nature methods

Resource

https://doi.org/10.1038/s41592-024-02511-3

Deep learning-based models for preimplantation mouse and human embryos based on single-cell RNA sequencing

Received: 19 February 2024

Accepted: 15 October 2024

Published online: 14 November 2024

Check for updates

Martin Proks [1,2], Nazmus Salehin [1,2] & Joshua M. Brickman [1]

The rapid growth of single-cell transcriptomic technology has produced an increasing number of datasets for both embryonic development and in vitro pluripotent stem cell-derived models. This avalanche of data surrounding pluripotency and the process of lineage specification has meant it has become increasingly difficult to define specific cell types or states in vivo, and compare these with in vitro differentiation. Here we utilize a set of deep learning tools to integrate and classify multiple datasets. This allows the definition of both mouse and human embryo cell types, lineages and states, thereby maximizing the information one can garner from these precious experimental resources. Our approaches are built on recent initiatives for large-scale human organ atlases, but here we focus on material that is difficult to obtain and process, spanning early mouse and human development. Using publicly available data for these stages, we test different deep learning approaches and develop a model to classify cell types in an unbiased fashion at the same time as defining the set of genes used by the model to identify lineages, cell types and states. We used our models trained on in vivo development to classify pluripotent stem cell models for both mouse and human development, showcasing the importance of this resource as a dynamic reference for early embryogenesis.

Mammalian development begins at fertilization, producing a totipotent zygote that gives rise to the embryo and associated supporting structures. Zygotic transcription begins at relatively early stages, with zygotic genome activation occurring in mouse at the two-cell (2C) stage and in human embryo at the eight-cell (8C) stage[1]. The process of early lineage specification is a highly dynamic, regulative and self-organizing process. The first lineage segregation event occurs at the morula stage (16C), when the outer cells polarize to form the trophectoderm (TE)[2,3] and the inner cells become the inner cell mass (ICM). As or shortly after (depending on the species) these cell types are established, the embryo transitions to the blastocyst stage and produces a fluid-filled blastocoel cavity. During blastocyst maturation, a second lineage specification event is observed as ICM cells differentiate to either epiblast (EPI) or primitive endoderm (PrE). This is followed by the segregation of these lineages, with the PrE becoming positioned between the EPI and blastocyst cavity. At this stage, the embryo is ready to hatch from the zona pellucida and implant into the uterine wall. During postimplantation development, the TE gives rise to the placenta, PrE to the visceral and parietal endoderm, and EPI to the embryo proper[4–6].

Since the explosion in single-cell RNA sequencing (scRNA-seq) techniques, this technology has been extensively applied to these accessible stages of embryonic development. Numerous datasets produced using different technologies have been applied to understanding these early cell fate choices[7–24]. These transcriptional profiles have been an essential resource for defining the developmental stage equivalent of in vitro cell types including human primed[25], naive embryonic stem

[1]The Novo Nordisk Foundation Center for Stem Cell Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [2]These authors contributed equally: Martin Proks, Nazmus Salehin. e-mail: joshua.brickman@sund.ku.dk

https://doi.org/10.1038/s41592-024-02511-3

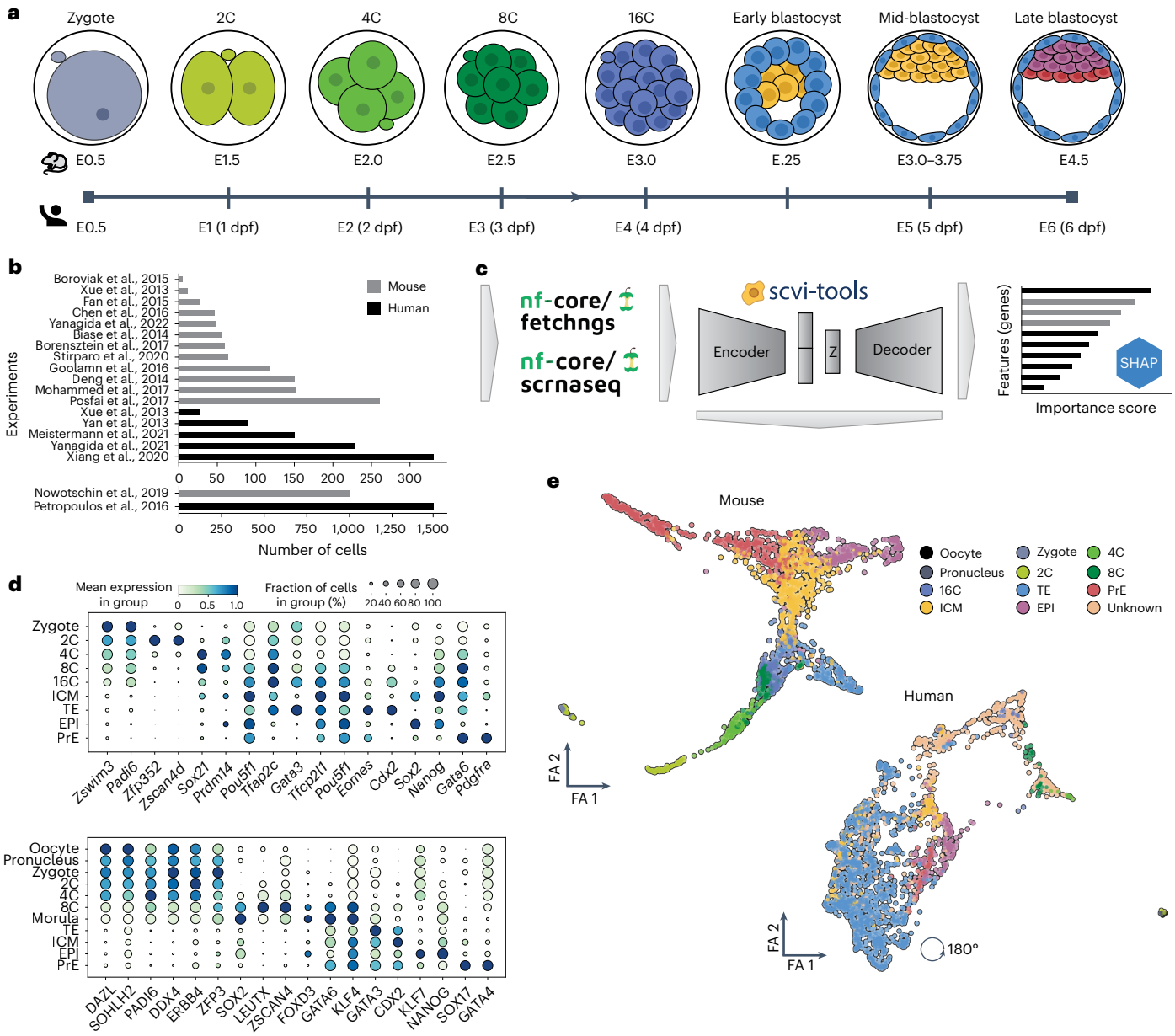


Fig. 1 | Summary of datasets used to build reference models. a, A schematic overview of mouse and human preimplantation development. **b**, Quantification of cells per publication that were collected for building the mouse (gray) and human (black) reference. **c**, A computational schematic of tools used to build and interpret the reference models. **d**, The gene expression of canonical markers for each developmental stage in mouse (top) and human (bottom) preimplantation. **e**, A reduced dimensional representation of preimplantation mouse (left) and human (right) datasets. dpf, days post fertilization; E, embryonic day.

(ES) cells[26–28], trophoblast stem cells[29–31], and PrE stem cells[32] and various three-dimensional in vitro models of development[24,33–36]. However, the need for manual isolation and small numbers of cells in each of the different lineages at these stages means that there is a limited amount of high-quality material when compared to adult organ studies. In addition, the ethical challenges associated with obtaining human embryos means that information gained from even a limited number of cells is extremely valuable. As a result, there is a need for coherent approaches to combine existing datasets and generate a useful and evolving resource that can be used to benchmark the ever-increasing number of cell culture models and cell types. One approach to overcome these challenges and strengthen downstream analyses is to collate and integrate multiple scRNA-seq experiments together.

Traditional data integration techniques assume a linear relationship between datasets, and this is sufficient when batch effects are small and biological complexity low[37]. However, the regulative and dynamic nature of early embryogenesis makes it a robust but variable process that introduces intrinsic variation into each dataset. Moreover, individual sequencing techniques provide varying sequencing depth and different levels of technical noise that distort dataset integration[37]. Lastly, computational demand scales linearly with the amount of data, suggesting current approaches will soon become intractable[38]. These shortcomings have been addressed by deep learning integration techniques employing neural networks and graphical processing units (GPUs) to collapse cells into a shared lower-dimensional latent space that can be used for downstream analyses[39–41]. These approaches have already been applied in the construction of single-cell atlases, such as The Human Cell Atlas[42], The Tabula Sapiens[43] and Human Lung Cell Atlas[44], in addition to organ- or disease-specific atlases[45–48]. While these integrations are promising it is unclear how best to apply them

to study the highly dynamic stages of early development. Moreover, regardless of the model system, the underlying neural networks suffer from a lack of interpretability.

In this Resource, we employed state-of-the-art computational tools to build transcriptomic models of mouse and human preimplantation development. We collated single-cell transcriptomic datasets of mouse and human preimplantation embryos and, by taking advantage of scvi-tools[49] for probabilistic modeling of single-cell omics data, built cell type and time point classifiers. We overcome a critical disadvantage of 'black box' deep learning models by implementing a Shapley additive explanations (SHAP)[50] algorithm to interpret the logic behind lineage classification. Finally, we display the utility of these models to classify lineages generated from in vitro differentiation of mouse and human stem cells, suggesting that these models will be a valuable resource for the community, providing an evolving model to probe phenotype and benchmark increasing numbers of in vitro cell culture models.

