT (green) and ALB (magenta) markers of hepatocytes differentiation;

Figure S2: A) Genes' expression time course during DE differentiation related to **i)** Pluripotency, **ii)** Endoderm, **iii)** Mesoderm and **iv)** Ectoderm. C32 is highlighted in purple. B) Heatmap of differentially expressed genes (DEGs) of C32 cell lines versus the rest of the cell line used in the study. **C)** Venn diagram of up and down DEGs compared with relevant ontologies. **D)** PCA representing data of bulk RNAseq of Day1 samples. **E)** DEGs between C32 and C7. Dots that are colored have a p.value < 0.05.

Figure S3: A) Schematic of the annotation transfer performed in our study; **B)** Result of the annotation transfer applied on our scRNA-seq dataset; **C)** The same results facetted for each cell line; **D)** Stacked barplot representing the proportion of each cell type for each cell line.

Figure S4: A) PCA representing the proteomic signature of hiPSC**; B)** Volcano plot representing peptides significantly differentially expressed between C32 and C7; **C)** Allocation of peptides to proteins and respective LogFC

Figure S5: Staining for SOX2 (green) and OCT4 (magenta) on three colonies of **A)** C16; **B)** C16-MKO; **C)** C32; **D)** C32-MKO**.** Scale bar = 100µm

Figure S6: A) Volcano plot representing Differentially Accessible Chromatins (DACs) region between C16 and C16-MKO. **B and C)** Heatmap showing ATAC-seq signal from C16 and C16-KO for DACs **(B)** more accessible in C16 and **(C)** in C16-MKO. **D and E)** Motifs enriched in DACs more accessible in **(D)** C16 and **(E)** C16-MKO. **F)** Genomic region of 350bp upstream of Mixl1 TSS. 7 guides RNA predicted by Benchling are highlighted. **G)** RT-qPCR results on *Mixl1* expression in HEK cells transfected with single guides or tandems.

Figure S7: Panels of 10 micropatterns at 48h post BMP4 treatment, stained with FOXA2, SOX17 and DAPI for **A)** C32, **B)** C32-MKO, **C)** C16 and **D)** C16-MKO. **E)** Indicates the average cell density for all micropattern analyzed.

Figure S8: A) Panels of 9 micropatterns at 48h post BMP4 treatment, stained with FOXA2, SOX17 and DAPI for C32-iMXL1 at 0, 1 and 2µg/mL of Doxycycline. **B)** Indicates the average cell density for all micropattern analyzed.

References

1. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell* **131**, 861–872 (2007).
2. Dimos, J. T. *et al.* Induced Pluripotent Stem Cells Generated from Patients with ALS Can Be Differentiated into Motor Neurons. *Science* **321**, 1218–1221 (2008).
3. Friedman, C. E. *et al.* Single-Cell Transcriptomic Analysis of Cardiac Differentiation from Human PSCs Reveals HOPX-Dependent Cardiomyocyte Maturation. *Cell Stem Cell* **23**, 586-598.e8 (2018).
4. Nayler, S. P. *et al.* Human iPSC-Derived Cerebellar Neurons from a Patient with Ataxia-Telangiectasia Reveal Disrupted Gene Regulatory Networks. *Front. Cell. Neurosci.* **11**, 321 (2017).
5. Park, I.-H. *et al.* Disease-Specific Induced Pluripotent Stem Cells. *Cell* **134**, 877–886 (2008).
6. Takasato, M. *et al.* Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis. *Nature* **526**, 564–568 (2015).
7. Cheng, X. *et al.* Self-Renewing Endodermal Progenitor Lines Generated from Human Pluripotent Stem Cells. *Cell Stem Cell* **10**, 371–384 (2012).
8. D'Amour, K. A. *et al.* Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nat. Biotechnol.* **23**, 1534–1541 (2005).
9. Loh, K. M. *et al.* Efficient Endoderm Induction from Human Pluripotent Stem Cells by Logically Directing Signals Controlling Lineage Bifurcations. *Cell Stem Cell* **14**, 237–252 (2014).
10. Green, M. D. *et al.* Generation of anterior foregut endoderm from human embryonic and induced pluripotent stem cells. *Nat. Biotechnol.* **29**, 267–272 (2011).

911 11. McCracken, K. W., Howell, J. C., Wells, J. M. & Spence, J. R. Generating human intestinal tissue
912 from pluripotent stem cells in vitro. *Nat. Protoc.* **6**, 1920–1928 (2011).
913 12. Miura, S. & Suzuki, A. Brief summary of the current protocols for generating intestinal organoids.
914 *Dev. Growth Differ.* **60**, 387–392 (2018).
915 13. Spence, J. R. *et al.* Directed differentiation of human pluripotent stem cells into intestinal tissue
916 in vitro. *Nature* **470**, 105–109 (2011).
917 14. D'Amour, K. A. *et al.* Production of pancreatic hormone–expressing endocrine cells from human
918 embryonic stem cells. *Nat. Biotechnol.* **24**, 1392–1401 (2006).
919 15. Hannan, N. R. F., Segeritz, C.-P., Touboul, T. & Vallier, L. Production of hepatocyte-like cells from
920 human pluripotent stem cells. *Nat. Protoc.* **8**, 430–437 (2013).
921 16. Varghese, D. S., Alawathugoda, T. T. & Ansari, S. A. Fine Tuning of Hepatocyte Differentiation
922 from Human Embryonic Stem Cells: Growth Factor *vs.* Small Molecule-Based Approaches. *Stem*
923 *Cells Int.* **2019**, 1–18 (2019).
924 17. Basma, H. *et al.* Differentiation and Transplantation of Human Embryonic Stem Cell–Derived
925 Hepatocytes. *Gastroenterology* **136**, 990-999.e4 (2009).
926 18. Hay, D. C. *et al.* Highly efficient differentiation of hESCs to functional hepatic endoderm requires
927 ActivinA and Wnt3a signaling. *Proc. Natl. Acad. Sci.* **105**, 12301–12306 (2008).
928 19. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic
929 genetic effects on gene expression. *Nat. Commun.* **11**, 810 (2020).
930 20. Burrows, C. K. *et al.* Genetic Variation, Not Cell Type of Origin, Underlies the Majority of
931 Identifiable Regulatory Differences in iPSCs. *PLOS Genet.* **12**, e1005793 (2016).
932 21. Kajiwara, M. *et al.* Donor-dependent variations in hepatic differentiation from human-induced
933 pluripotent stem cells. *PNAS* **109**, 12538–12543 (2012).

