Supplementary Information Contents

Methods

Collection of Human Gastrula Cells

The CS7 embryo was provided by the Human Developmental Biology Resource (HDBR - https://www.hdbr.org/general-information). HDBR has approval from the UK National Research Ethics Service (London Fulham Research Ethics Committee (18/LO/0822) and the Newcastle and North Tyneside NHS Health Authority Joint Ethics Committee (08/H0906/21+5)) to function as a Research Tissue Bank for registered projects. The HDBR is monitored by The Human Tissue Authority (HTA) for compliance with the Human Tissue Act (HTA; 2004). This work was done as part of project #200295 registered with the HDBR. The material was collected after appropriate informed written consent from the donor by medical termination. The sample was collected and transported in cold L15 media. It was then transferred to M2 media and imaged on a Leica Stereo microscope. The sample was micro-dissected using tungsten needles and dissociated into single cells using 200µl Accutase (ThermoFisher, Cat No. A1110501) for 12 minutes at 37°C, being agitated every 2 minutes, before adding 200µl heat-inactivated FBS (ThermoFisher, Cat No. 10500) to quench the reaction. Cells were then centrifuged at 1000rpm for 3 minutes at 4°C before being suspended in 100µl HBSS (ThermoFisher, Cat No. 14025) + 1% FBS, and stored on ice. Single cells were collected using a Sony SH800 FACS machine with a stringent single-cell collection protocol and sorted into 384 well plates containing SMART-seq2 lysis buffer [1] plus ERCC spike-ins (1:10M). To ensure we collected good quality cells, a live/dead dye (Abcam, Cat No. ab115347) was used; 100µl was added to the cell suspension at a 2x concentration in HBSS 10 minutes before collection, and live cells were collected based on their FITC intensity. Once cells were collected, plates were sealed, spun down, and frozen using dry ice before being stored at -80°C. This complete process, from dissection to single-cell collection, took approximately 2-3 hours. The embryo was karyotypically normal (Region specific assay: (13, 15, 16, 18, 21, 22) x 2, (X, Y) x 1).

Single-cell RNA sequencing

mRNA from single cells was isolated and amplified (21 PCR cycles) using the SMART-seq2 protocol[1]. Multiplexed sequencing libraries were generated from cDNA using the Illumina Nextera XT protocol and 125 bp paired-end sequencing was performed on an Illumina HiSeq 2500 instrument (V4 chemistry).

Raw data processing and normalization

In order to quantify the abundance of transcripts from 1,719 cells, salmon v0.17[2] was used. After indexing the human transcriptome (GRCh38.p13) in quasi-mapping-based mode, we quantified the transcripts with salmon using the --seqBias and --gcBias flags. We combined the transcript level abundances to the corresponding gene level counts, which were aggregated into a gene count matrix. Then, for downstream analyses, we only retained cells with more than 2,000 detected genes, with overall mapping rate greater than 55% and with relatively low mapping rate to mitochondrial genes (<0.02) and to ERCC spike-ins (<0.2). After this step, we obtained 1,195 good quality cells. The data were normalized using the 'quickcluster' and 'normalize' functions from scran package in R[3]. This was followed by pseudocount addition of 1 and natural-log transformation of the count matrix.

Clustering and cell type identification

To identify clusters of cells, we applied a graph-based algorithm. First, we selected the top 4,000 highly variable genes (HVGs) using the ‘high_variable_genes’ function from scanpy v1.4.4[4]. We constructed the cell-cell distance matrix as $\sqrt{(1-\rho)/2}$, where ρ is the Spearman’s correlation coefficient between cells. Next, a k-nearest neighbour graph was built with the first 30 principal components (PCs) and k=50. This was accomplished by the ‘neighbors’ function in scanpy, which computes the connectivity between cells based on UMAP[5]. To identify clusters, we applied the Leiden algorithm for community detection to the resulting graph (with a resolution of 0.75), as it has been shown to be a superior alternative to Louvain[6]. The same algorithm and resolution were used for subclustering the Endoderm, the Ectoderm and the Hemogenic Endothelial Progenitors clusters with top 2,000 HVGs in each. However, in this case the knn graph was built with the first 10 PCs and k=20. We visualized the resulting clusters in two dimensions by computing a UMAP representation with default parameters in scanpy (‘tl.umap’ function). To check the robustness of the clustering, we also computed the shared nearest neighbour (SNN)[7] graph and applied Leiden to it (resolution = 1.75), which produced very similar clusters (the adjusted mutual information score calculated with Python’s sklearn module was 0.8).

We identified marker genes for the clusters with the Wilcoxon rank-sum test in scanpy (‘rank_genes_groups’ function), by comparing the gene expression levels in a given cluster with the rest of the cells in the dataset. The genes were ranked according to their FDR, after p-values were corrected with the Benjamini-Hochberg method. We visualized the expression values of marker genes on a heatmap, after scaling the log-normalized counts between 0 and 1 by using ‘standard_scale=var’ option in scanpy heatmap plotting function sc.pl.heatmap.

Isoform analysis

We obtained isoform-level count matrix from Salmon, considering TPM-normalized counts and the ENSEMBL database (GRCh38.p13) for annotation. We compared transcript levels between pairs of clusters. First, we removed genes with more than 80% counts mapped to a single isoform. Then, for each gene, we built a contingency table including the average normalized levels of each isoform in the two clusters being compared. A chi-squared test was then used to check whether the isoform abundances differ between the two clusters of cells for a given gene, as in[8].

Trajectory analysis using diffusion pseudotime and RNA-velocity

For the whole embryo diffusion map, we built the k-nearest neighbor graph as described above (with k=50 and using the first 30 PCs) to find the connectivity kernel width. We then used the ‘diffmap’ function to build the diffusion map.

To estimate the trajectory of epiblast differentiation, we took 2,000 HVGs from epiblast, primitive streak (PS), ectoderm and nascent mesoderm clusters combined. Finally, the diffusion components were computed from the first 15 PCs with k=15.

To illustrate the estimated direction of differentiation of epiblast cells, we embedded the RNA velocities[9] of single cells on the above diffusion map. For this task, we aligned reads from each cell using STAR v2.7[10] to the human reference genome (GRCh38.p13), which was obtained from ENSEMBL. The aligned bam files were processed with velocyto v0.17.17[9] with the default 'run-smartseq2' mode, to create a count matrix made of spliced and unspliced read counts.

After filtering genes with less than 10 spliced and un-spliced counts from this matrix, we calculated the moments for velocity estimation by utilizing a built-in function from scVelo python module v01.20[11]. Subsequently, we inferred the splicing kinetic dynamics of the genes by applying the 'recover_dynamics' function. The velocity of each gene was estimated by solving splicing kinetics in the 'dynamical' mode with the 'velocity' function. Finally, we embedded the resulting velocities on the diffusion space calculated above by means of the 'velocity_embedding' function from the scVelo module. The diffusion map and the RNA velocities for the mesoderm specification analysis were computed in the same way.

We defined a diffusion pseudotime (dpt) coordinate on the diffusion map of epiblast differentiation in order to visualize gene expression trends. First, we fixed the cell with the highest value of the first diffusion component (DC1) as root, so that the middle point of pseudotime would fall roughly into epiblast. We fitted the expression levels of the genes as a function of the pseudotime with a generalized additive model, by utilizing the 'gam' package in R (v1.16.1). For visualisation purpose, we transformed the pseudotime values as (1-dpt), so that ectoderm cells would fall onto the left side, and primitive streak and nascent mesoderm on the right side of the pseudotime plot (Extended Data Figure 4a-c). Both fitted and unfitted values of the genes were scaled by dividing each by its maximum value of expression.

