Supplementary Information

Methods

Data availability

Published human embryonic datasets used in this study were obtained from E-MTAB-3929 (ArrayExpress), GSE136447 (publication), http://www.human-gastrula.net/ (raw reads were kindly provided by authors), respectively. Amnion-like cells were obtained from GSE134571, two blastoids datasets were downloaded from GSE150578 and GSE156596, respectively.

Pre-processing single-cell data and gene expression quantification.

All 10X Genomics single-cell data were processed using the Cell Ranger pipeline (v3.0.0, http://software.10xgenomics.com/single-cell/overview/welcome) with default parameters, which uses the STAR aligner (v2.5.1b) to map reads to GRCh38 reference genome (v.3.0.0, GRCh38, downloaded from the 10X Genomics website). To minimize platform and processing differences, published Smart-Seq2 datasets were also mapped on the same reference using the same aligner with default settings, and only uniquely mapped reads were kept for gene expression quantification. Raw read counts were further estimated using rsem-calculate-expression from RSEM tool with the option of “--single-cell-prior”.

Quality control and normalization

Cut-off based on number of expressed genes (nGene) and percentage of mitochondrial genes (percent.mito) were used to filter out the low quality cells. High quality cells from Smart-Seq2 datasets
were required to have at least 2000 nGene and percent.mito less than 0.125. Cells for Amnion-like 10X datasets were filtered using the same cut-off reported in their paper (around 3200 < nGene < 6400, percent.mito < 0.06). For two blastoids datasets, cut-off was chosen based on the general distribution. Cells with percent.mito less than 0.125 were further retained by requiring to have 1000 < nGene < 6500 and 2000 < nGene < 5000 for Stem-blastoids and Iblastoids, respectively. Cells belong to hemogenic endothelial progenitors and erythroblasts from Carnegie Stage 7 were excluded in the following analysis. After quality control and excluding mitochondria genes, we focused on genes with one or more count in at least five cells (assessed for each dataset separately) and calculated log-normalized counts using the deconvolution strategy implemented by the computeSumFactors function in R scran package (v.1.14.6)\textsuperscript{16} and followed by rescaled normalization using the multiBatchNorm function in the R batchelor package (v.1.2.4)\textsuperscript{17} so that the size factors were comparable across batches. log-normalized expression after rescaling were further used in integration and marker-gene detection.

**Restore previous cell annotation for published dataset**

Cells annotation for embryonic Smart-Seq2 datasets were retrieved from their previous publication. For Amnion-like cells and Iblastoids, cell annotation was recaptured by reanalyzing following the standard Seurat tutorial\textsuperscript{18} (For further detail see Code availability section). Cells annotation for Stem-blastoids was determined by integrating with Petropoulos et al., dataset according to the original publication and only cells from blastoids grown on 5iLA collected at Day 9 (labeled as LW60) were used for whole datasets integration. All reanalyzed cell annotations were further confirmed by the marker gene expression reported in their own publication.
**Integrated analysis of multiple datasets**

In total, we have 6 different datasets (3 of them are Smart-Seq2 datasets, the others are 10X datasets). We aligned those 6 datasets based on their mutual nearest neighbours (MNNs) using the RunfastMNN function wrapped in R SeuratWrappers package (v.0.3.0) (https://github.com/satijalab/seurat-wrappers). In detail, this was done by performing a principal component analysis (PCA) on the top 2000 highly variable genes selected by RunFastMNN function (we also tried to use up to 4000 variable genes, it made no difference in general) and then correcting the principal components (PCs) according to their MNNs. We selected the corrected top 25 PCs for downstream UMAP dimensional reduction and clustering analysis using the RunUMAP function and FindClusters function in the R Seurat package. The clustering resolution was set to 0.4 when using the FindClusters function.

**Clustering, marker gene detection and visualization**

Cluster identities for 13 individual clusters were assigned manually using co-localization of cell identities from all embryonic datasets and Amnion-like cells dataset as evidence, as well as marker and signature expression. In detail, cells from Iblastoids and Stem-blastoids were reclassified as Trophectoderm-like cells (TLC; C3), Epiblast-like cells (ELC; C4 and C8), hypoblast-like cells (HLC; C5), Amnion-like cells (AMLC; C1 and C6) and Mesoderm-like cells (MeLC; C0, C2 and C11), respectively. The remained blastoids cells including 4 cells belonging to C10 were assigned as undefined cells. FindMarkers function from R package Seurat performed paired-wise differential expression analysis using ‘roc’ test between lineages. The top 10 up-regulated marker genes with at least average power more than 0.6 conserved in all comparisons were selected for expression dotplot.
and Heatmap. TE, cytotrophoblast, syncytiotrophoblast, and extravillous trophoblast cells were merged as the TE cells. And the nascent mesoderm, emergent mesoderm, advanced mesoderm, and axial mesoderm were lumped as the mesoderm cell in expression dot plot and violin plot. During analysis, we found 50 cells labelled as “ICM” from Xiang et al., dataset highly expressed GATA2 and GATA3 but with low expression levels of NANOG, POU5F1 and BMP2. Also 92% of those cells were localized in TE-majority cluster C3. Therefore, we inferred those cells were misclassified in previous annotation and excluded them in UMAP visualization, marker gene detection and blastocyst classification. In addition, 41 cells labelled as “PSA-EPI” from Xiang et al., 2020 dataset were also excluded for UMAP visualization for potentially mixed with mesoderm cells.

**Code availability**

All data were analysed with standard programs and packages as detailed above. Code used in this project is provided at https://github.com/zhaocheng3326/CheckBlastoids_scripts
Extended Data

Extended Data Fig. 1 Reannotating blastoid cells based on single-cell transcriptomic profiling. a,
UMAP projection of integrated datasets colored for different cell clusters. Cluster distribution for E6 and E7 blastocyst was indicated by the pie plot. From inner to outer, the semi-circle represents the cluster distribution of Epiblast, Primitive Endoderm, and TE cells. **b,** Detailed cluster distribution for embryonic cells based on the previous annotation. The cell numbers of each lineage were labeled as below. **c,** Cell type distribution of blastoid cells based on reclassified annotations. **d,** Dot plot showing expression of the top 10 lineage-specific genes used in Figure 1B in blastoid cells, based on reclassified annotation. The size and colors of dots indicate the proportion of cells expressing the corresponding genes and scaled values of log-transformed expression, respectively.

**Supplementary Reference**