934 22. Kyttälä, A. *et al.* Genetic Variability Overrides the Impact of Parental Cell Type and Determines
935 iPSC Differentiation Potential. *Stem Cell Rep.* **6**, 200–212 (2016).
936 23. Nishizawa, M. *et al.* Epigenetic Variation between Human Induced Pluripotent Stem Cell Lines Is
937 an Indicator of Differentiation Capacity. *Cell Stem Cell* **19**, 341–354 (2016).
938 24. Ohi, Y. *et al.* Incomplete DNA methylation underlies a transcriptional memory of somatic cells in
939 human iPS cells. *Nat. Cell Biol.* **13**, 541–549 (2011).
940 25. Phetfong, J. *et al.* Cell type of origin influences iPSC generation and differentiation to cells of the
941 hematoendothelial lineage. *Cell Tissue Res.* **365**, 101–112 (2016).
942 26. Shao, K. *et al.* Induced Pluripotent Mesenchymal Stromal Cell Clones Retain Donor-derived
943 Differences in DNA Methylation Profiles. *Mol. Ther.* **21**, 240–250 (2013).
944 27. Sanchez-Freire, V. *et al.* Effect of Human Donor Cell Source on Differentiation and Function of
945 Cardiac Induced Pluripotent Stem Cells. *J. Am. Coll. Cardiol.* **64**, 436–448 (2014).
946 28. Pearce, J. J. H. & Evans, M. J. Mml, a mouse Mix-like gene expressed in the primitive streak. *Mech.*
947 *Dev.* **87**, 189–192 (1999).
948 29. Wolfe, A. D. & Downs, K. M. Mixl1 localizes to putative axial stem cell reservoirs and their
949 posterior descendants in the mouse embryo. *Gene Expr. Patterns* **15**, 8–20 (2014).
950 30. Hart, A. H. *et al.* Mixl1 is required for axial mesendoderm morphogenesis and patterning in the
951 murine embryo. *Development* **129**, 3597–3608 (2002).
952 31. Ng, E. S. *et al.* The primitive streak gene *Mixl1* is required for efficient haematopoiesis and BMP4-
953 induced ventral mesoderm patterning in differentiating ES cells. *Development* **132**, 873–884
954 (2005).
955 32. Lim, S. M. *et al.* Enforced Expression of *Mixl1* During Mouse ES Cell Differentiation Suppresses
956 Hematopoietic Mesoderm and Promotes Endoderm Formation. *Stem Cells* **27**, 363–374 (2009).

957 33. Tewary, M. *et al. High-Throughput Micro-Patterning Platform Reveals Nodal-Dependent*
958 *Dissection of Peri-Gastrulation-Associated versus Pre-Neurulation Associated Fate Patterning*.
959 http://biorxiv.org/lookup/doi/10.1101/465039 (2018) doi:10.1101/465039.
960 34. Becherel, O. J. *et al.* A new model to study neurodegeneration in ataxia oculomotor apraxia type
961 2. *Hum. Mol. Genet.* **24**, 5759–5774 (2015).
962 35. Briggs, J. A. *et al.* Integration-Free Induced Pluripotent Stem Cells Model Genetic and Neural
963 Developmental Features of Down Syndrome Etiology. *Stem Cells* **31**, 467–478 (2013).
964 36. Tyser, R. C. V. *et al.* Single-cell transcriptomic characterization of a gastrulating human embryo.
965 *Nature* **600**, 285–289 (2021).
966 37. Wang, X. *et al.* Fn1 Regulates the Third Pharyngeal Pouch Patterning and Morphogenesis. *J. Dent.*
967 *Res.* **101**, 1082–1091 (2022).
968 38. Omachi, K. & Miner, J. H. Comparative analysis of dCas9-VP64 variants and multiplexed guide
969 RNAs mediating CRISPR activation. *PLoS ONE* **17**, e0270008 (2022).
970 39. Jinek, M. *et al.* A programmable dual RNA-guided DNA endonuclease in adaptive bacterial
971 immunity. *Science* **337**, 816–821 (2012).
972 40. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR-Cas9-based transcription factors.
973 *Nat. Methods* **10**, 973–976 (2013).
974 41. Izzi, L. *et al.* Foxh1 recruits Gsc to negatively regulate Mixl1 expression during early mouse
975 development. *EMBO J.* **26**, 3132–3143 (2007).
976 42. Ota, M. & Sasaki, H. Mammalian Tead proteins regulate cell proliferation and contact inhibition
977 as transcriptional mediators of Hippo signaling. *Development* **135**, 4059–4069 (2008).

978 43. Morgani, S. M., Metzger, J. J., Nichols, J., Siggia, E. D. & Hadjantonakis, A.-K. Micropattern
979 differentiation of mouse pluripotent stem cells recapitulates embryo regionalized cell fate
980 patterning. *eLife* **7**, e32839 (2018).
981 44. Warmflash, A., Sorre, B., Etoc, F., Siggia, E. D. & Brivanlou, A. H. A method to recapitulate early
982 embryonic spatial patterning in human embryonic stem cells. *Nat. Methods* **11**, 847–854 (2014).
983 45. Plouhinec, J.-L., Simon, G., Vieira, M., Collignon, J. & Sorre, B. Dissecting signaling hierarchies in
984 the patterning of the mouse primitive streak using micropatterned EpiLC colonies. *Stem Cell Rep.*
985 S2213-6711(22)00262–4 (2022) doi:10.1016/j.stemcr.2022.05.009.
986 46. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the
987 organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
988 47. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308
989 (2013).
990 48. Bauer, D. E., Canver, M. C. & Orkin, S. H. Generation of Genomic Deletions in Mammalian Cell
991 Lines via CRISPR/Cas9. *J. Vis. Exp.* 52118 (2014) doi:10.3791/52118.
992 49. Etard, C., Joshi, S., Stegmaier, J., Mikut, R. & Strähle, U. Tracking of Indels by DEcomposition is a
993 Simple and Effective Method to Assess Efficiency of Guide RNAs in Zebrafish. *Zebrafish* **14**, 586–
994 588 (2017).
995 50. Baxter, M. *et al.* Phenotypic and functional analyses show stem cell-derived hepatocyte-like cells
996 better mimic fetal rather than adult hepatocytes. *J. Hepatol.* **62**, 581–589 (2015).
997 51. Lee, L. H. *et al.* Micropatterning of human embryonic stem cells dissects the mesoderm and
998 endoderm lineages. *Stem Cell Res.* **2**, 155–162 (2009).
999 52. Weigert, M., Schmidt, U., Haase, R., Sugawara, K. & Myers, G. Star-convex Polyhedra for 3D Object
1000 Detection and Segmentation in Microscopy. in *2020 IEEE Winter Conference on Applications of*