Human-Mouse EMT Comparison

For this analysis, we considered published single-cell RNA-seq data from mouse embryos during mid-streak stage (E7.25)[12], but we also checked that the results remain largely unaffected if data from E7.0 or E7.5 are used. Epiblast, primitive streak and nascent mesoderm clusters were selected from the human and the mouse datasets for downstream analysis, and they were analyzed separately as detailed below.

After constructing diffusion maps as described above with default parameters, we defined pseudotime starting from the cell with lowest DC1 value in both cases (Extended Data Figure 5a). After fitting gene expression values along pseudotime with generalized additive models (see above), we calculated the p-values using the ANOVA non-parametric test from the 'gam' R package and we then obtained the FDR values (Benjamini-Hochberg method). Genes with FDR<0.1 were clustered according to their expression pattern. This was achieved by hierarchical clustering with Spearman's correlation distance as described above ('hclust' function in R). For estimating the number of clusters, the dynamic hybrid cut method was used ('cutreeDynamic' function, in the package dynamicTreeCut, version 1.63, with 'deepslit'= 0 and 'minclustersize'= 50). In both human and mouse, we found three clusters of genes, two of which were characterized by a clear upward or downward average trend with an absolute log2-fold change greater than 1 between the fitted values at the end and at the beginning of the trajectory.

For the human-mouse comparison, we converted mouse genes to human equivalents (one-to-one homologous genes only) with the biomaRt R package[13]. We compared the trends of genes in human and mouse, and in particular we looked at genes coding for signalling molecules, as listed in the curated database of the CellPhoneDB package[14]. To visualize the trend of selected genes, we normalized the expression values by the maximum in both mouse and human. We set fitted values to zero for the genes that were expressed in less than 10 cells.

Mouse cluster comparison and blood staging analysis

We mapped the cells from the human gastrula against the mouse clusters at E7.25 available from[12]. To do that, we took the median levels of genes as a representation of the typical expression pattern of a given mouse cluster, and then, for each cell in the human gastrula, we used the "scmapCluster" function from the "scmap" R package (with 1,000 genes and similarity threshold parameter set to 0)[15] to identify the mouse cluster that was most similar to it. We performed the same procedure for human Endoderm (Figure 3c) and HEP (Extended Data Figure 9g) subclusters.

For staging analysis, we selected epiblast, primitive streak, endothelium, blood progenitors (1 and 2), and erythroid (1, 2 and 3) mouse clusters across the 9 stages, from E6.5 until E8.5. We merged the two blood progenitor clusters as well as 3 erythroid clusters and we obtained 4 mouse blood-related clusters that were used in downstream analyses. After verifying that the human blood-related clusters map onto the corresponding mouse clusters, we built a representative expression pattern for mouse for each cluster/stage, and calculated the median expression value of the genes per cluster/stage. Cells from human gastrula blood (Erythroblasts, Myeloid Progenitors, Endothelium, Blood Progenitors, EMPs), epiblast and primitive streak clusters were projected onto the corresponding mouse clusters (human Erythroblasts to mouse Erythroid; human Myeloid Progenitors, Blood Progenitors and EMPs to mouse Blood Progenitors; human Endothelium to mouse Endothelium; human PS to mouse PS; human epiblast to mouse epiblast) using scmap with the same parameters specified above.

Human and non-human primate (NHP) gastrulation comparison

We considered single-cell NHP gastrulation data[16] at 16 days post fertilization (d.p.f), since Primordial Germ Cells (PGCs) were only identified at that stage. Seurat integration method was applied to human and NHP single-cell data with 3,000 features used to find anchors (anchor.features parameter) and 70 neighbors to filter anchors (k.filter parameter). After obtaining the corrected expression values, we calculated the mean expression level of each gene per cluster. Finally, we performed hierarchical clustering with Spearman's correlation-based distance (see above) and the average aggregation method, using the linkage function from python's scipy module (v 1.5.2).

Primordial Germ Cell (PGC) identification and cross-species comparison

To single out the PGCs, we ran the RaceID algorithm ("RaceID" package v0.1.5)[17], which can identify rare cell types, on the cells in the primitive streak cluster. We used these parameter values: k=1, outlg=8 and probthr=0.005. This resulted in the identification of 9 subclusters of outlier cells. Among these, the PGCs were identified as the only cluster of outlier cells that had a median expression of PGC marker genes (NANOS3, SOX17, DND1, LAMA4, DPPA5) above 0.

To perform cross-species comparison of PGCs, we considered epiblast, primitive streak and PGC cells from human and mouse (E7.5 stage), and Late-epiblast (L-epi), late gastrulating cells 1 (L-gast1) and PGC clusters from non-human primate (16 d.p.f. stage) single-cell datasets. Z-scores were calculated for each gene per species by using rank_genes_groups function from scanpy with Wilcoxon-rank-sum test (method='wilcoxon') applied to PGC versus all others. The genes shown in the heatmaps of Figure 3d are selected from the top differentially expressed between PGC and the other clusters.

Cross-species signalling comparison

We obtained the gene sets for FGF, WNT and BMP signalling pathways from MSigDB database[18]. Here, we considered Epiblast, primitive streak and nascent mesoderm clusters from mouse and human gastrula data, and L-epi, L-gast1 and L-gast2 clusters from the non-human primate dataset. We computed the z-scores for each cluster per organism separately with Wilcoxon-rank-sum test as described above. The genes that were expressed in less than 10 cells across all clusters in a species were labelled as undetected (Extended Data Figure 7).

Cell cycle prediction

We estimated the cell cycle phase of each cell by applying the "pairs" algorithm described in[19]. A python implementation of this algorithm, 'pypairs' v3.1.1 was used in this analysis (https://pypairs.readthedocs.io/en/latest/documentation.html). After determining marker pairs from a training dataset[20] with the 'sandbag' function, we applied the function 'cyclone' to assign a cell cycle phase to each cell.

Indel analysis

Using our transcriptomic data, we estimated the sizes of genomic insertions and deletions (indels) in our data as well as in a dataset from human fetal liver cells[21]. This dataset was also processed with SMART-seq2 protocol and paired-end sequencing, although read lengths (75bp) were smaller than in our data (125bp). Hence, to minimize confounding effects in the results, we trimmed the reads in our data before processing it for this analysis. We aligned the data to the reference genome (GRCh38.p13), using bwa-mem v0.6[22] with default parameters. We then merged the aligned data from each single cell into one bam file and performed indel calling with a pipeline for insertion and deletion detection from RNA-seq data called 'transIndel' v0.1[23]. We kept the parameters at default values, except the minimum deletion length to be detected, which was set to 1 (-L flag set to 1).

Differential gene expression analysis between rostral and caudal mesoderm

We used the R packages DESeq2 v3.11[24] and Seurat v3.0[25] to identify the genes differentially expressed between rostral and caudal parts of the mesoderm cluster. After creating a Seurat object with the mesoderm cells, their anatomical and plate information, we converted it to DESeq2 object with “convertTo” function. We found differentially expressed genes (with FDR<0.1) between caudal and rostral parts of the mesoderm with “DESeqDataSet” and “DESeq” functions, while controlling for the plate effect.

Human embryonic stem cells comparison

For this comparison, we considered previously published single-cell RNA-seq data from pre-implantation human embryos[26] and from hESC[27]. In the pre-implantation embryo data, we removed cells from extra-embryonic tissues, from immunosurgery samples and with unannotated stage. Moreover, we only kept cells with a log10 total number of reads greater than 5.5. This resulted in 442 cells distributed between E3 and E7 stages.

In the hESC dataset, only cells in batch 1 (including both primed and naïve hESC) that passed the quality test performed in the original publication were taken.

These data from pre-implantation embryos and hESC were combined with the epiblast cells in our dataset, and count per million (CPM) normalization was performed. To assess the relationship between the datasets, we also used two different integration methods: Harmony[28](with the same HVGs and default parameters) and Seurat (using the same procedure as in the comparison with NHP data described above).