Results

An integrated annotated model of preimplantation development

To build a reference model we collected in vivo preimplantation scRNA-seq datasets for both mouse and human embryos covering the stages depicted in Fig. 1a. The datasets were selected only if they were part of a published peer-reviewed article and contained cell metadata for the time of collection as well as cell type annotations. These datasets were then filtered to select only wild-type embryos, rather than experimental genotypes or in vitro stem cells reported alongside these datasets. Based on these criteria, we built a ground truth reference model. Thirteen mouse and six human datasets satisfied these criteria which represent 11 years of studies employing five different sequencing techniques (Table 1 and Extended Data Figs. 1a and 2a). The datasets containing the largest numbers of cells were ref. 15 and ref. 21, for mouse and human, respectively (Fig. 1b and Extended Data Figs. 1b and 2b).

We placed an emphasis on automating the preprocessing by taking advantage of nf-core pipelines[51] to download, align and quantify the datasets (Fig. 1c) so that it takes advantage of improvements in aligner accuracy, performance, updated genome assemblies and gene annotations. In addition, this setup allows simple expansion with new in vivo datasets. To ensure our reference datasets are maintained, future iterations and their refinements will be versioned and accessible on Zenodo and Hugging Face. Steps taken after preprocessing to gene transcript quantification diverged for the mouse and human datasets (Methods).

***Mus musculus*.** As the mouse dataset contained a mixture of full-length and unique molecular identifier (UMI)-based single-cell sequencing, we normalized datasets generated using the SMART-seq1/2 protocol by gene length[52–54]. We discarded ribosomal mitochondrial genes owing to their possible contribution to variance and identification as highly variable genes (HVGs), as well as *Ct010467.1* owing to its high fraction of counts. Based on the quality control, cells containing a minimum of 20,000 transcripts per cell were retained (Methods). These steps yielded a final mouse dataset of 2,004 cells and 34,346 genes.

***Homo sapiens*.** In human, a number of datasets contained cell labels that were ambiguous or intermediate. To avoid introducing uncertainty for classifiers, we set these labels to 'Unknown' despite the availability of annotation on time point and speculation of possible identity. These cells were later used as an internal validation set during model optimization and to test our classifiers. Compared with the mouse dataset, the human dataset contained a disproportionate number of TE annotated cells, and although this can create an imbalance for our classifiers, we decided to retain all the cells. Given the scarcity of material, we sought to bolster cell numbers for classification. Here, we reasoned a minimum of 15 cells per label would be required, and so we collapsed all stages before the 8C stage to 'prelineage'. A similar collation was performed for all cells annotated as PrE regardless of developmental stage. Unlike the mouse studies, those of human relied exclusively on full-read sequencing technologies. However, to ensure the model could be used to integrate cells sequenced using UMI-based technologies, we similarly transformed read counts by gene length.

Table 1 | List of published datasets used to train reference models

Organism	Publication	Technology	GEO accession number
Mus musculus	7	SMART-seq	GSE57249
	8	qRT-PCR	GSE80810
	9	SMART-seq2	E-MTAB-2958
	10	SMART-seq2	GSE74155
	11	SMART-seq 1/2	GSE45719
	12	SUPeR-seq	GSE53386
	13	SMART-seq2	E-MTAB-3321
	14	SMART-seq2	GSE100597
	15	10X v2	GSE123046
	16	SMART-seq	GSE84892
	17	qRT-PCR	GSE44183
	18	SMART-seq2	GSE159030
	19	SMART-seq2	GSE148462
Homo sapiens	20	SMART-seq2	PRJEB30442
	21	SMART-seq2	E-MTAB-3929
	22	SMART-seq2	GSE136447
	23	SMART-seq	GSE36552
	24	SMART-seq2	GSE171820
	17	qRT-PCR	GSE44183

Integration. Based on benchmarking integration strategies for scRNA-seq datasets[37,55], we used single-cell variational inference (scVI)[39] and scGen[56] to integrate existing datasets. We fine tuned parameters during training (Supplementary Table 1) using the autotune feature in scvi-tools. The overall best performance was achieved using two hidden layers and fitting to negative binomial distribution with early stoppage during training for both species. To assess the performance we tracked the evidence lower bound per epoch and calculated batch and biological conservation (Extended Data Figs. 1c and 2c) using the scib-metrics package[37]. scGen ranked first, but for mouse only. This tool was designed to integrate control and perturbed experiments and in trajectory analysis produced disconnected cell types, suggesting an overcorrected batch effect (Extended Data Figs. 1d and 2d). As single-cell annotation using variational inference (scANVI) performed the best in human (Extended Data Fig. 2c) and did not have this problem, we continued with scVI for integration and scANVI[57] for all cell type classification.

Validation

To validate models, we performed a series of downstream analyses using the learned scVI latent space (Fig. 1c, bottleneck layer annotated as Z). We computed the nearest-neighbor graph ($k = 15$), followed by force-directed graph (FA) and Uniform Manifold Approximation and Projection (UMAP)[58] dimension reduction methods (Extended Data Figs. 1f and 2f). Finally, we identified populations of cells using unsupervised Leiden clustering[59] and inferred differentiation trajectories

using partition-based graph abstraction (PAGA)[60] (Extended Data Figs. 1g and 2g).

In the mouse model, branching trajectories covering the first lineage decisions (TE, EPI and PrE) coincide with our understanding of in vivo development. Of the 15 identified clusters, the most dominant and heterogeneous population was the ICM (Extended Data Fig. 1c), which spanned five clusters (Extended Data Fig. 1e). As ICM is a transient differentiating progenitor of both the EPI and PrE, this behavior is expected. In addition, cleavage states (8C to 16C stages) were contained in cluster 1, indicating that these stages are transcriptionally similar (Extended Data Fig. 1e). PAGA trajectory inference correctly connected all developmental stages, confirming the underlying connected graph was consistent with development (Extended Data Fig. 1g). We used diffusion pseudotime (dpt)[61] and scFates[62] to infer pseudotime. However, both tools predicted the TE as a lagging lineage compared with EPI and PrE (Extended Data Fig. 1h), but based on the correct lineage segregation observed with hierarchical trajectory inference using scFates (Extended Data Fig. 1i), we view this as a limitation of pseudotime algorithms.

Given the scarcity of human embryos and disproportionate quantity of specific cell types (the majority of the sampled cells, 55% were originally annotated as TE; Extended Data Fig. 2b), this model proved more complex. As zygotic transcription in human does not start until the 8C stage[1], these cells collapsed together removed from other cell types in the FA plot (Extended Data Fig. 2f). Segregation of TE, EPI and PrE populations was confirmed using PAGA (Extended Data Fig. 2g), but unsupervised clustering (18 clusters) was not able to distinguish individual cell types owing to TE oversaturation (Extended Data Fig. 2e). When pseudotime inference was applied with the prelineage cells as the initial state, dpt failed to assign any temporal sequence. scFates managed to infer a relatively correct pseudotime, but it failed to correctly assign terminal values to either EPI or PrE (Extended Data Fig. 2h,i). We hypothesize that refinement of the underlying scVI representation as well as identification of the unknown cells might align the model better with our understanding of human development.

Classification

One of the most difficult and arduous tasks in scRNA-seq analysis is cell type classification; it requires an extensive knowledge of the studied system and is usually an intuitive process that is difficult to automate. Before training a computational classifier, we compiled a list of canonical markers from in vivo studies for each developmental stage (Fig. 1d). To accelerate and automate this task, we trained two machine learning classifiers using gradient boosting decision trees and neural networks with scANVI. scANVI is a semisupervised model that extends scVI by incorporating cell labels as it refines the underlying scVI latent space, learning the features that enables it to predict a cell type. By default, scANVI outputs the cell type classification with the highest score. For each cell, we use entropy as a measure of the classification uncertainty by subtracting the predicted score from 1.0. Gradient boosting decision trees require a preexisting count matrix, which we generated from denoised RNA expression of HVGs, for training the cell type classifier. Datasets were split 80:20 for training and testing of each classification strategy.

***Mus musculus*.** Given the size of the mouse versus human datasets, we tested both classifiers in mouse. XGBoost classifiers performed the best in terms of the balanced accuracy metric with XGBoost[scVI] (0.96), XGBoost[scANVI] (0.91) and XGBoost[scGEN] (0.91) compared with scANVI (0.64) (Fig. 2a, Extended Data Fig. 3a,b and Supplementary Table 2). scANVI performed relatively poorly (Fig. 2a, middle), but as scANVI is a neural net that looks at the proximity of all cells in latent space and our datasets are not large, we asked if balancing the number of cells in the training set ($n = 15$, per cell type in each training epoch) would improve classification. This adjustment yielded a 23% increase in the balanced accuracy (0.87), strengthening the predictive power for E3.25-ICM/TE, E3.75-ICM and E4.5-TE, which were previously misclassified. As in the clustering analysis, the most difficult cell type to predict was E3.5-ICM ($n = 457$) with only a 46% prediction score (Fig. 2b and Extended Data Fig. 3c). As discussed above, this is probably due to the ICM being heterogeneous, representing multiple stages of EPI and PrE differentiation[63–65].

***Homo sapiens*.** Based on our observation in mouse, we settled on validating only scANVI ($n = 15$) on a human dataset. As a result of this training exercise, scANVI was now able to predict the identity of the unknown cells in the human dataset (Fig. 1e). The majority of unannotated cells at E3.0 and E4.0 were predicted to be from the prelineage embryo and morula, respectively (Fig. 2c, left). The accuracy of these predictions is also apparent in their ability to predict the time points on which these samples were collected in the TE lineage (Fig. 2c).

A second round of prediction with scANVI was performed on the total dataset and used to determine if any of previously classified cells might be misclassified (Fig. 2d). In the third iteration, scANVI reannotated a cluster of ICM cells toward early stages of TE (Fig. 2e). Figure 2e (right) shows a subset of the complete dataset containing only cells from the ICM, EPI and PrE. The majority of the reannotated cells originated from the ref. 22 (Fig. 2e,f) dataset, where other groups have observed inconsistencies[66] and either omitted or manually reannotated this dataset[20,66,67]. As these cells do not cluster with the rest of ICM (Fig. 2e), or express ICM, EPI (*SOX2*, *NANOG*, *KLF17* and *POU5F1*) (Fig. 2f and Extended Data Fig. 3d) or PrE (*SOX17*) (Extended Data Fig. 3d) markers and instead express the TE marker *GATA3* (Fig. 2f), we have updated their annotation in our model. This reannotation results in an inferred trajectory that aligns closer to canonical views of ICM specification of EPI and PrE, in addition to suggesting that EPI can differentiate directly to PrE and retains the capacity to differentiate to TE (Extended Data Fig. 3e), consistent with the ability of naive human ES cells to differentiate to both lineages[32,68].