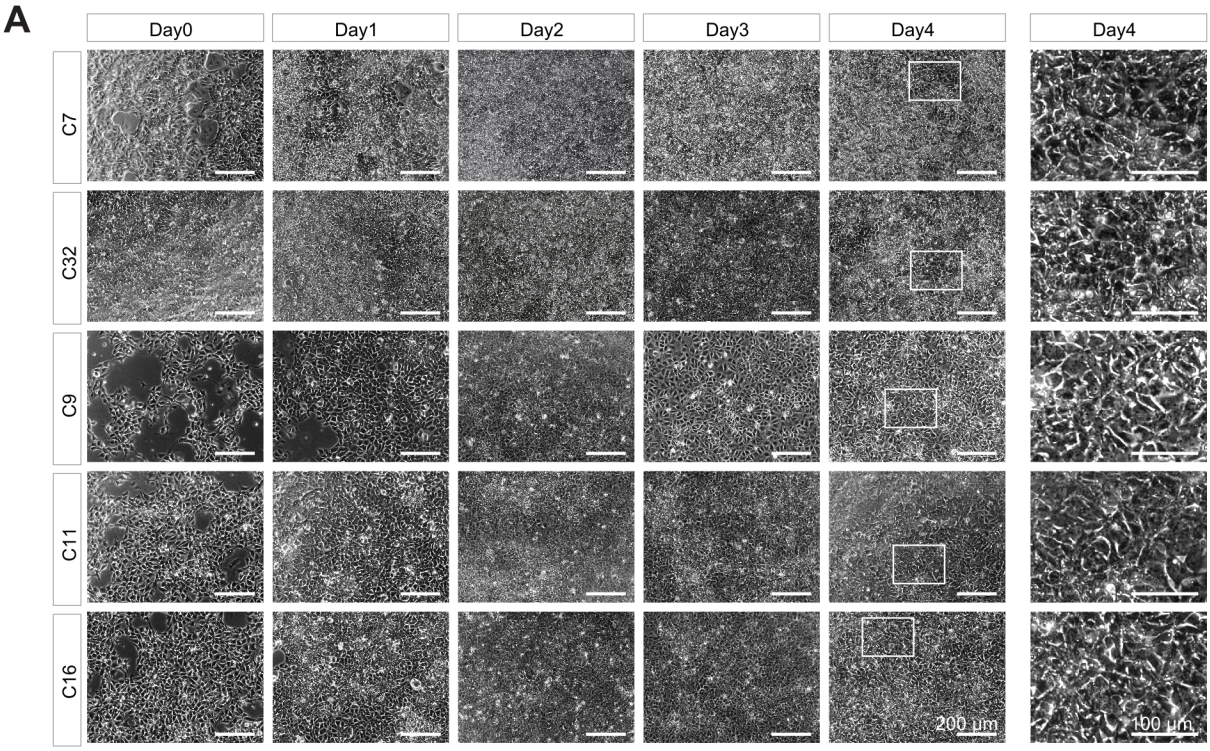
1001 *Computer Vision (WACV)* 3655–3662 (IEEE, Snowmass Village, CO, USA, 2020).
1002 doi:10.1109/WACV45572.2020.9093435.
1003 53. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**,
1004 676–682 (2012).
1005 54. Salehin, N., Tam, P. P. L. & Osteil, P. Prenet: Predictive network from ATAC-SEQ data. *J. Bioinform.*
1006 *Comput. Biol.* **18**, 2040003 (2020).
1007 55. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–
1008 359 (2012).
1009 56. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic
1010 Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).
1011 57. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079
1012 (2009).
1013 58. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS
1014 alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
1015 59. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
1016 60. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning
1017 sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
1018 61. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-
1019 seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
1020 62. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis.
1021 *Nucleic Acids Res.* **44**, W160–W165 (2016).
1022 63. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-
1023 performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

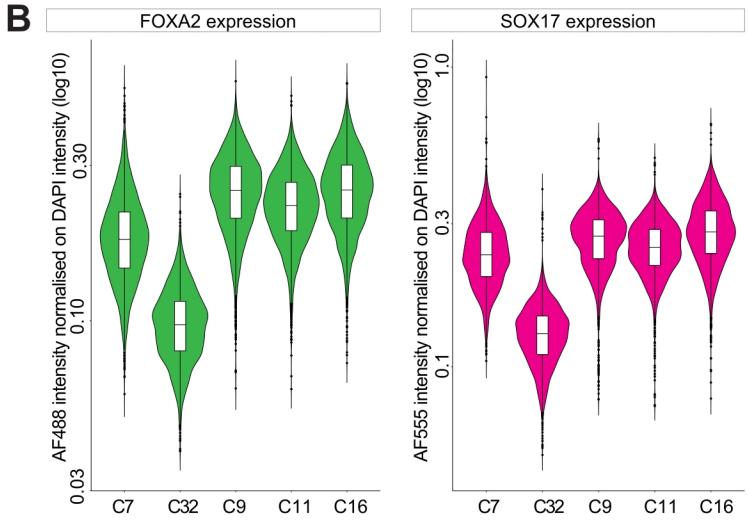
1024 64. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-
1025 regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589
1026 (2010).
1027 65. Zerbino, D. R., Johnson, N., Juettemann, T., Wilder, S. P. & Flicek, P. WiggleTools: parallel
1028 processing of large collections of genome-wide datasets for visualization and statistical analysis.
1029 *Bioinformatics* **30**, 1008–1009 (2014).
1030 66. Aryamanesh, N. A Reproducible and Dynamic Workflow for Analysis and Annotation of scRNA-
1031 Seq Data. *Methods Mol. Biol.* **2490**, 101–140 (2022).
1032 67. Lucas, N. *et al.* Accelerated Barocycler Lysis and Extraction Sample Preparation for Clinical
1033 Proteomics by Mass Spectrometry. *J. Proteome Res.* **18**, 399–405 (2019).
1034 68. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-
1035 based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
1036 69. Pang, Ignatius C, Waardenberg, Ashley J, Aryamanesh, Nader, & Graham, Mark E. ProteomeRiver
1037 pipeline. (2023).
1038 70. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-
1039 range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–
1040 1372 (2008).
1041 71. Engholm-Keller, K. *et al.* The temporal profile of activity-dependent presynaptic phospho-
1042 signalling reveals long-lasting patterns of poststimulus regulation. *PLOS Biol.* **17**, e3000170
1043 (2019).
1044 72. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in
1045 microarray data. *Biostatistics* **13**, 539–552 (2012).

1046 73. Molania, R., Gagnon-Bartsch, J. A., Dobrovic, A. & Speed, T. P. A new normalization for Nanostring
1047 nCounter gene expression data. *Nucleic Acids Res.* **47**, 6073–6083 (2019).
1048 74. Storey, J. D. A Direct Approach to False Discovery Rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**,
1049 479–498 (2002).
1050 75. Chiu, D. S. & Talhouk, A. diceR: an R package for class discovery using an ensemble driven
1051 approach. *BMC Bioinformatics* **19**, 11 (2018).
1052 76. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The*
1053 *Innovation* **2**, 100141 (2021).
1054 77. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids*
1055 *Res.* **51**, D523–D531 (2023).
1056 78. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for
1057 taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
1058 79. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–
1059 D692 (2022).