To compare changes in gene expression levels between the naïve and primed state in epiblast and in hESC, we took cells from E6 stage, given that they were closest to the naïve (see Extended Data Figure 3a and Figure 2a). Then, the log-fold changes of the previously identified HVGs (after removal of genes with less than mean log count of 1) were calculated between CS7 vs E6 cells and primed vs naïve hESC, after adding a pseudocount of 0.1 to the mean expression values. The line in Extended Data Figure 3b is obtained through a linear regression (LinearRegression function from sklearn python module).

Human gastruloid comparison

Recently published spatial transcriptomic data from human gastruloids were considered for the comparison[29]. Specifically, we took the z-scores of the genes that were found to be reproducible across the two replicates of the spatial transcriptomic experiment (Source Data Fig. 3c of [29]). For these genes, we calculated z-scores also in each cluster of our human gastrula data using rank_genes_groups function from scanpy with Wilcoxon-rank-sum test (method=’wilcoxon’) applied to cells in a given cluster versus all other cells.

Then, we compared the human gastrula with the gastruloid data by computing:

$$\rho_{ij} = corr(\boldsymbol{G}_i, \boldsymbol{S}_j)$$

i.e., the Pearson’s correlation coefficient between the z-scores of i-th gastrula cluster G_i and the z-scores of a gastruloid slice taken at the j-th position along the anterior/posterior axis S_j .

A null distribution for ρ_{ij} , $\mathcal{P}(\rho_{ij})$, was estimated by computing the Pearson's correlation coefficient after shuffling the z-scores of the gastruloid dataset across slices 500 times. We estimated a p-value p_{ij} as:

$$p_{ij} = \sum_{\rho^*{}_{ij} > \rho_{ij}} \mathcal{P}(\rho^*{}_{ij})$$

Maintenance and differentiation of hESC

Human ESCs (H9/WA09 line; WiCell) were cultured on plates coated with 10 µg/ml vitronectin (Stem Cells Technologies) in 37°C and 5% CO2. Pluripotent hESCs were plated as single cells at 4.0-5.0x10^4 cells/cm^2 using accutase (Gibco) and 10 µM Y27632 (Selleck), and maintained for two days in E6 media[30] supplemented with 2 ng/mL TGF-beta (bio-techne) and 25 ng/mL FGF2 (Dr. Marko Hyvönen, Cambridge University). These cells were sampled as "D0 PLU". Then, the cells were cultured for one day in CDM/PVA media[31], 1 mg/ml polyvinyl alcohol (Sigma) instead of BSA) with 100 ng/ml Activin A (Dr. Marko Hyvönen, Cambridge University), 80 ng/ml FGF2, 10 ng/ml BMP4 (bio-techne), 10 µM LY294002 (Promega) and 3 mM CHIR99021 (Tocris), and sampled as "D1 ME" or "D1 ME+PD". PD0325901 (Stem Cell Institute) was added at 1 µM. Bright field pictures were taken with Axiovert microscope (200M, Zeiss).

Immunocytochemistry

Cells plated on vitronectin-coated round coverslips (Scientific Laboratory Supplies) were washed once with PBS, and fixed with 4% paraformaldehyde (Alfa Aesar) in PBS at RT for 10 min. Following another PBS wash, cells were incubated with 0.25% Triton in PBS at 4°C for 15-20 min, 0.5% BSA (Sigma) in PBS at room temperature for 30 min, primary antibodies at 4°C overnight and secondary antibodies at room temperature for one hour. Anti-Ecadherin antibody (3195, Cell Signaling Technology, 1:200), and anti-Rabbit IgG Alexa Fluoro 568 antibody (A10042, Invitrogen, 1:1000) together with 10 µg/ml Hoechst33258 (B2883) were diluted in 0.5% BSA in PBS and each staining was followed by three washes with 0.5% BSA in PBS. Coverslips were preserved on slide glasses (Corning) with ProLong Gold Antifade Mountant (Life Technologies) and nail polisher, and observed with Zeiss inverted confocal system (LSM 710, Zeiss).

Quantitative RT-PCR for hESC samples

Total RNA was extracted from cells using the GenElute Mammalian Total RNA Miniprep Kit (Sigma-Aldrich) and the On-Column DNase I Digestion set (Sigma-Aldrich). Complementary DNA was synthesized from the RNA using random primers (Promega), dNTPs (Promega), RNAseOUT (Invitrogen) and SuperScript II (Invitrogen). Real-time PCR was performed with KAPA SYBR FAST qPCR Master Mix (Kapa Biosystems) on QuantStudio 12K Flex Real-Time PCR System machine (Thermo Fisher Scientific). Molecular grade water (Thermo Fisher Scientific) was used when necessary. Each gene expression level was normalized by the average expression level of *PBGD* and *RPLP0*. Primer sequences are shown in SI Table 15 and source data is provided in SI Table 17. Statistical analysis was performed using GraphPad Prism.

Mouse strains, husbandry and embryo collection

All animal experiments complied with the UK Animals (Scientific Procedures) Act 1986, approved by the local Biological Services Ethical Review Process and were performed under UK Home Office project licenses PPL 30/3420 and PCB8EF1B4. To obtain wild-type embryos, C57BL/6 males (in house) were crossed with 8-16 week old CD1 females (Charles River, England). All mice were maintained in a 12-hr light-dark cycle. Noon of the day when a vaginal plug was found was designated E0.5. To dissect the embryos, the pregnant females were culled by cervical dislocation in accordance with schedule one of the Animals (Scientific Procedures) Act. Embryos of the appropriate stage were dissected in M2 medium (Sigma-Aldrich, Cat No. M7167).

In Situ Hybridization Chain reaction (HCR)

In situ HCR kit (ver.3) containing amplifier set, hybridization, amplification, wash buffers, and DNA probe sets, were purchased from Molecular Instruments (molecularinstruments.org) and the protocol described in [32] was followed with slight modifications [33]. Probe libraries were designed and manufactured by Molecular Instruments using *Mus musculus* sequences from NCBI database. Following HCR embryos were then placed into 87% glycerol solution and imaged on a Zeiss 880 confocal microscope with a 40x oil (1.36 NA) objective. Images were captured at 512 × 512 pixel dimension using multiple tiles with a Z-step of 1.5 µm. Each HCR was repeated on at least 3 embryos.

Supplementary Notes

Supplementary Note 1 - Annotation of gastrula cell types

The Epiblast could be detected by the expression of *SOX2, OTX2, CDH1* and was represented in both the caudal and rostral regions of the embryo (55% caudal, 45% rostral, 0% yolk sac; Figure 1c, Extended Data Figure 2a and SI Table 2). In contrast, the Ectoderm (Amniotic/Embryonic) came predominantly from the rostral portion of the embryo and did not express pluripotency markers but was characterised by high expression of key markers such as *DLX5*, *TFAP2A* and *GATA3,* representative of both the extra-embryonic ectoderm of the amnion as well as the embryonic ectoderm at the rostral boundary of the neural plate[34,35] ((Figure 1d, and Extended Data Figure 2b).

The Primitive Streak was identified by the archetypal marker *TBXT* (*Brachyury*) in combination with *CDH1* and *FST* (Figure 1c and Extended Data Figure 2a). As expected, these cells originated almost exclusively from the caudal portion of the embryo (Figure 1d, and Extended Data Figure 2b). While the majority of Nascent Mesoderm cells were also located in the caudal region and expressed *TBXT*, they could be distinguished from PS cells by the expression of key mesodermal markers *MESP1 and PDGFRA* (Figure 1d, Extended Data Figure 2a and 10a). This co-expression of both PS and mesoderm markers led us to define this mesoderm as 'nascent', representing the forming mesoderm cells in the process of delaminating from the PS (Figure 1c). Axial Mesoderm could be detected by the expression of *TBXT*, *CHRD* and *NOTO* (Figure 1c, and Extended Data Figure 2a).