Robustness of classifiers to HVG dropouts. To assess the robustness of the classifiers, we compared scANVI models with those generated with XGBoost-based classifier, which typically uses fewer features for classification. To test this, we benchmarked the accuracy of both classifiers when they were provided with normalized gene expression after removing the top HVGs used to build them (Fig. 2g). Dropping as few as ten HVGs reduced the performance of the XGBoost-based classifier to close to 10% for all measured accuracies in both species (Fig. 2g and Supplementary Table 3). The scANVI classifier was more robust, losing accuracy only after the removal of the top 200 HVGs. The reduction of robustness observed with XGBoost probably reflects its underlying logic that predicts identity on the basis of the cumulative score from each classifier, making a sequence of binary judgments with the narrowest possible feature set and, therefore, ignoring the stochastic variation in single markers[69], observed in scRNA-seq. The scANVI classifier takes advantage of integrating datasets into one latent space and then fits a multimodal model on annotated cells taking advantage of probabilistic sampling when predicting a cell type.

Explaining scANVI models. We next sought to understand what the features are that define a specific cell type, to uncover the genes used to assign cell type identity and to determine how these genes align with the known markers. scANVI suffers from a 'black-box' issue derived from its neural network architecture consisting of difficult-to-interpret learned weights. This shortcoming can be addressed with methods such as SHAP[50,70] or local interpretable model-agnostic explanations[71] that test the importance of individual features (genes) for each prediction by exclusion. Although SHAP has been applied to XGBoost, it has never been applied to neural networks like scANVI. We therefore devised an scANVI explainer (scANVIExplainer), which estimates SHapley values to quantify the feature contributions used in cell type prediction.

https://doi.org/10.1038/s41592-024-02511-3

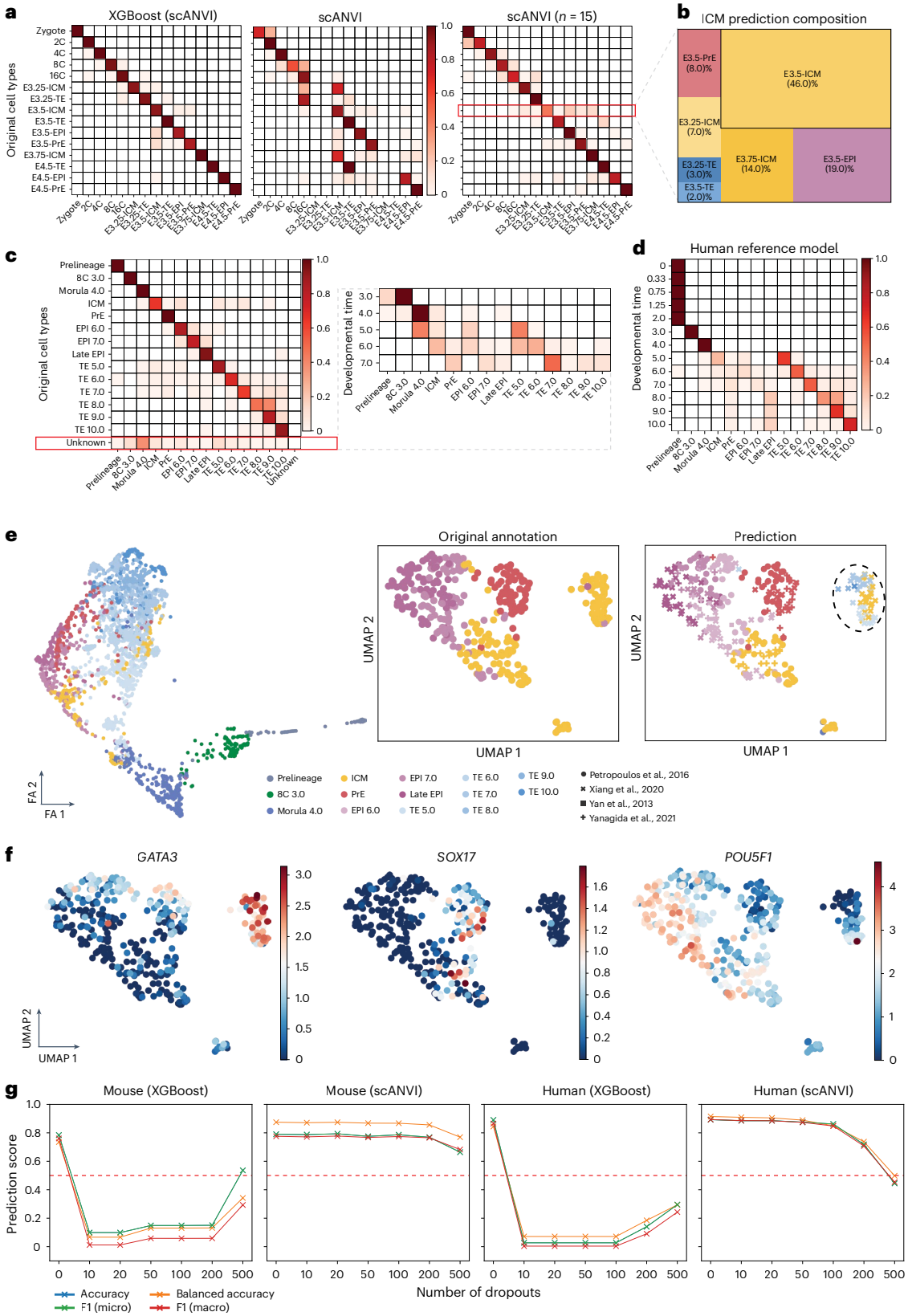


Fig. 2 | Cell type classification. a, The accuracy of predictions from three different mouse classifiers XGBoost (left), scANVI (middle) and scANVI with cell type subsampling (scANVI ($n = 15$), right) (x axis, predicted; y axis, observed). The scale represents the prediction score for each individual cell type. **b**, A closer inspection of how annotated E3.5-ICM were predicted by scANVI ($n = 15$). **c**, The accuracy of the subsampled scANVI classifier (scANVI ($n = 15$), left) for the human reference, including reannotation of previously unannotated cells (right) (x axis, predicted; y axis, observed). **d**, Classifier annotations for cells sampled at known developmental times **e**, FA graph (left) and UMAP dimensional reduction (middle and right) displaying cells originally annotated as ICM but predicted to be TE. **f**, The expression of *GATA3*, *SOX17* and *POU5F1* in the ICM and ICM derivative subset. **g**, The impact of removing the top dispersion HVGs on the classification performance of XGBoost and subset scANVI classifiers.

https://doi.org/10.1038/s41592-024-02511-3

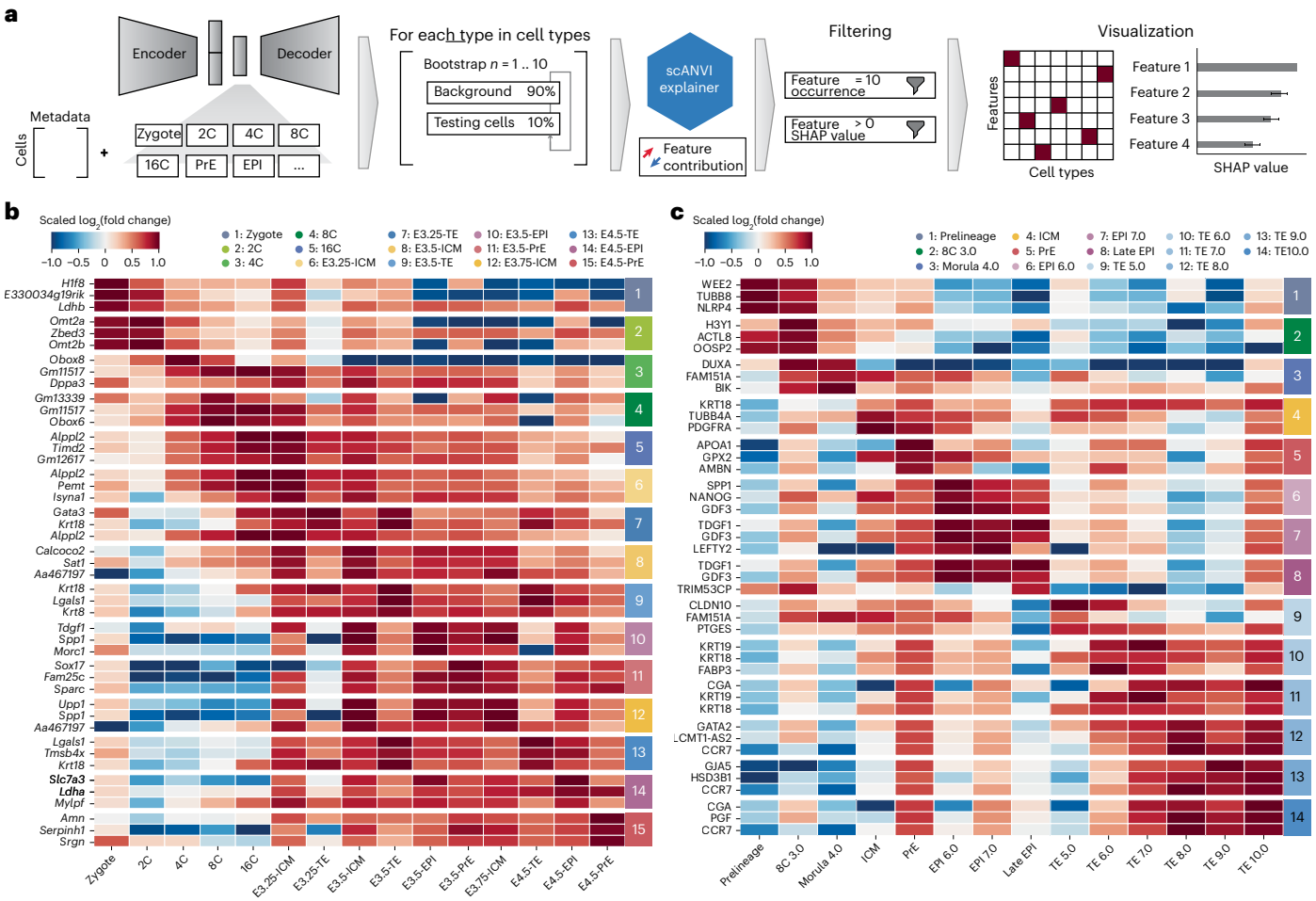


Fig. 3 | Extracting key predicting features with scANVIExplainer. **a**, A schematic overview of how scANVIExplainer works. In brief, each cell type is split randomly into 90/10 (train/test) subsets to identify the importance of each individual feature in its respective cell type. This is repeated with bootstrapping ten times using shifting populations of cells, and only features that are present in each bootstrap are considered identifiers. **b**,**c**, The differential expression analysis (one versus all) of genes identified as top three predictors for each cell type in mouse (**b**) and human (**c**) classifiers. The heatmap displays $\log_2$(fold change) of cell type versus all other cell types. The legend for vertical cell type identification is given at the top of each heatmap.