1060

Figure S1


A
Day0
Day1
Day2
Day3
Day4
Day4
C7
C32
C9
C11
C16
200 µm
100 µm


B
FOXA2 expression
SOX17 expression
AF488 intensity normalised on DAPI intensity (log10)
0.03
0.10
0.30
AF555 intensity normalised on DAPI intensity (log10)
0.1
0.3
1.0
C7
C32
C9
C11
C16

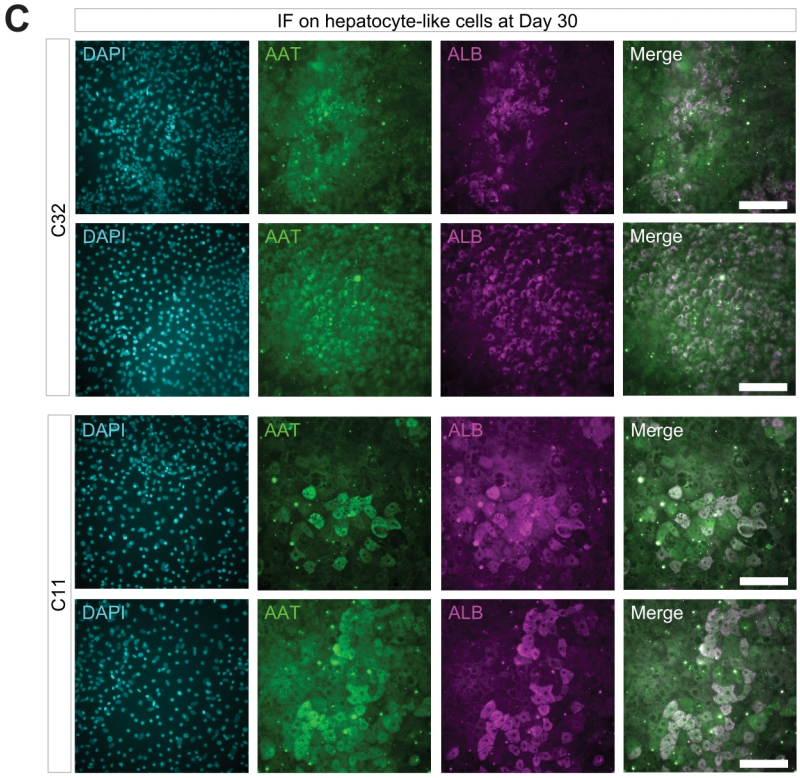

C
IF on hepatocyte-like cells at Day 30
C32
C11
DAPI
AAT
ALB
Merge

Figure S2

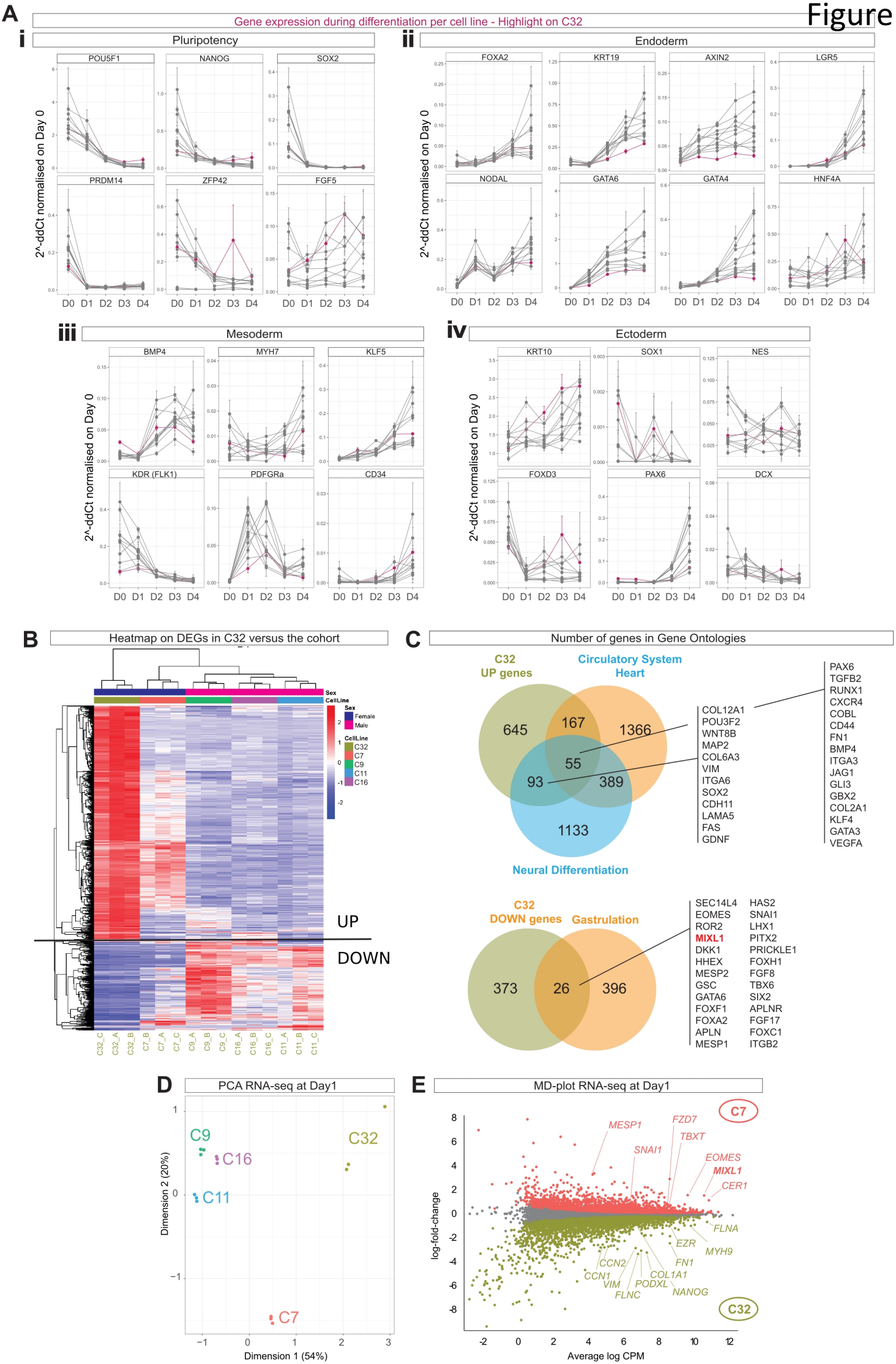

A
Gene expression during differentiation per cell line - Highlight on C32
i
Pluripotency
POU5F1
NANOG
SOX2
PRDM14
ZFP42
FGF5
ii
Endoderm
FOXA2
KRT19
AXIN2
LGR5
NODAL
GATA6
GATA4
HNF4A
iii
Mesoderm
BMP4
MYH7
KLF5
KDR (FLK1)
PDFGRa
CD34
iv
Ectoderm
KRT10
SOX1
NES
FOXD3
PAX6
DCX
2^-ddCt normalised on Day 0
D0 D1 D2 D3 D4
B
Heatmap on DEGs in C32 versus the cohort
Sex
CellLine
Female
Male
C32
C7
C9
C11
C16
UP
DOWN
C32_C C32_A C32_B C7_B C7_A C7_C C9_A C9_B C9_C C16_A C16_B C16_C C11_A C11_B C11_C
C
Number of genes in Gene Ontologies
C32
UP genes
Circulatory System
Heart
645
167
1366
55
93
389
1133
Neural Differentiation
COL12A1
POU3F2
WNT8B
MAP2
COL6A3
VIM
ITGA6
SOX2
CDH11
LAMA5
FAS
GDNF
PAX6
TGFB2
RUNX1
CXCR4
COBL
CD44
FN1
BMP4
ITGA3
JAG1
GLI3
GBX2
COL2A1
KLF4
GATA3
VEGFA
C32
DOWN genes
Gastrulation
373
26
396
SEC14L4
EOMES
ROR2
MIXL1
DKK1
HHEX
MESP2
GSC
GATA6
FOXF1
FOXA2
APLN
MESP1
HAS2
SNAI1
LHX1
PITX2
PRICKLE1
FOXH1
FGF8
TBX6
SIX2
APLNR
FGF17
FOXC1
ITGB2
D
PCA RNA-seq at Day1
C9
C16
C11
C32
C7
Dimension 2 (20%)
Dimension 1 (54%)
E
MD-plot RNA-seq at Day1
C7
C32
MESP1
FZD7
TBXT
SNAI1
EOMES
MIXL1
CER1
FLNA
EZR
MYH9
FN1
CCN2
CCN1
VIM
FLNC
PODXL
COL1A1
NANOG
log-fold-change
Average log CPM