Two other clusters of embryonic mesoderm could be distinguished by their relative degree of maturation and location within the embryo. We annotated the first as Emergent Mesoderm since it expressed the highest levels of *MESP1* but was negative for *TBXT*, thereby representing a transition from the Nascent Mesoderm towards the more mature Advanced Mesoderm (Figure 1c, Extended Data Figure 2a and 10b). It also expressed *LHX1* and *OTX2* as well as the highest levels of *LEFTY2*, which in the mouse is expressed in mesoderm arising from the mid-distal region of the PS[36] (Extended Data Figure 10). The Advanced Mesoderm cluster was relatively more mature based on the decreased expression of *MESP1* and the highest expression of mesoderm markers *PDGFRA* and GATA6 (Extended Data Figure 2a and 10c). The Advanced Mesoderm also expressed *HAND1*, *BMP4*, *FOXF1* and *SNAI2,* all markers of relatively more mature mesoderm (Extended Data Figure 2a and 10c). The cells that we annotated as Nascent, Emergent and Advanced Mesoderm did not correspond directly to any of the established embryonic mesoderm sub-types such as paraxial or lateral plate mesoderm. Combinatorial marker signatures of such mesodermal sub-types spanned multiple clusters or were seen only within a subset of a cluster (Extended Data Figure 10a). For example, co-expression of *TBX6* and *MSGN1*, which marks presomitic mesoderm [37], was only detected in a subset of the Nascent Mesoderm cluster. In contrast, co-expression of *HAND1* and *GATA6*, which marks precardiac lateral-plate mesoderm, could be detected in multiple clusters including Advanced Mesoderm, Emergent Mesoderm, Extraembryonic Mesoderm as well as the Ectoderm (Amniotic/Embryonic). This suggests that at this stage, the embryonic mesodermal clusters identified do not represent specified mesodermal subtypes and instead correspond to transitional mesodermal states.

Extraembryonic Mesoderm was identified based both on its anatomical origin (69% yolk sac, 29% rostral, 2% caudal) as well as the expression of extraembryonic mesoderm markers such as *POSTN* and *ANXA1* (Extended Data Figure 2a, 2b and SI Table 16). In mice, *POSTN* is a marker of extraembryonic mesoderm overlying not only the amnion but also the yolk sac [38]. This cluster is likely to represent the yolk sac mesoderm given the spatial origin of the majority of cells. However, it may also include some amniotic mesoderm cells, given the commonalities in marker expression between these two cell types.

Consistent with the possibility that some cells of the Advanced Mesoderm might contribute later in development to the Extraembryonic Mesoderm, there was a degree of overlap in marker expression between the two (e.g. *HAND1* and *SNAI2*), and proximity between them on the diffusion map. However, we noted that while other mesoderm populations formed a well-connected trajectory, there was a clear separation between Advanced and Extraembryonic Mesoderm populations (Extended Data Figure 3d and SI Table 16), suggesting the Extraembryonic mesoderm may not be as closely related as other mesoderm populations. This might be because, unlike in rodents, in humans the extra-embryonic mesoderm is already present pre-gastrulation at CS 5[39,40]. This early extraembryonic mesoderm is thought to arise from the hypoblast or parietal endoderm of the bilaminar disk stage embryo[41]. The high DC1 value of Extraembryonic Mesoderm is consistent with the notion of an early origin, with cells arising both prior to and during gastrulation[42].

The information we retained on the spatial origin of these cells also allowed us to track the progression of mesoderm maturity (Extended Data Figure 3d). Cells that had more time (since they emerged from the PS) to mature might also, in that time, be expected to have migrated further from the PS. Consistent with this, Nascent Mesoderm was almost entirely collected from the caudal portion of the embryo that encompassed tissue immediately adjacent to the PS (99% caudal, 0% rostral, 1% yolk sac; Figure 1d, Extended Data Figure 2b and SI Table 2). Similarly, Axial Mesoderm was only located in the caudal region, consistent with it having just emerged from the PS when the embryo was collected. In contrast, Emergent Mesoderm and Advanced Mesoderm were collected from both the rostral and caudal regions of the embryo (Emergent: 70% caudal, 30% rostral; Advanced: 58% caudal, 42% rostral), highlighting that they had migrated rostrally away from the PS. These two mesoderm clusters also showed evidence of sub-structure based on Rostral-Caudal differences in origin (See below)

Other mesoderm-derived clusters included primitive erythroblasts, characterised by the expression of globin genes including the embryonic globins *HBZ* and *HBE1* as well as the eryhroid-related transcription factor *GATA1*. The majority of erythroblasts were collected from the yolk sac (81% yolk sac, 19% caudal, 0% rostral). We annotated a Hemato-Endothelial Progenitor population based on the expression of both endothelial makers (*PECAM1* and *MEF2C*) as well as hematopoietic markers (*RUNX1* and *GATA1*). Hemato-Endothelial Progenitor cells were also located predominantly in the yolk sac, although some cells also came from caudal and rostral regions (72% yolk sac, 15% caudal, 12% rostral).

Both blood-related populations expressed erythroid marker genes[43]; however the HEP population had a mixed expression profile of endothelial, myeloid and erythroid markers, suggesting a higher order substructure (Extended Data Figure 9b and c). Unsupervised clustering of the HEP population revealed four different subpopulations with distinct transcriptional and isoform signatures (Figure 4d, Extended Data Figure 9c and d). One subpopulation represented Endothelium (Endoth) based on the high expression of *PECAM1*, *CDH5*, *KDR* and *TEK*. Another expressed both megakaryocyte (*GP1BB*, *ITGA2B* (CD41), *NFE2*) and erythroid (*GATA1*, *KLF1*, *GYPB*, *HBE1*) markers, which we annotated as Megakaryocyte-Erythroid Progenitors (MEP). A Myeloid Progenitor sub-population could be identified on the basis of high levels of monocyte/macrophage markers *CD36*, *CSF1R*, and *LYVE1.* The final subcluster had an unusual transcriptional profile given the early stage of the sample, expressing a range of myeloid and erythroid markers including *KIT*, *CSF1R*, *MYB*, *SPI1* (PU.1), *CD34*, *PTPRC* (CD45), CD52 and *NFE2*[44]. Based on these markers and the expression of *MYB,* which in the mouse marks Erythro-Myeloid Progenitors (EMP), we annotated this cluster as EMP. Supporting this annotation is their co-expression of *CD34*, *CD45* and *CD44,* which have recently been used to define a yolk sac-derived myeloid-biased progenitor in CS11 human embryos[45]. These CS11 cells have been shown to have multi-lineage potential and are thought to correspond with murine EMPs. Our results therefore indicate that such progenitors might already start to emerge as early as CS7.

Endoderm could be identified by the expression of *SOX17*, *GATA6*, *FOXA2* and *TTR*. Endoderm was collected from all three anatomical regions (64% rostral, 19% yolk sac, 17% caudal; Extended Data Figure 8b and e). YS endoderm was identified based on spatial location (47% yolk-sac, 45% rostral, 8% caudal) and expression of the established marker genes *AFP* and *TTR* (Extended Data Figure 8e). In addition, we identified *GJB1* as a new marker of the Yolk Sac Endoderm and validated this by Hybridization Chain Reaction in the mouse (Extended Data Figure 8e and h).