To do this, we adapted DeepExplainer[72], a tool designed for deep learning models. Although we had initially used KernelExplainer[50] its runtime scaled poorly with increasing size of datasets. As DeepExplainer appeared more efficient, we modified it to employ scANVI architecture to develop scANVIExplainer. scANVIExplainer attempts to weight each feature contribution (positive or negative) in predicting a cell type. It first splits the input data into 90/10 (train/background) to estimate approximate conditional expectation SHAP values (background). Next, using these background estimates, each feature is assigned an importance value based on the weighted difference between the background and tested feature. To extract a robust feature set for each cell type, scANVIExplainer performed ten bootstrapped runs and discarded features that did not have positive weights throughout all ten iterations (Fig. 3a). To assess the physiological relevance of these features, we performed differential expression analysis. Figure 3b,c shows a heatmap of gene expression changes for the top three ranked features. A full list of features alongside their corresponding SHAP value is presented in Extended Data Fig. 5a,b and Supplementary Table 4.

In both mouse and human, the classifier used a combination of canonical and noncanonical markers. In mouse, we found that some prominent markers used for staining preimplantation lineages, such as *Cdx2* (TE), *Gata6* (PrE) and *Nanog* (ICM/EPI), were not reported in the top list. The mouse model used both genes associated with early development such as *Omt2a*, *Obox8* and *Dppa3* and canonical markers such as *Gata3* (TE), *Sox17* (PrE) and *Spp1* (ICM) (Extended Data Fig. 5a). Similar analysis in the human model revealed the classification of the 8C stage relied on the previously identified *NLRP4* and the oocyte factor oocyte secreted protein 2 (*OOSP2*) (Extended Data Fig. 5b). As with the mouse, the human model also utilized traditional markers, such as *PDGFRA* for PrE, *NODAL* and *GDF3* for the EPI and *KRT18*, *CGF* and *PGF* for TE (Supplementary Table 4 and Extended Data Fig. 5b).

Query integration using the optimized scANVI models

One of the key advantages of the scANVI model is simple integration and classification of new datasets; below, we discuss three instances where we have probed the utility of the model.

Model extension. To demonstrate that our model could readily be extended, we added data from in vitro embryo culture through E14 of human development[22]. Extended Data Fig. 4 shows that including later stages of development nicely integrates with the existing model and positions postimplantation EPI adjacent to preimplantation EPI. Moreover, in the PAGA, the primitive streak anlage is linked to late EPI, consistent with primitive streak induction from EPI at the beginning of gastrulation[73].

Mouse in vitro PrE differentiation. We developed an in vitro model for PrE differentiation using ES cells[74] and, with massively parallel

https://doi.org/10.1038/s41592-024-02511-3

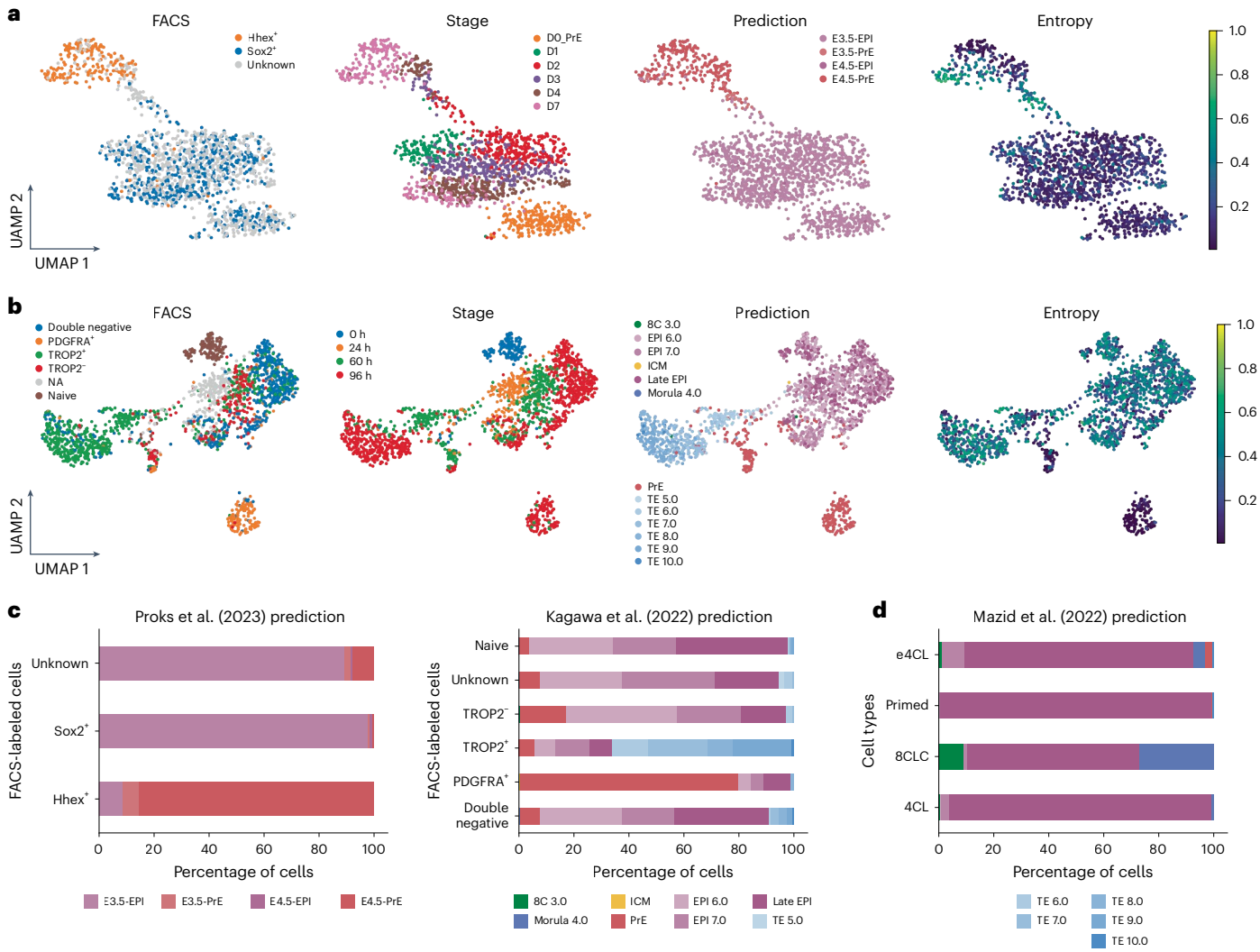


Fig. 4 | Classifying in vitro datasets. a, The prediction of cell types generated during mouse in vitro PrE differentiation in HHex/Sox2 double-reporter ES cells. **b**, The prediction of cell types generated in a human stem cell-based model of blastocyst development. **c**, Left: predicted cell type proportions within mouse PrE differentiation from ES cells compared with reporter expression. Right: predicted cell type proportions within in vitro human blastoids compared with cell surface marker expression. **d**, Shifts in predicted cell type proportions in human primed ES cell (primed) cultures transferred to conditions to produce naive (4CL) and 4C-like (e4CL) cells or sorted to enrich for 8C-like cells (8CLC)[82].

single-cell RNA-seq (MARS-seq)[75], profiled single-cell gene expression in a line harboring fluorescent protein reporters for EPI(Sox2–GFP) and PrE(Hhex–mCherry)[76,77]. Naive ES cells differentiated toward PrE over 7 days[74] were sorted by fluorescence-activated cell sorting (FACS), based on reporter expression, for equal quantities of each population before sequencing. In these experiments, PrE differentiation is initiated by HHEX-expressing clusters of cells[76,77] that expand and further differentiate with time. Figure 4a shows the dataset, flow cytometry and predicted scANVI cell types indicating that these cells pass through an E3.5 PrE state to a terminal E4.5 PrE state, while the population that fails to differentiate retains an EPI-like identity. The remaining SOX2$^+$ cells were classified as E3.5/E4.5 EPI-like cells, consistent with our previous findings that these cells express a pluripotent gene expression signature. Figure 4a (ref. 76) also shows an entropy score (uncertainty) for the scANVI prediction that suggests accurate predictions (Fig. 4c, left and Supplementary Table 5).

Human blastoids. A number of recent studies have demonstrated that naive human ES cells are capable of self-organizing into integrated stem cell-based models of human blastocysts called blastoids[33,34,78,79]. While scRNA-seq datasets have been generated from these structures in a range of conditions, we chose a dataset that contained both scRNA-seq and cell surface marker annotations[33] and asked our scANVI model to infer identity. These structures were generated from naive human ES cells cultured in PXGL (media containing PD0325901, XAV939, Gö6983 and human LIF)[80] that were transferred into microwells for blastoid formation and sorted by FACS for lineage markers (PDGFRA$^+$ PrE, TROP2$^+$ TE and double-negative EPI) before scRNA-seq. Figure 4b shows the flow cytometry and time course for differentiation into blastoids in the first two panels, alongside the predicted scANVI identity and entropy. While some predictions have lower certainty (higher entropy) than we achieved in mouse, scANVI reliably predicted PrE emergence and TE maturation. As expected, we find a large EPI-like population that has a similar identity to the starting naive human ES cells. While it has been suggested that naive human ES cells resemble early EPI and dedifferentiate to ICM identity to generate TE in blastoids[33,68], scANVI did not predict the existence of ICM at any time point in blastoid formation. Taken together, we demonstrated the capacity of our scANVI classifier to act in the absence of manual curation or extensive knowledge of developmental biology to assign cell identity solely on the basis of single-cell transcriptomes (Fig. 4c right and Supplementary Table 5).