Figure S3

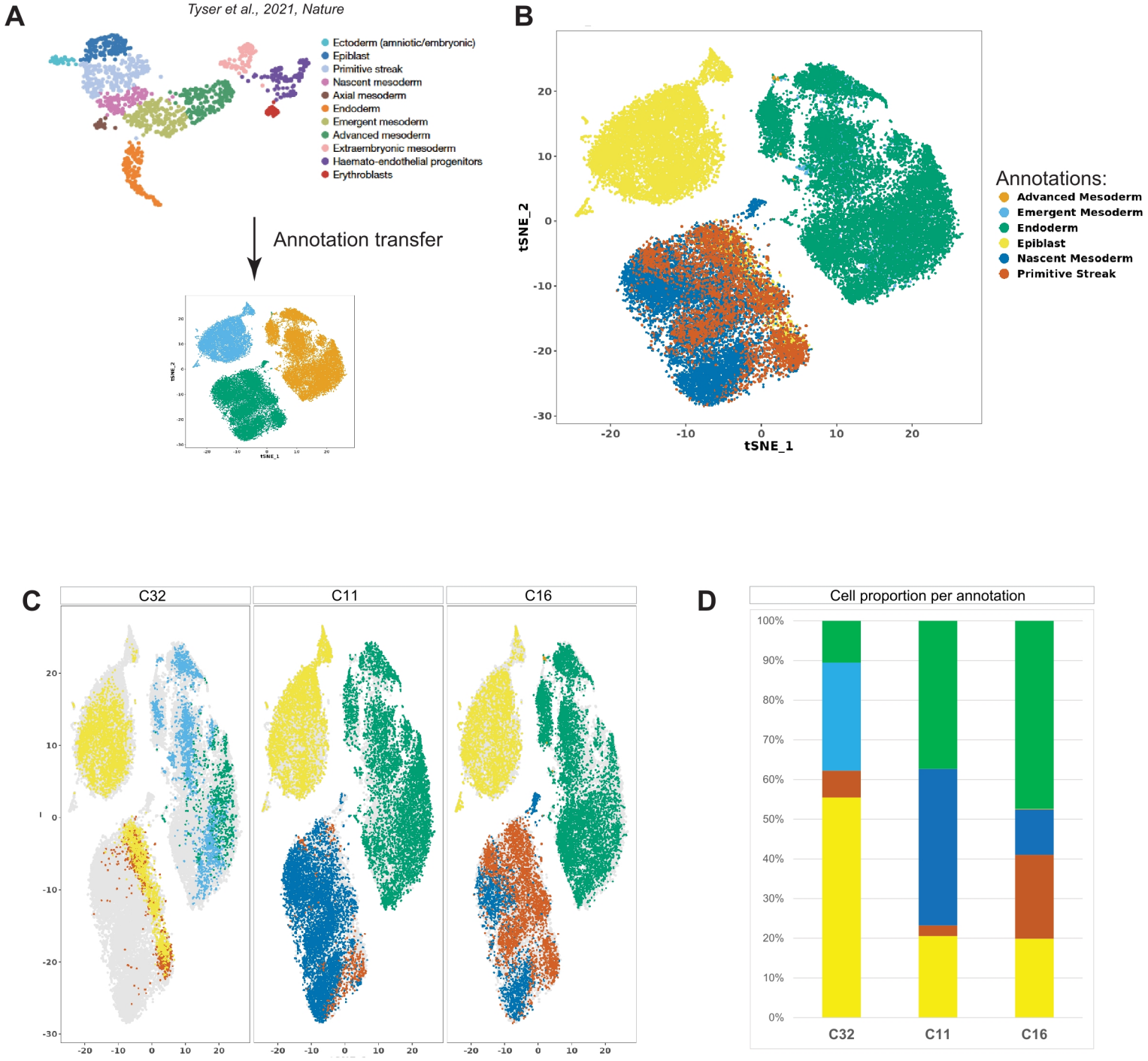
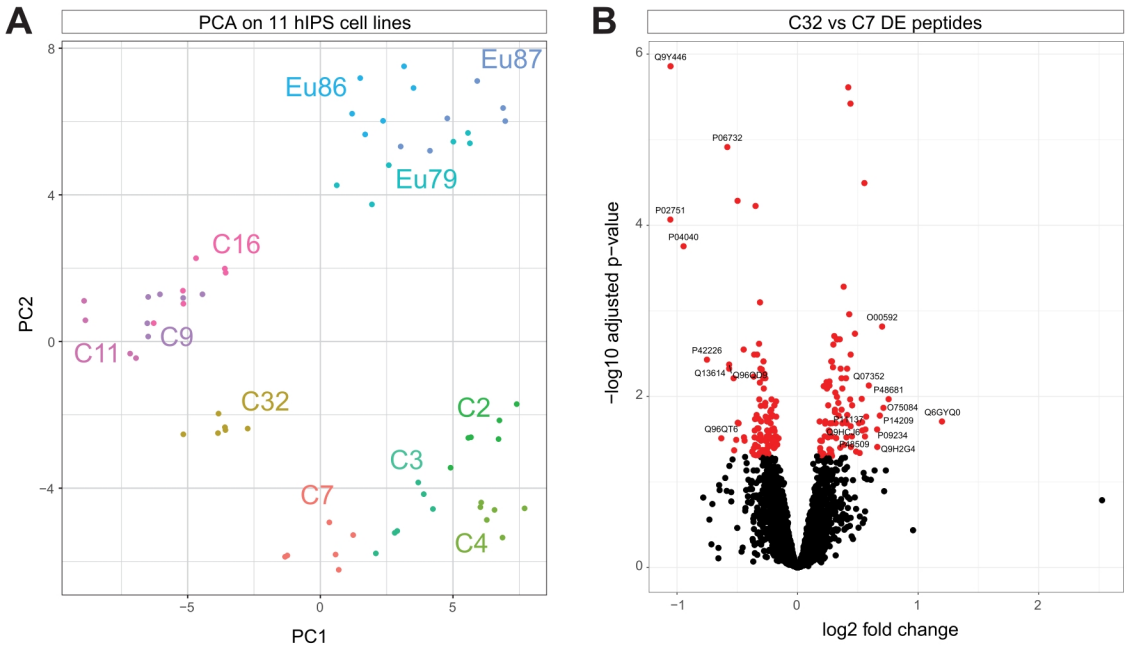