The Smart-Seq2 protocol also allowed us to differentiate between transcript isoforms (SI Table 3). An analysis of this was able to detect the cluster specific expression of particular isoforms of genes such as *MEST* and *GCNT2* (Extended Data Figure 2e). *GCNT2* is known to be differentially spliced during mouse embryonic development, validating this approach [46]. We observed that *GCNT2* isoform *201* was expressed in most clusters except Axial mesoderm and Erythroblasts, *202* was more epiblast and PS specific, whilst *225* was expressed most strongly in the Advanced Mesoderm and Endoderm populations. This approach also identified *MEST* as a gene which had cell type specific isoform variation. Whilst MEST 202 was expressed in all clusters, 201 expression was low in the Endoderm and Erythroblasts clusters and 203 was more strongly expressed in the Nascent Mesoderm and Extraembryonic Mesoderm. This analysis further validates our clustering and highlights the depth of this data set.

Supplementary Note 2 - Comparison with Gastruloids

A recent advance in developing *in vitro* models of human gastrulation are hESC derived gastruloids[29]. We used the human gastrula to benchmark the rostral-caudal patterning of gastruloids, by comparing the transcriptional patterns of sections collected along the rostral-caudal axis of human gastruloids[29] with the cell type and the spatial information from our human gastrula dataset (Extended Data Figure 3c). This verified an overall agreement between the rostral-caudal gene expression patterns in gastruloids compared to the gastrula. In future, once single-cell data from gastruloids become available, in order to further refine protocols for producing gastruloids, more detailed analyses can be performed, for example, by comparing cell types produce *in vitro* with their natural counterparts (see Methods and Extended Data Figure 3a-c).

Supplementary Note 3 - Rostral and Caudal differences in diversification of mesodermal subtypes

The earliest mesoderm to emerge from the PS gives rise to the *extra-embryonic mesoderm*, that form the early blood islands[41,47]. Later cells to stream through the PS remain embryonic and give rise to the: *lateral plate mesoderm* (LPM), *intermediate mesoderm*, *paraxial mesoderm*, and *axial mesoderm*. On the basis of just marker gene expression, the mesodermal clusters at this stage did not correspond specifically to any of these established mesodermal sub-types, but rather had mixed expression of markers. This suggests that the clusters capture transient *states* of maturation of mesoderm as they transitioning towards more clearly specified sub-types.

To investigate the differentiation trajectory of human mesoderm, we took all mesoderm-related clusters together with the PS and performed a diffusion map and RNA-velocity analysis (Extended Data Figure 10b). Diffusion component 1 captured the time-course of mesoderm formation from the Primitive Streak through Nascent to Extra-embryonic Mesoderm. This trajectory was further highlighted by the RNA-velocity analysis, which showed vectors leading from PS through the Nascent Mesoderm towards the more mature mesoderm types. This computational analysis was supported by marker expression and the anatomical location from which cells were collected (Extended Data Figure 10b). The differentiation trajectory could also be observed by the transition along DC1 in the expression of transcription factors *TBXT*, *MESP1* and *HAND1* (Extended Data Figure 10c) that are respectively markers of the PS, early mesoderm and later relatively more mature mesoderm[48–50]. These results suggest that the pattern of emergence and maturation of mesoderm in the human is similar to that in chick and mouse.

Advanced Mesoderm predominantly represented LPM based on the expression of markers such as *HAND1*, *BMP4, GATA6* and *PDGFRA* (Extended Data Figure 10b). Within the Advanced Mesoderm cluster we could observe a separation in cells collected from rostral and caudal regions in both the UMAP and the diffusion map (Extended Data Figure 10b). To investigate this difference, we analyzed the differentially expressed genes between Advanced Mesoderm cells derived from these two regions (Extended Data Figure 10d). Rostral cells of the Advanced Mesoderm consistently showed elevated levels of expression for cardiomyocyte related contractile genes such as *TNNT2*, *MYL7*, *TNNI1* and *MYH10* (Extended Data Figure 10d and f). Consistent with this region of the LPM giving rise to the cardiac crescent.

However, several canonical markers of cardiac progenitors in the mouse were expressed at surprisingly low levels, if at all. The archetypal cardiac progenitor marker *NKX2-5* was expressed in very few cells throughout the mesoderm clusters, with no difference in caudal or rostral expression [51](Extended Data Figure 10d and e). *TBX5* was also expressed in very few cells, but in contrast to *NKX2-5*, expression was restricted specifically to rostral cells. *ISL1* was more strongly expressed but had a relatively broad distribution. In contrast *HOPX*, a homeodomain protein required for cardiac development in Zebrafish and Mouse, acting downstream of *NKX2-5* [52] was strongly and specifically expressed in caudal Advanced Mesoderm cells (Extended Data Figure 10d and e), despite the absence of *NKX2-5*. The cardiac-related marker *MAB21L2* [33,53], was strongly expressed in the Advanced Mesoderm, particularly in cells collected from the rostral region (Extended Data Figure 10d and e). Together these data suggest that in the human embryo, early cardiac contractile gene expression is initiated independently of *NKX2-5* and that other factors may play an important role in generating the cardiac progenitor state.

In the caudal cells of the Advanced Mesoderm cluster, there was an increase in PS markers such as *HOXA1* and *CDX1* consistent with their location (Extended Data Figure 10d and g). This region also strongly expressed *CDX2*, a marker of the early allantoic bud [54]. *CDH1*, required for placenta formation, is also expressed at elevated levels in the caudal cells. The combined expression of *CDX2* and *CDH1* suggests the caudal population of Advanced Mesoderm represents mesoderm emerging from the PS, to form allantoic mesoderm, consistent with the presence of an allantoic bud structure (Extended Data Figure 10d and g).

We could also detect the earliest signatures of other mesodermal subtypes. Chordin (*CHRD*) and *NOTO*, which are strongly expressed in the node and notochord could be detected in the axial mesoderm cluster, (Extended Data Figure 10h). The Emergent Mesoderm cluster had a broad and heterogenous marker gene profile, expressing genes reported in a number of different mesoderm types including, paraxial, intermediate, lateral plate mesoderm. The intermediate and axial midline marker, *LHX1*, was also expressed in this cluster along with other axial midline genes such as *OTX2* and *EOMES* [55] (Extended Data Figure 10j). The paraxial mesoderm marker *TBX6* and *SFRP2* was restricted to the Nascent Mesoderm and the caudal cells of the Emergent Mesoderm (Extended Data Figure 10i). The broad mesoderm signature of this cluster highlights its relatively immature state and suggests they emerged from the mid to anterior PS, supported by the specific expression of the anterior PS marker *LEFTY2* [36](Extended Data Figure 10j).

Supplementary Note 4 - Comparison of EMT pathway member expression during human and mouse gastrulation

During gastrulation, epithelial cells of the epiblast undergo an epithelial to mesenchymal transformation (EMT) by downregulating adherens junction molecules such as E-Cadherin (*CDH1*) so they can delaminate from the epiblast and migrate away as mesenchymal cells. Given this sample contains cells actively undergoing gastrulation, we sought to examine the transcriptional changes which occur during gastrulation in the human and compare them to the mouse, the leading model for studying mammalian gastrulation.

Pseudotime analysis showed that in the CS7 human gastrula, the PS marker *TBXT* (*Brachyury*) increased during the transition from Epiblast to Nascent Mesoderm, peaking in the Primitive Streak, while as expected, the mesoderm marker *MESP1* increased with the formation of Nascent Mesoderm (Figure 4c and Extended Data Figure 4b). A core event during EMT in mouse is a switch in the adherens junction molecules Cadherins, from E- to N-Cadherin (*CDH1* to *CDH2*)[56,57]. In the human, we could detect a similar trend in these genes, with CDH1 decreasing towards Nascent Mesoderm while *CDH2* increased (Extended Data 4b). In total, we found 3,350 genes that were differentially expressed along the developmental trajectory between Epiblast and Nascent Mesoderm (SI Table 5), of which 449 genes correlated with the expression profile of *TBXT* (FDR < 0.01; SI Table 6).