Human 8C-like induction. A subpopulation of human ES cells was recently identified that appear to express markers of the 8C stage, when zygotic transcription is initiated, and resemble the mouse 2C state normally found in naive mouse ES cell culture[6,81,82]. Are these cells representative of the 8C stage of development, and do human ES cells, which are EPI like, revert to an 8C population and vice versa? A cocktail of small molecules (PD0325901, TSA, DZNep and IWR1) and LIF (4CL) and an enhanced version of the same cocktail (e4CL) were reported to support 8CLCs, and we therefore sought to predict the identity of these cells with scANVI. The model correctly identified a small 8CLC population in the e4CL medium and found that culture of primed human ES cells in 4CL, and then further in e4CL, produced a progressive shift from late EPI toward the morula and 8C identity, suggesting this cytokine cocktail may drive stepwise developmental reprogramming and vice versa (Fig. 4d).

Discussion

In this Resource, we applied deep learning tools to scRNA-seq from early embryos to generate dynamic models with predictive power that can be used to benchmark in vitro cell types. We do this from a readily assembled, compartmentalized set of tools wrapped in nf-core pipelines and scvi-tools. These models can both predict cell type identity and, when coupled with our newly built scANVIExplainer, estimate SHapley values or produce novel marker sets for the unbiased identification of cell types. To facilitate the uptake of our model, we have developed a portal (https://brickman-preimplantation.streamlit.app) for inspecting and visualizing reference models. We also provide a Google Colab notebook containing code to easily retrain or query new datasets. One of the limitations of our models is the number of cells and imbalanced cell types. This is complicated by the nontrivial task of embryo dissection and the scarcity of human material. Despite this, our downstream analyses provide a snapshot of development that is consistent with existing knowledge of in vivo development. However, our classifiers will improve as the models expand, and we invite others to augment these models with further reference datasets to strengthen the robustness of integration and cell type classification.

Cell type classification has traditionally been based on morphology, functionality and position within an embryo or adult organ[83,84]. In the wake of the molecular revolution in developmental and stem cell biology, identification became based on gene expression exploiting historically defined marker sets. Markers were discovered on the basis of gene expression and functionality, and then came to define specific cell types. In preimplantation development, lineages were defined on the basis of marker expression such as *Nanog* or *Sox2* for EPI, *Cdx2* or *Eomes* for TE or *Gata4* and *Gata6* for PrE[6,85]. As a result, phenotype analysis has relied on accumulated knowledge, rather than a systematic and unbiased approach to defining cell type identity. Recent advances in our ability to capture whole genome expression in single cells suggest that RNA best describes a range of gene expression identities that comprise a cell state rather than discrete cell types[86,87]. These cell states are defined using unsupervised clustering as populations of cells that are then annotated on the basis of differential expression analysis and manual curation. Although some historically defined markers have been useful in defining cell types, these factors are not selected on the basis of their ability to provide unequivocal identity and ignore markers that have as yet been discovered. By using an unbiased deep learning approach, we find a mix of previously identified factors including *Gata3* (TE), *Sox17* (PrE), *PDGFRA* (PrE) and *NODAL* (EPI) as well as new ones. Taken together, this suggests that, although experimental developmental biology has identified some key markers, the identification of cell types using single-cell transcriptomes may better define cell types in an unbiased fashion where the criteria may not be biological function in the lineage but rather the unbiased assignment of cell type identity based on the best combination of markers.

While our current models are limited to preimplantation development, the model can still be used to benchmark in vitro cell types not contained in our dataset on the basis of the entropy score[57]. In both the human and the mouse, we observe that naive ES cells are EPI like, although multiple studies have suggested they can form different extraembryonic lineages[6,29–32,88–92], and at least in human, we found this is not a result of regression to an ICM-like identity[6,88–92].

Although our models only use HVGs limiting the gene space, refs. 37,57 demonstrated that similar integrations using the full genome did not substantially improve overall performance. One of the drawbacks of this study is the imbalanced dataset and limited number of cells, due to technical and ethical restrictions; despite the abundance of TE in our human dataset, our approach to balancing the classifier produced a reasonably high-fidelity model. In addition, our downstream data analysis is consistent with current knowledge of in vivo development. However, in the absence of experimental validation, we cannot exclude the possibility that our models are unable to properly define and identify new cell types. We believe that further expansion with new data will improve the robustness of the models in future.

The past year has seen the development of an abundance of in vitro stem cell-based model systems and revised ex vivo culture conditions for human embryos[93,94]. The online resource accompanying this paper, while useful, highlights the need for new approaches to both classification and phenotype analysis. We believe our models are an excellent complement to already established large organ atlases[42–48] and attempts to model gastrulation[15,95,96]. Given the scarcity and difficulty in obtaining human material, models like these represent essential resources for computational interrogation of genetic and biochemical perturbation of early development and in vitro models.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02511-3.

References

1. Yuan, S. et al. Human zygotic genome activation is initiated from paternal genome. *Cell Discov.* **9**, 13 (2023).
2. Nishioka, N. et al. The Hippo signaling pathway components lats and yap pattern tead4 activity to distinguish mouse trophectoderm from inner cell mass. *Dev. Cell* **16**, 398–410 (2009).
3. Gerri, C. et al. Initiation of a conserved trophectoderm program in human, cow and mouse embryos. *Nature* **587**, 443–447 (2020).
4. Gilbert, S. F. *Developmental Biology* 6th edn (Sinauer Associates, 2000); https://www.ncbi.nlm.nih.gov/books/NBK10052/
5. Saiz, N. & Plusa, B. Early cell fate decisions in the mouse embryo. *Reproduction* **145**, R65–R80 (2013).
6. Riveiro, A. R. & Brickman, J. M. From pluripotency to totipotency: an experimentalist's guide to cellular potency. *Development* https://doi.org/10.1242/dev.189845 (2020).
7. Biase, F. H., Cao, X. & Zhong, S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.* **24**, 1787–1796 (2014).
8. Borensztein, M. et al. Xist-dependent imprinted X inactivation and the early developmental consequences of its failure. *Nat. Struct. Mol. Biol.* **24**, 226–233 (2017).
9. Boroviak, T. et al. Lineage-specific profiling delineates the emergence and progression of naive pluripotency in mammalian embryogenesis. *Dev. Cell* **35**, 366–382 (2015).
10. Chen, G. et al. Single-cell analyses of X chromosome inactivation dynamics and pluripotency during differentiation. *Genome Res.* **26**, 1342–1354 (2016).

11. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
12. Fan, X. et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* **16**, 148 (2015).
13. Goolam, M. et al. Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* **165**, 61–74 (2016).
14. Mohammed, H. et al. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep.* **20**, 1215–1228 (2017).
15. Nowotschin, S. et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361–367 (2019).
16. Posfai, E. et al. Position- and Hippo signaling-dependent plasticity during lineage segregation in the early mouse embryo. *eLife* **6**, e22906 (2017).
17. Xue, Z. et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593–597 (2013).
18. Stirparo, G. G. et al. OCT4 induces embryonic pluripotency via STAT3 signaling and metabolic mechanisms. *Proc. Natl Acad. Sci. USA* **118**, e2008890118 (2021).
19. Yanagida, A. et al. Cell surface fluctuations regulate early embryonic lineage sorting. *Cell* **185**, 777–793.e20 (2022).
20. Meistermann, D. et al. Integrated pseudotime analysis of human pre-implantation embryo single-cell transcriptomes reveals the dynamics of lineage specification. *Cell Stem Cell* **28**, 1625–1640. e6 (2021).
21. Petropoulos, S. et al. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026 (2016).
22. Xiang, L. et al. A developmental landscape of 3D-cultured human pre-gastrulation embryos. *Nature* **577**, 537–542 (2019).
23. Yan, L. et al. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
24. Yanagida, A. et al. Naive stem cell blastocyst model captures human embryo lineage segregation. *Cell Stem Cell* **28**, 1016–1022. e4 (2021).
25. Thomson, J. A. et al. Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147 (1998).
26. Gafni, O. et al. Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**, 282–286 (2013).
27. Takashima, Y. et al. Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* **162**, 452–453 (2015).
28. Bredenkamp, N., Stirparo, G. G., Nichols, J., Smith, A. & Guo, G. The cell-surface marker sushi containing domain 2 facilitates establishment of human naive pluripotent stem cells. *Stem Cell Rep.* **12**, 1212–1222 (2019).
29. Dong, C. et al. Derivation of trophoblast stem cells from naïve human pluripotent stem cells. *eLife* https://doi.org/10.7554/eLife.52504 (2020).
30. Cinkornpumin, J. K. et al. Naive human embryonic stem cells can give rise to cells with a trophoblast-like transcriptome and methylome. *Stem Cell Rep.* **15**, 198–213 (2020).
31. Okae, H. et al. Derivation of human trophoblast stem cells. *Cell Stem Cell* **22**, 50–63.e6 (2018).
32. Linneberg-Agerholm, M. et al. Naïve human pluripotent stem cells respond to Wnt, Nodal, and LIF signalling to produce expandable naïve extra-embryonic endoderm. *Development* https://doi.org/10.1242/dev.180620 (2019).
33. Kagawa, H. et al. Human blastoids model blastocyst development and implantation. *Nature* **601**, 600–605 (2021).
34. Liu, X. et al. Modelling human blastocysts by reprogramming fibroblasts into iblastoids. *Nature* **591**, 627–632 (2021).
35. Yu, L. et al. Blastocyst-like structures generated from human pluripotent stem cells. *Nature* **591**, 620–626 (2021).
36. Fan, Y. et al. Generation of human blastocyst-like structures from pluripotent stem cells. *Cell Discov.* https://doi.org/10.1038/s41421-021-00316-8 (2021).
37. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
38. Angerer, P. et al. Single cells make big data: new challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* **4**, 85–91 (2017).
39. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
40. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2021).
41. Erfanian, N. et al. Deep learning applications in single-cell genomics and transcriptomics data analysis. *Biomed. Pharmacother.* **165**, 115077 (2023).
42. Regev, A. et al. The human cell atlas. *eLife* https://doi.org/10.7554/eLife.27041 (2017).
43. The Tabula Sapiens Consortium et al. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* https://doi.org/10.1126/science.abl4896 (2022).
44. Sikkema, L. et al. An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
45. Swamy, V. S., Fufa, T. D., Hufnagel, R. B. & McGaughey, D. M. Building the mega single-cell transcriptome ocular meta-atlas. *GigaScience* https://doi.org/10.1093/gigascience/giab061 (2021).
46. Eraslan, G. et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* https://doi.org/10.1126/science.abl4290 (2022).
47. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* https://doi.org/10.1126/science.abl5197 (2022).
48. Suo, C. et al. Mapping the developing human immune system across organs. *Science* https://doi.org/10.1126/science.abo0510 (2022).
49. Gayoso, A. et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).
50. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. Preprint at https://arxiv.org/abs/1705.07874 (2017).
51. Ewels, P. A. et al. The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).
52. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
53. Phipson, B., Zappia, L. & Oshlack, A. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res.* **6**, 595 (2017).
54. Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-023-00586-w (2023).
55. Brendel, M. et al. Application of deep learning on single-cell RNA sequencing data analysis: a review. *Genomics Proteomics Bioinform.* **20**, 814–835 (2022).
56. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
57. Xu, C. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
58. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2020).