A
Tyser et al., 2021, Nature
Ectoderm (amniotic/embryonic)
Epiblast
Primitive streak
Nascent mesoderm
Axial mesoderm
Endoderm
Emergent mesoderm
Advanced mesoderm
Extraembryonic mesoderm
Haemato-endothelial progenitors
Erythroblasts
Annotation transfer
B
tSNE_2
tSNE_1
Annotations:
Advanced Mesoderm
Emergent Mesoderm
Endoderm
Epiblast
Nascent Mesoderm
Primitive Streak
C
C32
C11
C16
D
Cell proportion per annotation
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
C32
C11
C16

Figure S4



C

Idenfier	Name	logFC
Q6GYQ0	RALGAPA1, GARNL1, KIAA0884, TULIP1	1.19771
P48681	NES, Nbla00170	0.75524
O75084	FZD7	0.712103
O00592	PODXL, PCLP, PCLP1	0.701107
P14209	CD99, MIC2, MIC2X, MIC2Y	0.682289
Q9H2G4	TSPYL2, CDA1, DENTT, TSPX, HRIHFB2216	0.66118
P09234	SNRPC	0.659824
Q07352	ZFP36L1, BERG36, BRF1, ERF1, RNF162B, TIS11B	0.590807
P11137	MAP2	0.569085
Q9HCJ6	VAT1L, KIAA1576	0.562332
P48509	CD151, TSPAN24	0.55891
Q96QD9	FYTTD1, UIF	-0.56781
Q13614	MTMR2, KIAA1073	-0.56782
P06732	CKM, CKMM	-0.58283
Q96QT6	PHF12, KIAA1523	-0.6327
P42226	STAT6	-0.75175
P04040	CAT	-0.94585
Q9Y446	PKP3	-1.05386
P02751	FN1, FN	-1.05565

Figure S5

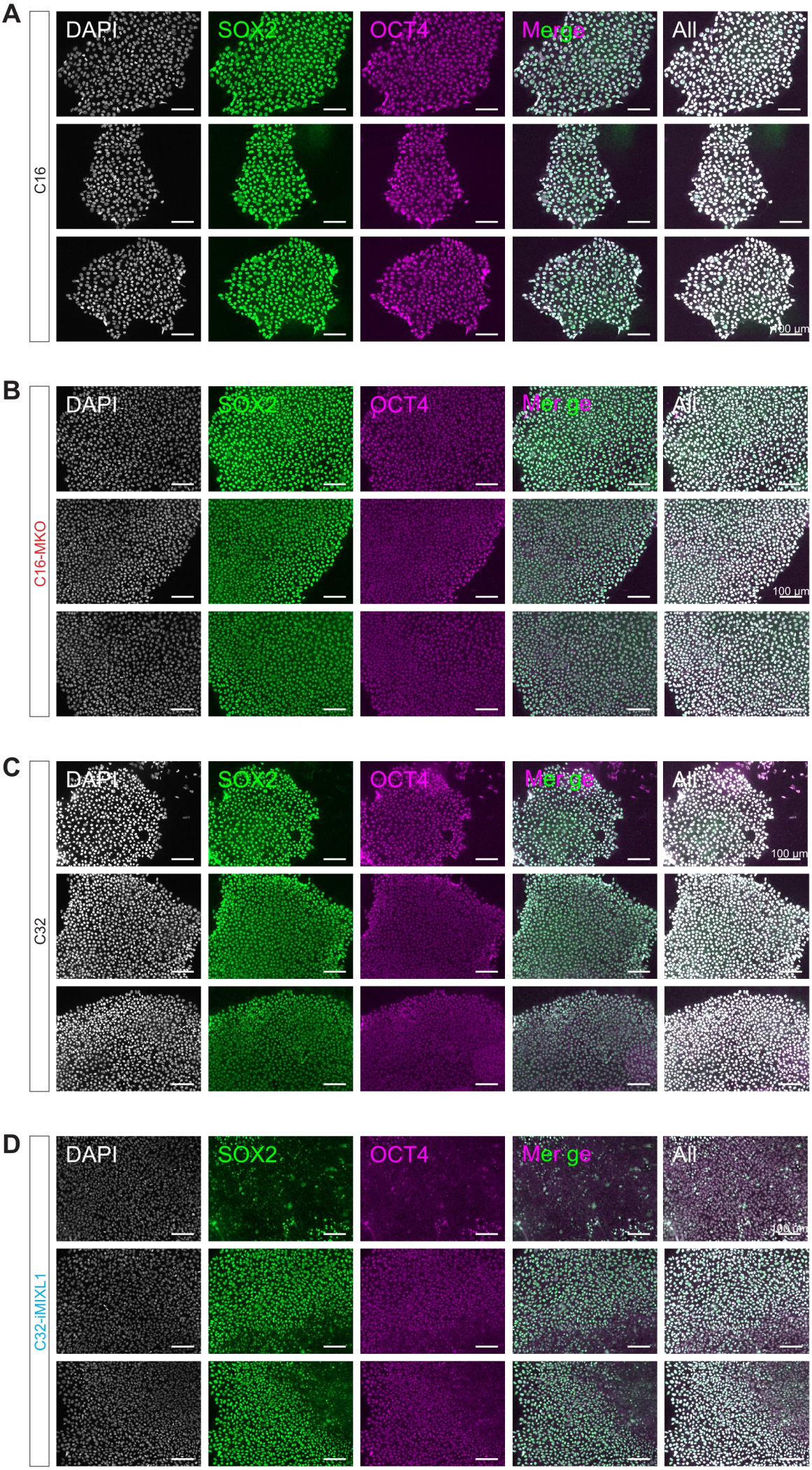
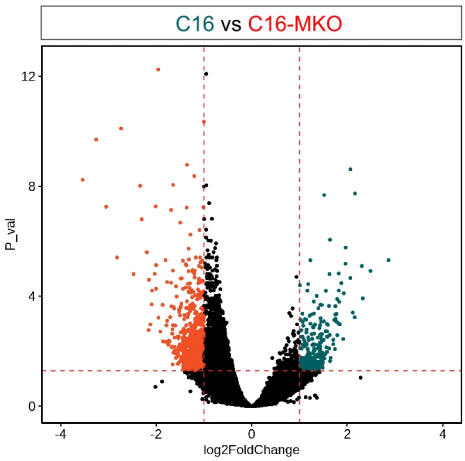
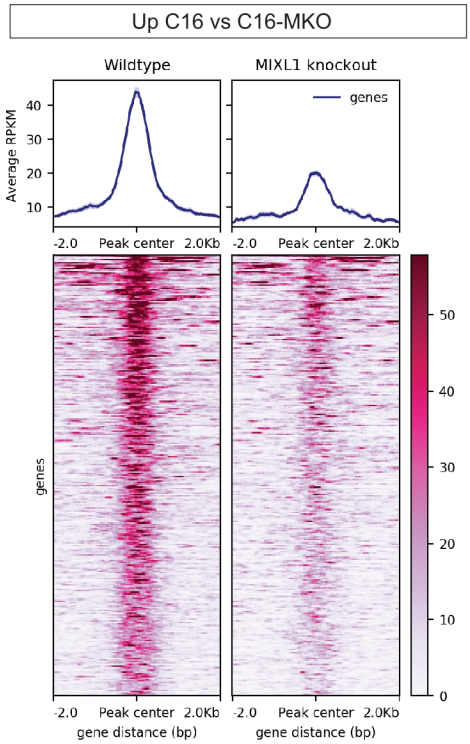