To test in an unbiased manner similarities and differences between human and mouse gastrulation, we used pseudotime analyses to compare the transition from epiblast to early mesoderm in the human cells with the equivalent populations in the Mouse Gastrula Single Cell Atlas[12] (Figure 2d and Extended Data Figure 5a). First, we compared the trends of genes that were differentially expressed in both human and mouse. For this analysis, we only considered genes with strong changes (i.e., with a log2-fold change > 1 along the trajectory, see Methods). By doing this, we identified 662 genes that were differentially expressed along the developmental trajectories from Epiblast to Nascent Mesoderm in both species (Extended

Data Figure 5b and SI Table 7). Of these genes, the vast majority (531) shared the same trend across pseudotime, either increasing (117) or decreasing (414). For example, in both mouse and human, *CDH1* decreased during transition from epiblast to nascent mesoderm, *TBXT* was transiently expressed and *SNAI1* continuously increased towards nascent mesoderm (Figure 2d and Extended Data Figure 5c). This verifies that the mouse represents a broadly faithful model of human gastrulation, despite the evolutionary distance between the two and the differences in gastrula morphology (human trilaminar disc compared to rodent specific egg cylinder).

Additionally, we also found some genes that were differentially expressed in only one of the two species. One example was in the expression of the zinc-finger transcription factor *SNAI2* (Slug), a regulator of EMT. In the human, *SNAI2* levels increased dramatically during Nascent Mesoderm formation, however *Snai2* was not detected during this transition in the mouse or cynomolgus macaque (Figure 2d). The absence of *Snai2* transcript in mouse during this transition was confirmed by additional independent mouse transcriptomic datasets from the same stages[58]. By contrast in the chick, as in the human, *SNAI2* is expressed within the PS and interfering in its expression results in the impaired emergence of mesoderm from the PS[59]. Together this suggests that in contrast to the mouse, in human, *SNAI2* may play a role in regulating EMT during gastrulation.

Another difference was in the expression of the transcription factor *MSGN1*. In the mouse *Msgn1* is expressed only weakly, if at all, in the streak and is expressed most strongly in the paraxial mesoderm[12,58]. Consistent with this, *Msgn1* null mouse embryos do not show a gastrulation phenotype [37,60], but show defects in somitogenesis[61]. Similar to mouse, *Msgn1* was not detected during gastrulation in the cynomolgus macaque. In contrast, the human gastrula showed robust and widespread expression of *MSGN1* in Nascent Mesoderm (Extended Data Figure 5c) raising the possibility that it may be functional more broadly in the human than the mouse.

In the mouse, the expression of various signaling molecules is crucial for EMT, germ layer specification and migration[62–64]. Hence, we analyzed specifically the expression trends of signaling molecules across the human, mouse and cynomolgus monkey. Our analyses again revealed broad similarities, but we also did observe some striking differences (Extended Data Figure 7). In the mouse, *TDGF1*, a NODAL co-receptor essential for normal mesodermal patterning, shows an increase in expression during Primitive Streak and Nascent Mesoderm formation. In contrast, in the human gastrula *TDGF1* expression showed the opposite trend, decreasing as Nascent Mesoderm formed (Figure 2d). FGF8 is the only known FGF directly required for gastrulation[65], playing a particularly important role in the migration of cells away from the PS[66]. In contrast, FGF8 was completely absent during the transition from Epiblast to Nascent Mesoderm in human (Figure 2d). We noted however that other FGF members were expressed in the human during this transition, raising the possibility that they may be serving the function that FGF8 does in the mouse. For example, we observed expression of *FGF4* (which is also expressed in the mouse), and *FGF2*, which is not required for gastrulation in the mouse[67,68] nor expressed, as confirmed in other datasets[58](Figure 2d). Consistent with the notion of an overlap in FGF function during gastrulation, treatment of *in vitro* cultured mouse epiblast with FGF2 results in altering the fate of these cells from ectoderm to mesoderm[69].

To experimentally validate these human specific transcriptional trends, we used an *in vitro* model of the transition from Epiblast to Nascent Mesoderm. For this, we differentiated pluripotent human ESC (PLU) to mesendoderm progenitors (ME) (Figure 2e and Extended Data Figure 6)[70–72]. ME colonies showed hallmarks of the EMT accompanying gastrulation, such as dispersed morphology and downregulated E-Cadherin (Figure 2e and Extended Data

Figure 6a), in addition to the upregulation of key EMT and gastrulation markers (Figure 2e and Extended Data Figure 6b). Serving as a negative control, blockage of FGF/ERK signaling by a MEK inhibitor prevented all of these responses and directed ME+PD cells towards a non-neural ectoderm state (Extended Data Figure 6c). These results indicated that our hESC model recapitulates the transition from Epiblast to Nascent Mesoderm in the diffusion map (Figure 2c).

Using this *in vitro* system, we tested the gene expression changes identified in the human gastrula. Consistent with our earlier findings, *SNAI2* and *MSGN1* increased, while *FGF2* and *TDGF1* decreased significantly during the transition from PLU to ME states. *FGF8* was expressed at very low level throughout the differentiation (Figure 2e and Extended Data Figure 6b). Upon MEK inhibition, all these markers were maintained at levels comparable to that in PLU cells, with the exception of *TDGF1*, possibly because it is regulated by multiple pathways. Together, these results indicate that there is a broad conservation of molecular players in human and mouse gastrulation, while the roles of specific members in these gene modules may vary between humans and mice.

Supplementary References

1. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* (2014). doi:10.1038/nprot.2014.006

2. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* (2017). doi:10.1038/nmeth.4197

3. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* (2016). doi:10.1186/s13059-016-0947-7

4. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* (2018). doi:10.1186/s13059-017-1382-0

5. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* (2018). doi:10.21105/joss.00861

6. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* (2019). doi:10.1038/s41598-019-41695-z

7. Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Trans. Comput.* (1973). doi:10.1109/T-C.1973.223640

8. Froussios, K., Mourão, K., Simpson, G., Barton, G. & Schurch, N. Relative abundance of transcripts (RATs): Identifying differential isoform abundance from RNA-seq [version 1; referees: 1 approved, 2 approved with reservations]. *F1000Research* (2019). doi:10.12688/f1000research.17916.1

9. La Manno, G. *et al.* RNA velocity of single cells. *Nature* (2018). doi:10.1038/s41586-018-0414-6

10. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013). doi:10.1093/bioinformatics/bts635

11. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *bioRxiv* (2019). doi:10.1101/820936

12. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).

13. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat. Protoc.* (2009). doi:10.1038/nprot.2009.97

14. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* (2020). doi:10.1038/s41596-020-0292-x

15. Kiselev, V. Y., Yiu, A. & Hemberg, M. Scmap: Projection of single-cell RNA-seq data across data sets. *Nat. Methods* (2018). doi:10.1038/nmeth.4644

16. Ma, H. *et al.* In vitro culture of cynomolgus monkey embryos beyond early gastrulation. *Science (80-.).* (2019). doi:10.1126/science.aax7890

17. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* (2015). doi:10.1038/nature14966

18. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* (2005). doi:10.1073/pnas.0506580102

19. Scialdone, A. *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* (2015). doi:10.1016/j.ymeth.2015.06.021

20. Leng, N. *et al.* Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* (2015). doi:10.1038/nmeth.3549

21. Segal, J. M. *et al.* Single cell analysis of human foetal liver captures the transcriptional profile of hepatobiliary hybrid progenitors. *Nat. Commun.* (2019). doi:10.1038/s41467-019-11266-x

22. Li, H. [Heng Li - Compares BWA to other long read aligners like CUSHAW2] Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr. arXiv* (2013). doi:arXiv:1303.3997 [q-bio.GN]

23. Yang, R., Van Etten, J. L. & Dehm, S. M. Indel detection from DNA and RNA sequencing data with transIndel. *BMC Genomics* (2018). doi:10.1186/s12864-018-4671-4

24. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* (2014). doi:10.1186/s13059-014-0550-8

25. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* (2019). doi:10.1016/j.cell.2019.05.031

26. Petropoulos, S. *et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* (2016). doi:10.1016/j.cell.2016.03.023

27. Messmer, T. *et al.* Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution. *Cell Rep.* (2019). doi:10.1016/j.celrep.2018.12.099

28. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* (2019). doi:10.1038/s41592-019-0619-0

29. Moris, N. *et al.* An in vitro model of early anteroposterior organization during human development. *Nature* (2020). doi:10.1038/s41586-020-2383-9

30. Chen, G. *et al.* Chemically defined conditions for human iPSC derivation and culture. *Nat. Methods* (2011). doi:10.1038/nmeth.1593

31. Johansson, B. M. & Wiles, M. V. Evidence for involvement of activin A and bone morphogenetic protein 4 in mammalian mesoderm and hematopoietic development. *Mol. Cell. Biol.* (1995). doi:10.1128/mcb.15.1.141

32. Choi, H. M. T. *et al.* Third-generation in situ hybridization chain reaction: Multiplexed, quantitative, sensitive, versatile, robust. *Dev.* (2018). doi:10.1242/dev.165753

33. Tyser, R. C. V. *et al.* Characterization of a common progenitor pool of the epicardium and myocardium. *Science (80-.).* (2021). doi:10.1126/science.abb2986

34. Yang, L. *et al.* An early phase of embryonic Dlx5 expression defines the rostral boundary of the neural plate. *J. Neurosci.* (1998). doi:10.1523/JNEUROSCI.18-20-08322.1998

35. Streit, A. The preplacodal region: An ectodermal domain with multipotential progenitors that contribute to sense organs and cranial sensory ganglia. *International Journal of Developmental Biology* (2007). doi:10.1387/ijdb.072327as

36. Meno, C. *et al.* Mouse lefty2 and zebrafish antivin are feedback inhibitors of nodal signaling during vertebrate gastrulation. *Mol. Cell* (1999). doi:10.1016/S1097-2765(00)80331-7

37. Nowotschin, S., Ferrer-Vaquer, A., Concepcion, D., Papaioannou, V. E. & Hadjantonakis, A. K. Interaction of Wnt3a, Msgn1 and Tbx6 in neural versus paraxial mesoderm lineage commitment and paraxial mesoderm differentiation in the mouse embryo. *Dev. Biol.* (2012). doi:10.1016/j.ydbio.2012.04.012

38. Dobreva, M. P. *et al.* Periostin as a biomarker of the amniotic membrane. *Stem Cells Int.* (2012). doi:10.1155/2012/987185

39. O'Rahilly, R. & Müller, F. Developmental Stages in Human Embryos. *Contrib. Embryol., Carnegie Inst. Wash* **637**, (1987).

40. Enders, A. C. & King, B. F. Formation and differentiation of extraembryonic mesoderm in the rhesus monkey. *Am. J. Anat.* (1988). doi:10.1002/aja.1001810402

41. Bianchi, D. W., Wilkins-Haug, L. E., Enders, A. C. & Hay, E. D. Origin of extraembryonic mesoderm in experimental animals: Relevance to chorionic mosaicism in humans. *Am. J. Med. Genet.* (1993). doi:10.1002/ajmg.1320460517

42. Ross, C. & Boroviak, T. E. Origin and function of the yolk sac in primate embryogenesis. *Nat. Commun.* (2020). doi:10.1038/s41467-020-17575-w

43. McGrath, K. E., Frame, J. M. & Palis, J. Early hematopoiesis and macrophage development. *Seminars in Immunology* (2015). doi:10.1016/j.smim.2016.03.013

44. Palis, J. Hematopoietic stem cell-independent hematopoiesis: emergence of erythroid, megakaryocyte, and myeloid potential in the mammalian embryo. *FEBS Letters* (2016). doi:10.1002/1873-3468.12459

45. Bian, Z. *et al.* Deciphering human macrophage development at single-cell resolution. *Nature* (2020). doi:10.1038/s41586-020-2316-7

46. Revil, T., Gaffney, D., Dias, C., Majewski, J. & Jerome-Majewska, L. A. Alternative splicing is frequent during early embryonic development in mouse. *BMC Genomics* (2010). doi:10.1186/1471-2164-11-399

47. Kinder, S. J. *et al.* The orderly allocation of mesodermal cells to the extraembryonic structures and the anteroposterior axis during gastrulation of the mouse embryo. *Development* (1999).

48. Wilkinson, D. G., Bhatt, S. & Herrmann, B. G. Expression pattern of the mouse T gene and its role in mesoderm formation. *Nature* **343**, 657–658 (1990).

49. Saga, Y. *et al.* MesP1: A novel basic helix-loop-helix protein expressed in the nascent mesodermal cells during mouse gastrulation. *Development* (1996).

50. Cserjesi, P., Brown, D., Lyons, G. E. & Olson, E. N. Expression of the Novel Basic Helix-Loop-Helix Gene eHAND in Neural Crest Derivatives and Extraembryonic Membranes during Mouse Development. *Dev. Biol.* (1995). doi:10.1006/dbio.1995.1245

51. Harvey, R. P. NK-2 homeobox genes and heart development. *Developmental Biology* (1996). doi:10.1006/dbio.1996.0212

52. Chen, F. *et al.* Hop is an unusual homeobox gene that modulates cardiac development. *Cell* (2002). doi:10.1016/S0092-8674(02)00932-7

53. Saito, Y., Kojima, T. & Takahashi, N. Mab21l2 is essential for embryonic heart and liver development. *PLoS One* (2012). doi:10.1371/journal.pone.0032991

54. Beck, F., Erler, T., Russell, A. & James, R. Expression of Cdx-2 in the mouse embryo and placenta: Possible role in patterning of the extra-embryonic membranes. *Dev. Dyn.* (1995). doi:10.1002/aja.1002040302

55. Costello, I. *et al.* Lhx1 functions together with Otx2, Foxa2, and Ldb1 to govern anterior mesendoderm, node, and midline development. *Genes Dev.* (2015). doi:10.1101/gad.268979.115

56. Smith, D. E., Franco Del Amo, F. & Gridley, T. Isolation of Sna, a mouse gene homologous to the Drosophila genes snail and escargot: Its expression pattern suggests multiple roles during postimplantation development. *Development* (1992).