59. Traag, V., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
60. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
61. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
62. Faure, L., Soldatov, R., Kharchenko, P. V. & Adameyko, I. scFates: a scalable Python package for advanced pseudotime and bifurcation analysis from single-cell data. *Bioinformatics* **39**, btac746 (2023).
63. Plusa, B., Piliszek, A., Frankenberg, S., Artus, J. & Hadjantonakis, A.-K. Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. *Development* **135**, 3081–3091 (2008).
64. Ohnishi, Y. et al. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell Biol.* **16**, 27–37 (2013).
65. Saiz, N., Williams, K. M., Seshan, V. E. & Hadjantonakis, A.-K. Asynchronous fate decisions by single cells collectively ensure consistent lineage composition in the mouse blastocyst. *Nat. Commun.* https://doi.org/10.1038/ncomms13463 (2016).
66. Zhao, C. et al. A comprehensive human embryo reference tool using single-cell RNA-sequencing data. *Nat. Methods* https://doi.org/10.1038/s41592-024-02493-2 (2024).
67. Radley, A., Corujo-Simon, E., Nichols, J., Smith, A. & Dunn, S.-J. Entropy sorting of single-cell RNA sequencing data reveals the inner cell mass in the human pre-implantation embryo. *Stem Cell Rep.* **18**, 47–63 (2023).
68. Guo, G. et al. Human naive epiblast cells possess unrestricted lineage potential. *Cell Stem Cell* **28**, 1040–1056.e6 (2021).
69. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
70. Strumbelj, E. & Kononenko, I. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.* **11**, 1–18 (2010).
71. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should i trust you?': explaining the predictions of any classifier. Preprint at https://arxiv.org/abs/1602.04938 (2016).
72. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *Proc. 34th International Conference on Machine Learning* 3145–3153 (PMLR, 2017).
73. Rossant, J. & Tam, P.P.L. Early human embryonic development: blastocyst formation to gastrulation. *Dev. Cell* **57**, 152–165 (2022).
74. Anderson, K. G. V. et al. Insulin fine-tunes self-renewal pathways governing naive pluripotency and extra-embryonic endoderm. *Nat. Cell Biol.* **19**, 1164–1177 (2017).
75. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
76. Perera, M. et al. Transcriptional heterogeneity and cell cycle regulation as central determinants of primitive endoderm priming. *eLife* **11**, e78967 (2022).
77. Proks, M., Herrera, J. A. R., Sedzinski, J. & Brickman, J. M. nf-core/marsseq: systematic pre-processing pipeline for MARS-seq experiments. Preprint at *bioRxiv* https://doi.org/10.1101/2023.06.28.546862 (2023).
78. Karvas, R. M. et al. 3D-cultured blastoids model human embryogenesis from pre-implantation to early gastrulation stages. *Cell Stem Cell* **30**, 1148–1165.e7 (2023).
79. Yu, L. et al. Large-scale production of human blastoids amenable to modeling blastocyst development and maternal-fetal cross talk. *Cell Stem Cell* **30**, 1246–1261.e9 (2023).
80. Bredenkamp, N. et al. Wnt inhibition facilitates RNA-mediated reprogramming of human somatic cells to naive pluripotency. *Stem Cell Rep.* **13**, 1083–1098 (2019).
81. Genet, M. & Torres-Padilla, M.-E. The molecular and cellular features of 2-cell-like cells: a reference guide. *Development* https://doi.org/10.1242/dev.189688 (2020).
82. Mazid, M. A. et al. Rolling back human pluripotent stem cells to an eight-cell embryo-like stage. *Nature* **605**, 315–324 (2022).
83. Mulas, C., Chaigne, A., Smith, A. & Chalut, K. J. Cell state transitions: definitions and challenges. *Development* https://doi.org/10.1242/dev.199950 (2021).
84. Fleck, J. S., Camp, J. G. & Treutlein, B. What is a cell type? *Science* **381**, 733–734 (2023).
85. Biondic, S., Canizo, J., Vandal, K., Zhao, C. & Petropoulos, S. Cross-species comparison of mouse and human preimplantation development with an emphasis on lineage specification. *Reproduction* **165**, R103–R116 (2023).
86. Huang, S. *Multistability and Multicellularity: Cell Fates as High-Dimensional Attractors of Gene Regulatory Networks* (eds Kriete, A. & Eils, R.) 293–326 (Elsevier, 2006); https://doi.org/10.1016/B978-012088786-6/50033-2
87. Enver, T., Pera, M., Peterson, C. & Andrews, P. W. Stem cell states, fates, and the rules of attraction. *Cell Stem Cell* **4**, 387–397 (2009).
88. Beddington, R. S. P. & Robertson, E. J. An assessment of the developmental potential of embryonic stem cells in the midgestation mouse embryo. *Development* **105**, 733–737 (1989).
89. Macfarlan, T. S. et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
90. Morgani, S. M. et al. Totipotent embryonic stem cells arise in ground-state culture conditions. *Cell Rep.* **3**, 1945–1957 (2013).
91. Gonzalez, J. M. et al. Embryonic stem cell culture conditions support distinct states associated with different developmental stages and potency. *Stem Cell Rep.* **7**, 177–191 (2016).
92. Redó-Riveiro, A. et al. Transcription factor co-expression mediates lineage priming for embryonic and extra-embryonic differentiation. *Stem Cell Rep.* **19**, 174–186 (2024).
93. Oldak, B. et al. Complete human day 14 post-implantation embryo models from naive es cells. *Nature* **622**, 562–573 (2023).
94. Weatherbee, B. A. T. et al. Pluripotent stem cell-derived model of the post-implantation human embryo. *Nature* **622**, 584–593 (2023).
95. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
96. Mittnenzweig, M. et al. A single-embryo, single-cell time-resolved model for mouse gastrulation. *Cell* **184**, 2825–2842.e22 (2021).

Methods

Downloading and preprocessing of scRNA-seq data

We compiled a curated list of datasets from different publications. Raw sequencing (FASTQ) files were downloaded from the National Center for Biotechnology Information Gene Expression Omnibus and European Nucleotide Archive repositories using the nf-core/fetchngs (v1.10.0) pipeline[97]. For gene expression quantification, we forked nf-core/scrnaseq[98] to brickmanlab/scrnaseq (branch: feature/scrnaseq) and adapted it to support SMART-seq 1/2 experiments. Transcript abundance was estimated using STARsolo using indexes built from the GRCm38 reference genome and Ensembl 102 gene annotations or the GRCh38 reference genome and Ensembl 110 gene annotations for mouse and human datasets, respectively.

Data harmonization and normalization

The metadata for each cell in the dataset contain information about the experiment/publication, sequencing technology, batches and original cell type annotation. We harmonized the original annotations as follows. (1) In the case of the mouse data, if a cell type was defined as 2C early/mid/late, these cells were reannotated to 2C cells. (2) In the case of the human data, where some cell types were transitionary, these were reannotated as unknown. These reannotations were saved as 'ct', for cell type, while the original annotations were saved as 'ct_orig' in the metadata.

Gene transcript quantification was further processed using the scanpy[99] library. To ensure compatibility between UMI-based and full-length technologies, full-length (SMART-seq) sequencing was normalized to mean gene length obtained from Ensembl gene annotations using gtftools[100] and rounded to the nearest integer. For mouse datasets, genes were filtered to exclude ribosomal, cell cycle, mitochondrial genes and those that were expressed in fewer than ten cells. In addition, we discarded data arising from oocytes owing to their low representation ($n = 13$). Cells were filtered to exclude those expressing more than 20,000 genes and 26,000,000 counts. Raw counts were then depth normalized to median counts in the filtered dataset and log1p transformed. For human datasets, genes were filtered to exclude ribosomal, mitochondrial genes and those that were expressed in fewer than ten cells. Cells collected until the 8C stage were reannotated as prelineage. Raw counts were depth normalized to 10,000 total counts and log1p transformed.

Integration and classification training

The top 3,000 HVGs were identified using the sc.pp.highly_variable_genes function within scanpy, with the following arguments: flavour='cell_ranger' and batch_key='experiment'. The scVI models were then built using two hidden layers and a negative binomial gene likelihood and trained for a maximum of 400 epochs, with early stoppage. The scANVI model was built on top of the scVI integration with provided cell type labels and trained for 15 epochs. scGEN was trained using normalized counts, with batch_key='batch' and labels_key='ct', for a maximum of 100 epochs, with a batch size of 32 and early stoppage enabled. To compare PCA, scVI, scANVI and scGen integration methods, we used the GPU-accelerated scib-metrics Python package (v0.4.1) to compute the evaluation metrics defined in ref. 37.