A
C16
DAPI
SOX2
OCT4
Merge
All
100 µm
B
C16-MKO
DAPI
SOX2
OCT4
Merge
All
100 µm
C
C32
DAPI
SOX2
OCT4
Merge
All
100 µm
D
C32-iMIXL1
DAPI
SOX2
OCT4
Merge
All
100 µm

Figure S6

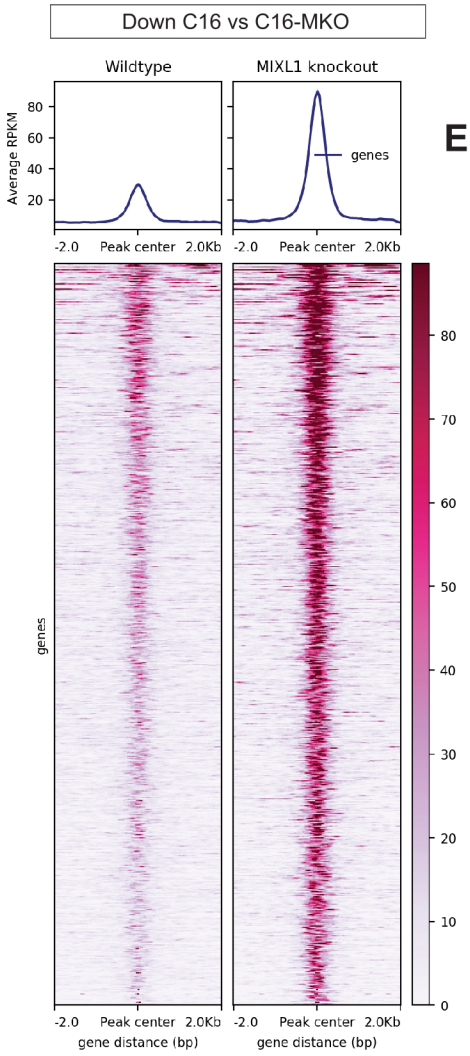
A



B



C



E

Known factor motifs

Motif	Factor	*P*-value
	TEAD1	10^{-9}
	FOXH1	10^{-8}
	JUN-AP1	10^{-8}
	BATF	10^{-7}
	CTCF	10^{-7}

De novo motifs

Motif	Best match	*P*-value
	FOXH1	10^{-18}
	CTCF	10^{-15}
	FOSL2	10^{-13}
	SPIB	10^{-13}
	TCF3	10^{-13}

t

D

Known factor motifs

Motif	Factor	*P*-value
	PROP1	10^{-13}
	PHOX2A	10^{-13}
	PAX7	10^{-8}
	LHX9	10^{-7}
	LHX2	10^{-7}

De novo motifs

Motif	Best match	*P*-value
	PHOX2B	10^{-21}
	SOX15	10^{-10}
	E2A	10^{-10}
	SPDEF	10^{-9}
	RBPJ	10^{-8}

F

```
1   AGCTTTGATG AGGACAGACG GGACTAGGGC AAGGCTCAAG   gRNA-1
41  GCCCAGGCCA TGTAAGGCAC AGGCCAGGCA AGGTGCTTCC   gRNA-2
81  GCCGCTCTTA ACGGGGATTT GACCCGGAGA AGAGAGTTCT   gRNA-3
121 GTCGCCTGCG GCCCGCGCAC CCGGGCCCCT CCGGGCTGCC   gRNA-4
161 CCTCCCAGGC AAATAGTCCT TCGGGGATGT GGATTGCGCC   gRNA-5
181 GTCCGGGCGG GTGGGCGGGG AAGAACACGG AGGGGGCCGG   gRNA-6
201 GACCCAAGTT CACGCCTCCC TCCTGCATCC CGCAGCCGGC   gRNA-7
241 GGAAGCGATT ATTCCCCGGC GTCTGGCGGG GCCGGGGGCG   Mixl1 TSS
281 GGCTCGCACC CGGAGGAGAC ACGGGTCTCC CCGAGCCCAC
321 CTTTGATCGC ACCGCCCCGC CTTCCCAGCC CCTGGCCCGG
361 GAGGGTATAA GTGCGGCCCG CGCC
```