57. Cano, A. *et al.* The transcription factor Snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. *Nat. Cell Biol.* (2000). doi:10.1038/35000025

58. Peng, G. *et al.* Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. *Dev. Cell* (2016). doi:10.1016/j.devcel.2016.02.020

59. Nieto, M. A., Sargent, M. G., Wilkinson, D. G. & Cooke, J. Control of cell behavior during vertebrate development by Slug, a zinc finger gene. *Science (80-.).* (1994). doi:10.1126/science.7513443

60. Jeong Kyo Yoon & Wold, B. The bHLH regulator pMesogenin1 is required for maturation and segmentation of paraxial mesoderm. *Genes Dev.* (2000). doi:10.1101/gad.850000

61. Chalamalasetty, R. B. *et al.* Mesogenin 1 is a master regulator of paraxial presomitic mesoderm differentiation. *Dev.* (2014). doi:10.1242/dev.110908

62. Ciruna, B. & Rossant, J. FGF Signaling Regulates Mesoderm Cell Fate Specification and Morphogenetic Movement at the Primitive Streak. *Dev. Cell* (2001). doi:10.1016/S1534-5807(01)00017-X

63. Ding, J. *et al.* Cripto is required for correct orientation of the anterior-posterior axis in the mouse embryo. *Nature* (1998). doi:10.1038/27215

64. Jin, J. Z. & Ding, J. Cripto is required for mesoderm and endoderm cell allocation

during mouse gastrulation. *Dev. Biol.* (2013). doi:10.1016/j.ydbio.2013.05.029

65. Ornitz, D. M. & Itoh, N. The fibroblast growth factor signaling pathway. *Wiley Interdiscip. Rev. Dev. Biol.* (2015). doi:10.1002/wdev.176

66. Sun, X., Meyers, E. N., Lewandoski, M. & Martin, G. R. Targeted disruption of Fgf8 causes failure of cell migration in the gastrulating mouse embryo. *Genes Dev.* (1999). doi:10.1101/gad.13.14.1834

67. Zhou, M. *et al.* Fibroblast growth factor 2 control of vascular tone. *Nat. Med.* (1998). doi:10.1038/nm0298-201

68. Ortega, S., Ittmann, M., Tsang, S. H., Ehrlich, M. & Basilico, C. Neuronal defects and delayed wound healing in mice lacking fibroblast growth factor 2. *Proc. Natl. Acad. Sci. U. S. A.* (1998). doi:10.1073/pnas.95.10.5672

69. Burdsal, C. A., Flannery, M. L. & Pedersen, R. A. FGF-2 alters the fate of mouse epiblast from ectoderm to mesoderm in vitro. *Dev. Biol.* (1998). doi:10.1006/dbio.1998.8898

70. Teo, A. K. K. *et al.* Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes Dev.* (2011). doi:10.1101/gad.607311

71. Mendjan, S. *et al.* NANOG and CDX2 pattern distinct subtypes of human mesoderm during exit from pluripotency. *Cell Stem Cell* (2014). doi:10.1016/j.stem.2014.06.006

72. Yiangou, L. *et al.* Method to Synchronize Cell Cycle of Human Pluripotent Stem Cells without Affecting Their Fundamental Characteristics. *Stem Cell Reports* (2019). doi:10.1016/j.stemcr.2018.11.020

Supplementary Table Legends

Supplementary Table 1 - Human Gastrula cluster marker genes

Table showing the marker genes ranked by statistical significance for the 11 different clusters identified, including: Epiblast (Epi), Ectoderm (Amniotic/Embryonic) (EAE), Primitive Streak (PS), Nascent Mesoderm (NM), Emergent Mesoderm (EM), Advanced Mesoderm (AM), Extraembryonic Mesoderm (ExM), Axial Mesoderm (AxM), Endoderm (Endo), Hemato-Endothelial Progenitors (HEP), Erythroblasts (Ery).

Supplementary Table 2 – Cell origin per cluster

Table showing the percentage of cells from a specific anatomical region for each cluster.

Supplementary Table 3 - Transcript isoform differences for all clusters comparisons

Tables showing transcript isoform comparisons between clusters. Each worksheet refers to the comparison of a single cluster with every other cluster of cell types, and includes the names of the genes whose isoforms are differentially expressed with the relative p-value. One-sided chi-square test. Extraembryonic Mesoderm (ExM), Hemato-Endothelial Progenitors (HEP), Ectoderm (Amniotic/Embryonic) (EAE).

Supplementary Table 4 - Top 30 genes with highest log-fold change in each quadrant of CS7 vs E6 embryos and naïve vs primed hESC correlation

Table showing the top 30 genes with the highest log2-fold changes in each quadrant of the CS7 vs E6 embryos and naïve vs primed hESC comparison (Figure 3b). The numeric values in the table represent log2-fold changes in shown comparisons.

Supplementary Table 5 - Differentially expressed genes along the Epiblast to Nascent mesoderm trajectory

List of differentially expressed genes and their trends during Epiblast to Nascent Mesoderm transition in the human gastrula. The trends are “up” or “down” when there is an increasing or decreasing trend with a log2-fold change greater than 1 between the expression values at the beginning and at the end of the trajectory; “flat” genes are those having a log-fold change less than 1 between initial and final expression value. False discovery rate, FDR.

Supplementary Table 6 - Genes correlating with TBXT along the Epiblast to Nascent mesoderm trajectory

List of genes which correlate with TBXT along the Epiblast to Nascent Mesoderm trajectory. Values represent correlation coefficient (coef), p-value (p-val) and false discovery rate (FDR). Two-sided Spearman’s rho test.

Supplementary Table 7 - DEGs in human and mouse during Epiblast to Nascent mesoderm differentiation

Comparison of differentially expressed genes (DEGs) in mouse and human along the Epiblast to Nascent Mesoderm differentiation trajectory. Genes are marked by whether they are differentially expressed in mouse (DE Mouse) or human (DE Human), their expression trends in mouse and human and false discovery rate (FDR) in these species. The trends are "up" or "down" when there is an increasing or decreasing trend with a log2-fold change greater than 1 between the expression values at the beginning and at the end of the trajectory; "flat" genes are those having a log-fold change less than 1 between initial and final expression value.

Supplementary Table 8 - Ectoderm (Amniotic/Embryonic) subcluster genes

Table of the top 50 marker genes for the Ectoderm (Amniotic/Embryonic) subclusters (Amnion and Non-Neural Ectoderm (NNE)).

Supplementary Table 9 - Primordial Germ Cell Primitive Streak DEGs

List of top differentially expressed genes (DEGs) between Primordial Germ cells (PGC) and Primitive Streak with false discovery rate (FDR). Two-sided Wilcoxon rank-sum test.

Supplementary Table 10 - Primordial Germ Cell (PGC) cross species gene expression analysis

Table showing 50 shared and disparate genes when comparing human Primordial Germ cells to Cynomolgus macaque and mouse. Only genes differentially expressed for PGCs in each species are shown.

Supplementary Table 11 - Endoderm subcluster marker genes

List of Endoderm subcluster marker genes. Definitive Endoderm, DE; Yolk Sac, YS.

Supplementary Table 12 - Transcript Isoform differences for Endoderm subclusters

Tables showing transcript isoform comparisons between Endoderm subclusters. Each worksheet refers to the comparison of a single subcluster with every other endoderm subclusters, and includes the names of the genes whose isoforms are differentially expressed with the relative p-value. Definitive Endoderm, DE; Yolk Sac, YS. One-sided chi-square test.

Supplementary Table 13 - Hemato-Endothelial Progenitor subcluster genes

List of Hemato-Endothelial Progenitor subcluster marker genes and associated false discovery rate (FDR).

Supplementary Table 14 - Transcript Isoform differences in Hemato-Endothelial Progenitor subclusters

Tables showing transcript isoform comparisons between Hemogenic/ Endothelial Progenitor (HEP) subclusters. Each worksheet refers to the comparison of a single subcluster with every other HEP subclusters, and includes the names of the genes whose isoforms are differentially expressed with the relative p-value. One-sided chi-square test.

Supplementary Table 15 - RT-PCR Primer details

List of real-time PCR primer sequences.

Supplementary Table 16 - Extraembryonic (EXE) and Advanced Mesoderm Differentially Expressed Genes

List of top 100 most differentially expressed genes between Advanced Mesoderm and Extraembryonic (Exe) mesoderm.

Supplementary Table 17 – Source data for RT-PCR hESC analysis

RT-PCR source data for hESC differentiation *in vitro* analysis in Figure 2f and Extended Data Figure 6b and c. Exact p-values are also provided. Day 0 hESC, PLU; D ay 1 Mesendoderm differentiation, ME; D1 MEK Inhibition, ME+PD. Each gene is shown on a separate sheet