To fine tune the arguments used to build the reference models, we took advantage of an experimental scvi.autotune package that allowed for rapid model generation using combinations of arguments. In both datasets, we adjusted the search space with the following parameters: gene_likelihood (nb and zinb), gene dispersion (gene and gene-batch), number of hidden layers (128, 144 and 256), number of layers (2–5) and learning rate range (between 1×10^{-4} and 0.6). For each hyperparameter optimization, we measured the validation loss with a maximum of 100 training epochs and generated, in total, 50 models (Supplementary Table 1).

The gradient boosted decision tree classifiers were built using the XGBoost library, trained on denoised expression matrices obtained using normalized expression values from the scVI, scANVI and scGen models: get_normalized_expression function (return_mean=True, return_numpy=True). Predictions were benchmarked by calculating accuracy, balanced accuracy, f1 (micro and macro) scores using the scikit-learn Python package[101], where the test set was the whole reference model.

Dimensional reduction visualization and trajectory inference

Nearest-neighbor graphs ($k = 15$) were calculated from the scVI normalized learned latent space (get_latent_representation function) using the scanpy function sc.pp.neighbors. To identify cell clusters, we used unsupervised Leiden clustering[59] with resolution set to 0.8. Next, we performed principal component analysis to visually inspect the spread of cell types. Using the nearest-neighbor graph, we performed dimension reduction using UMAP[58] and FA to visualize data in a two-dimensional plot.

Trajectory inference was performed by calculating a diffusion map (sc.tl.diffmap) followed by PAGA analysis[60] (sc.tl.paga) with default settings. Pseudotime was estimated using the dpt[61] package, with 'zygote' specified as the initial state for both species. For scFates[62], we recomputed the diffusion map and found the optimal sigma value (500 for both) using the scf.tl.explore_sigma function. Based on the scFates visualization scf.pl.graph, the initial states were set to 'zygote' and '4C' for mouse and human, respectively. Estimated pseudotime was scaled from 0 to 1 as is reported by the dpt package.

XGBoost classification

Gradient boosting decision trees using XGBoost classification requires denoised RNA expression generated from the decoder of scVI, scANVI and scGen. To mitigate overfitting, we allowed for early stoppage of the training if the logloss metric failed to improve in the previous ten iterations. Lastly, we performed 10-fold cross-validation to confirm the robustness of the classification model.

scANVIExplainer

To explain which features (genes) are used to determine a cell type, we used SHAP[50,102]. For XGBoost, we use shap.GPUTreeExplainer. For scANVI, we had to modify the original DeepExplainer because scANVI requires count matrix (X), batch and cell type annotation (labels) for initialization. The current implementation does not support additional covariates. We next modified code to call classifier for the learned latent space (Z). We provide the code for scANVIExplainer in deep_scanvi.py. For all classifiers, we split the reference dataset to 90:10 (background:test dataset) to obtain an expected value for each feature. The sum of weighted Shapley values from the tested dataset are then subtracted from the expected background to estimate the contribution score in each cell prediction. This is executed ten times, and we keep only the features that occurred in every iteration. Next, we subset for nonnegative features only and calculate mean and standard deviation for each of them. Lastly, we rank the features for each predicted cell type based on their mean value. The source code originally used for scANVIExplainer can be found at https://github.com/brickmanlab/proks-salehin-et-al/blob/master/scripts/deep_scanvi.py. Since then, we have formalized the scANVIExplainer into a Python package https://github.com/brickmanlab/scanvi-explainer.

Extension of the model to postimplantation stages of human development

scRNA-seq from ex vivo cultured later-stage (E12 and E14) human embryos[22] were preprocessed as with the preimplantation human embryo datasets. The preimplantation model was downloaded from Hugging Face, and count data from E12 and E14 embryos were appended to the downloaded version of the model data. Training, classification and trajectory inference were performed as with the preimplantation model.

In vitro predictions

For the mouse in vitro dataset, we downloaded the raw count matrix from ref. 77 and proceeded without reprocessing. The human blastoid scRNA-seq dataset was reanalyzed from raw sequencing reads. The list of accession numbers was fed into nf-core/fetchngs, which downloaded raw FASTQ files. These were then immediately preprocessed using the brickmanlab/scrnaseq pipeline. As the blastoid human dataset was sequenced using SMART-seq2, we normalized the raw count by mean gene length. Preprocessed counts from scRNA-seq data from the 8CLC induction dataset were downloaded from figshare at https://figshare.com/s/34110eebb58462a79dd5. All datasets were integrated to their species appropriate scANVI reference model with the following settings: max_epochs=100, plan_kwargs=dict(weight_decay=0.0) and check_val_every_n_epoch=10. For each integrated cell, we obtained the predicted cell type, defined as the cell type with the highest probability. Uncertainty (entropy) for the prediction was calculated as 1 − the highest cell type probability: 1 - lvae_q.predict(soft=True).max(axis=1).

Reporting summary

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

Data availability

Reference downloading and preprocessing pipelines can be found via GitHub at https://github.com/nf-core/fetchngs (revision 1.10.0) and https://github.com/brickmanlab/scrnaseq (revision: feature/smart-seq), respectively. Dataset and trained models with parameters were uploaded to Hugging Face at https://huggingface.co/brickmanlab. Publicly available datasets used in this study can be accessed from the National Center for Biotechnology Information Gene Expression Omnibus (GSE57249, GSE80810, GSE74155, GSE45719, GSE53386, GSE100597, GSE123046, GSE84892, GSE44183, GSE159030, GSE148462, GSE136447, GSE36552, GSE171820 and GSE44183) and the European Nucleotide Archive (E-MTAB-2958, E-MTAB-3321, E-MTAB-3929 and PRJEB30442).

Code availability

Data analysis notebooks were uploaded to GitHub at https://github.com/brickmanlab/proks-salehin-et-al. The web portal was uploaded to GitHub at https://github.com/brickmanlab/preimplantation-portal and deployed to https://brickman-preimplantation.streamlit.app.

References

97. Patel, H. et al. nf-core/fetchngs: nf-core/fetchngs v1.10.0 *Zenodo* https://doi.org/10.5281/zenodo.5070524 (2024).
98. Peltzer, A. et al. nf-core/scrnaseq: nf-core/scrnaseq 2.4.1 *Zenodo* https://doi.org/10.5281/zenodo.3568187 (2023).
99. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology* https://doi.org/10.1186/s13059-017-1382-0 (2018).
100. Li, H.-D., Lin, C.-X. & Zheng, J. Gtftools: a software package for analyzing various features of gene models. *Bioinformatics* **38**, 4806–4808 (2022).
101. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
102. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

Acknowledgements

We thank K. Niakan, S. Petropoulos and F. Lanner for valuable comments and feedback. We thank M. Linneberg-Agerholm for providing markers for identifying lineages in preimplantation embryos. Lastly, we thank M. P. Lowndes and M. Perera for critical comments on this paper and members of the Brickman lab for constructive feedback on the analysis and concepts described in this paper. Work in the Brickman laboratory was supported by Lundbeck Foundation (R198-2015-412, R370-2021-617 and R400-2022-769), Independent Research Fund Denmark (DFF-8020-00100B, DFF-0134-00022B and DFF-2034-00025B) and the Danish National Research Foundation (DNRF116) and European Union (ERC, SENCE, 101097979). The Novo Nordisk Foundation Center for Stem Cell Medicine (reNEW) is supported by the Novo Nordisk Foundation, grant number NNF21CC0073729, and previously NNF17CC0027852.

Author contributions

M.P., N.S. and J.M.B. wrote the paper. The computational work has been done by both M.P. and N.S.

Competing interests

The authors declare no competing interests.

Additional information

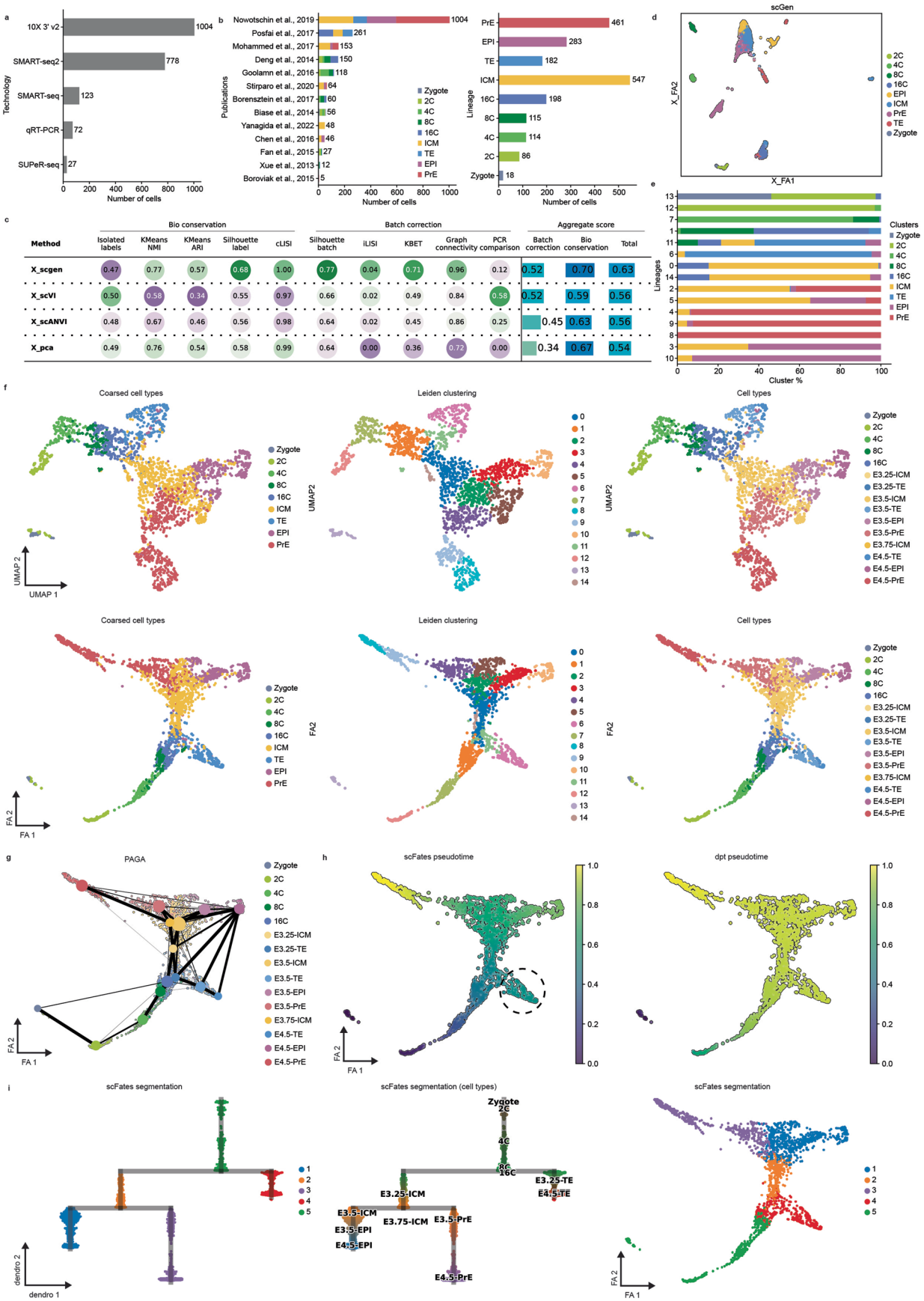
Extended data is available for this paper at https://doi.org/10.1038/s41592-024-02511-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02511-3.