G

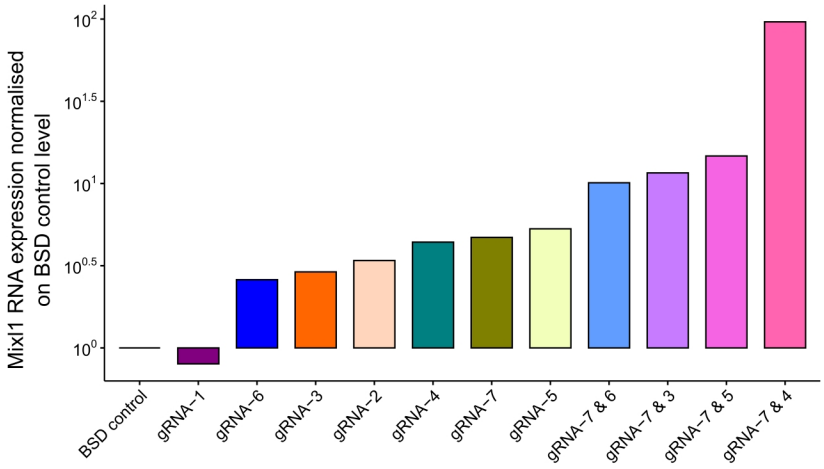


Figure S7

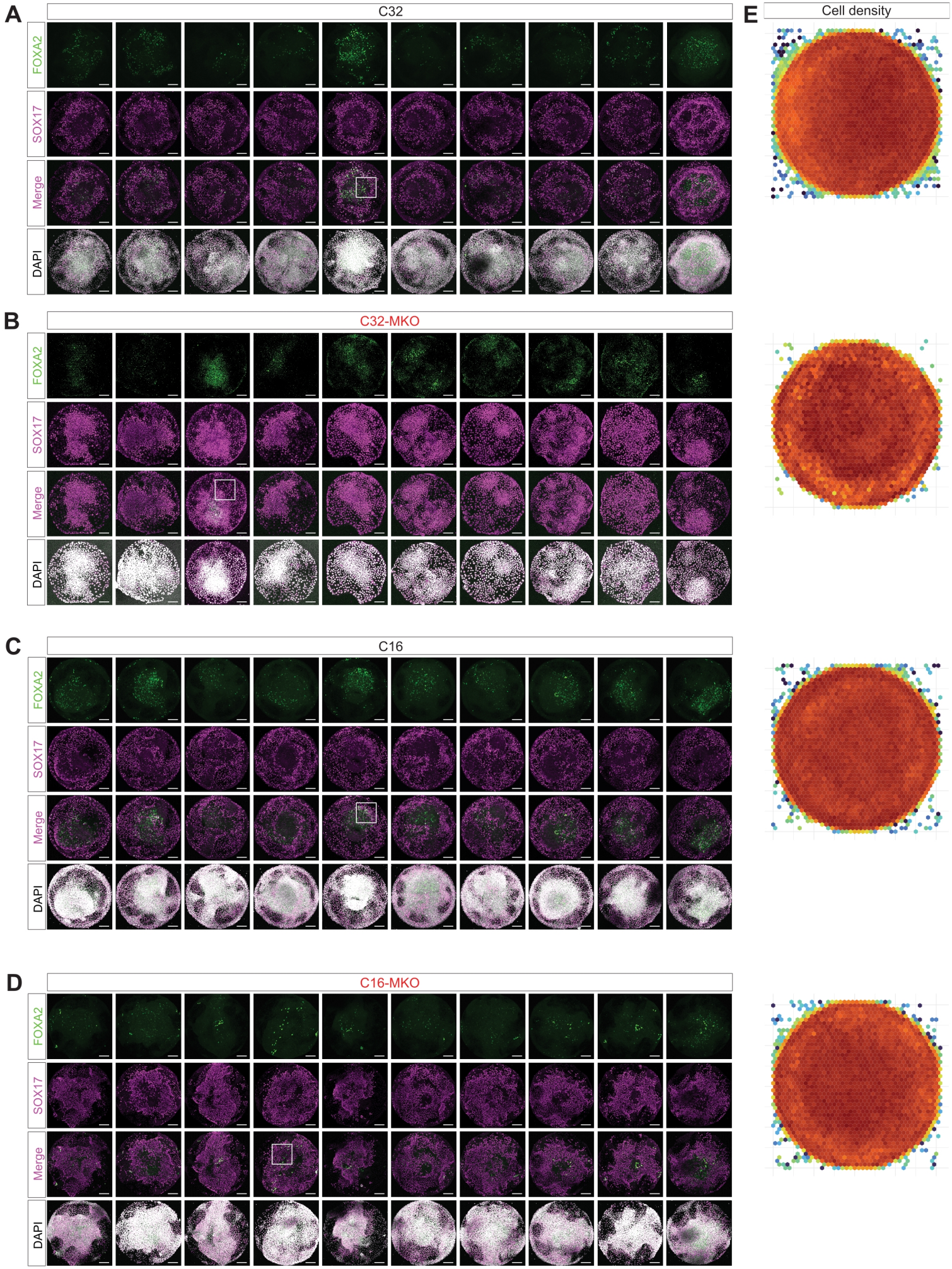
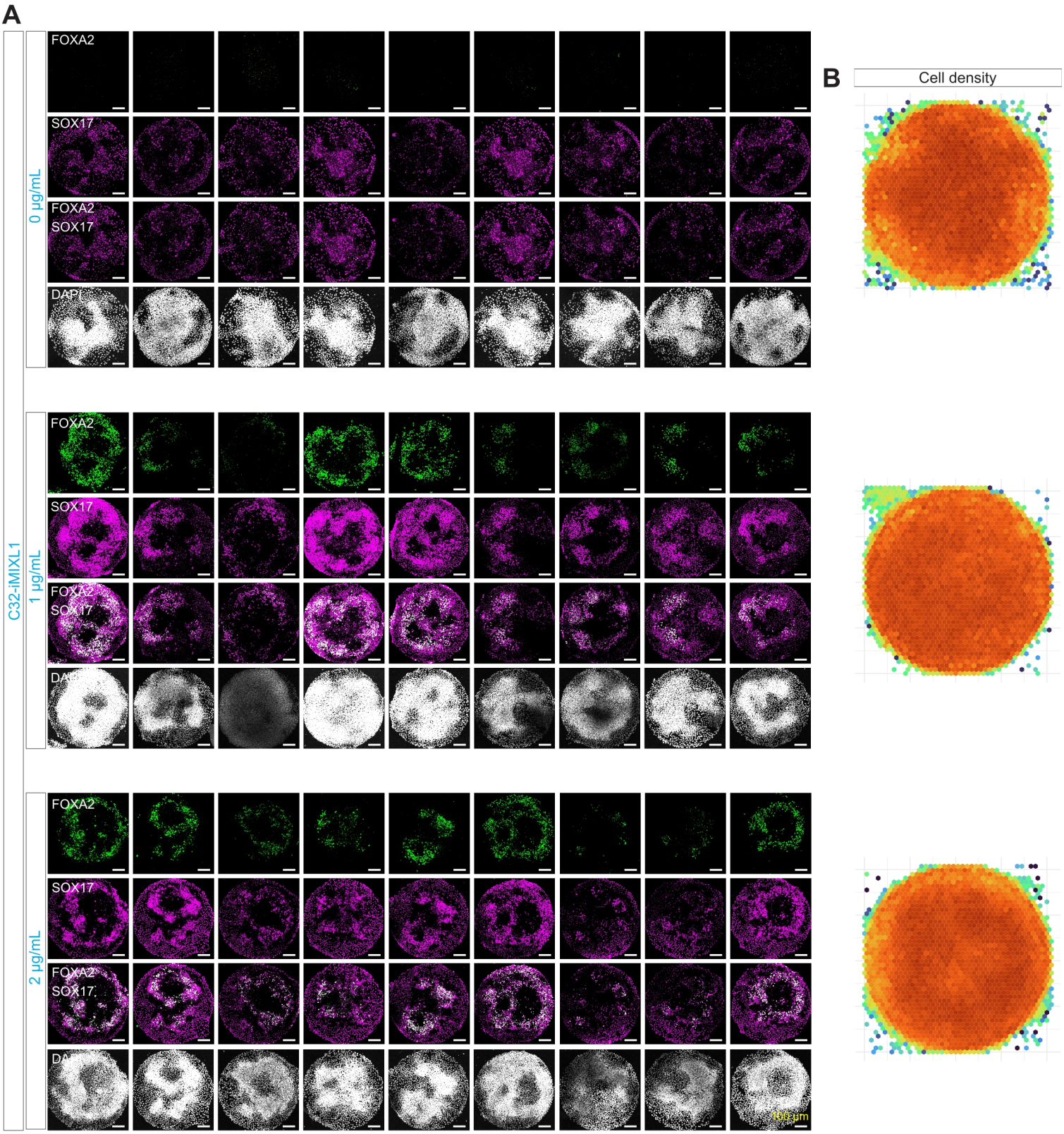

A
C32
FOXA2
SOX17
Merge
DAPI
B
C32-MKO
C
C16
D
C16-MKO
E
Cell density

Figure S8


A
B
Cell density
C32-iMIXL1
0 µg/mL
1 µg/mL
2 µg/mL
FOXA2
SOX17
FOXA2
SOX17
DAPI
1000 µm