Correspondence and requests for materials should be addressed to Joshua M. Brickman.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

https://doi.org/10.1038/s41592-024-02511-3

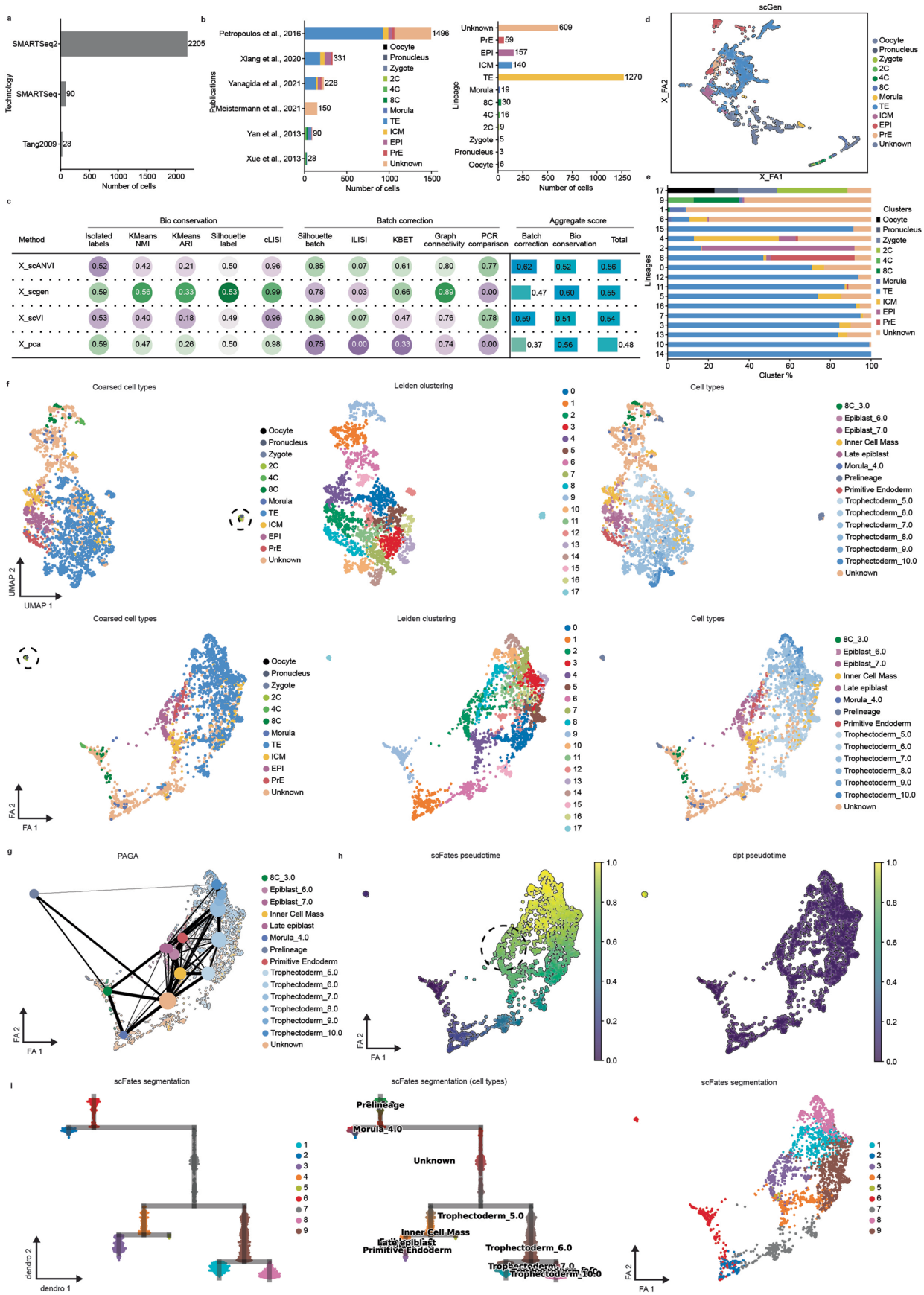


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41592-024-02511-3

Extended Data Fig. 1 | Mouse integration. a) Number of cell types per sequencing technology. **b**) Number of cell types per dataset (left) and lineage (right). **c**) Integration metrics for individual methods. **d**) Force directed graph of latent space inferred using scGEN. **e**) Cell type proportion per identified unsupervised clustering. **f**) Visualization of mouse reference dataset using UMAP (top row) and Force Directed Graph (bottom row) dimensional reduction techniques. **g**) Trajectory inference using PAGA. **h**) Pseudotime inference using scFates (left) and dpt (right) algorithms. **i**) Hierarchical clustering (left, middle) of segmentation (overlaid on Force directed graph representation, right) based on pseudotime from scFates.

https://doi.org/10.1038/s41592-024-02511-3

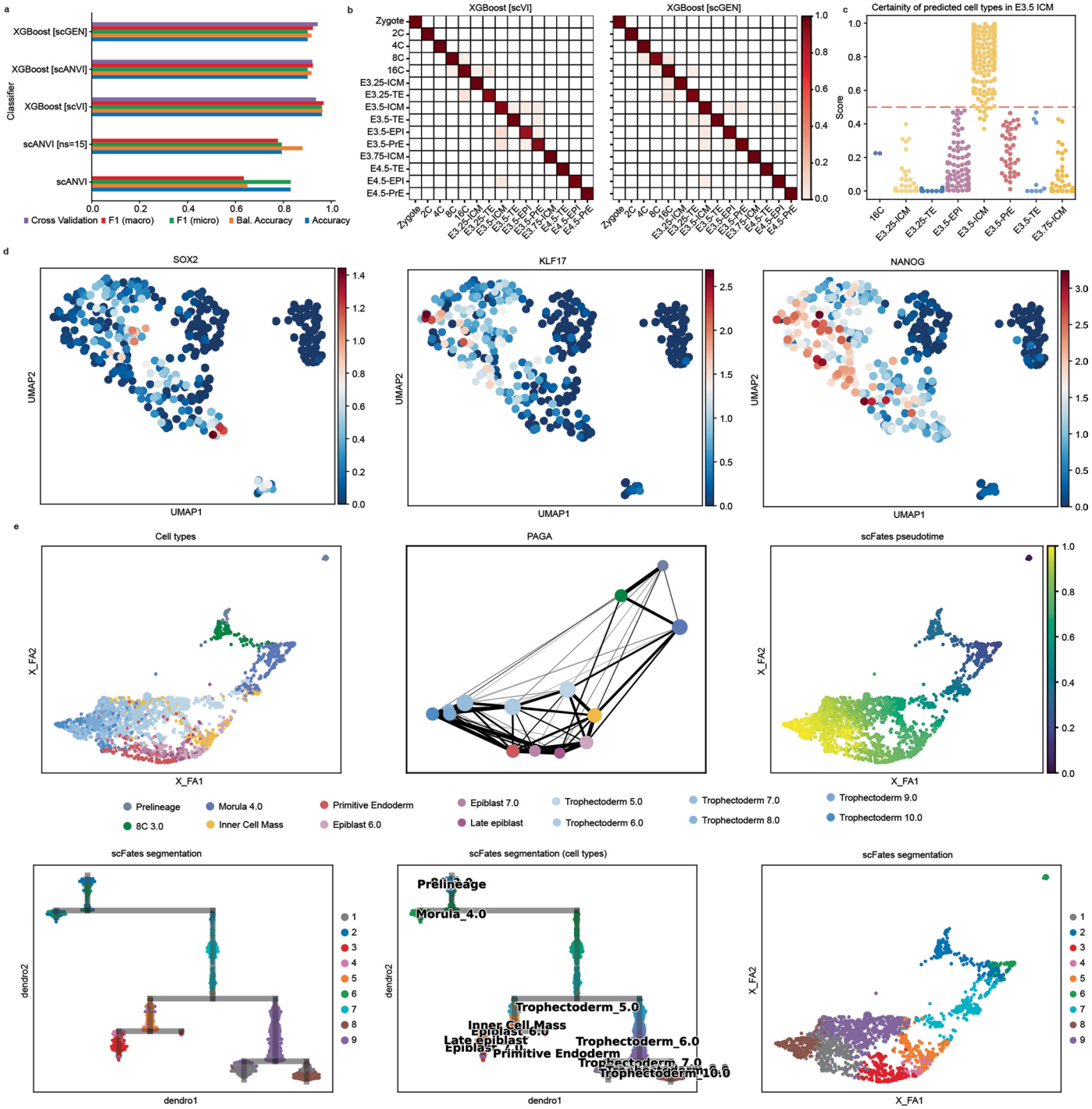


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41592-024-02511-3

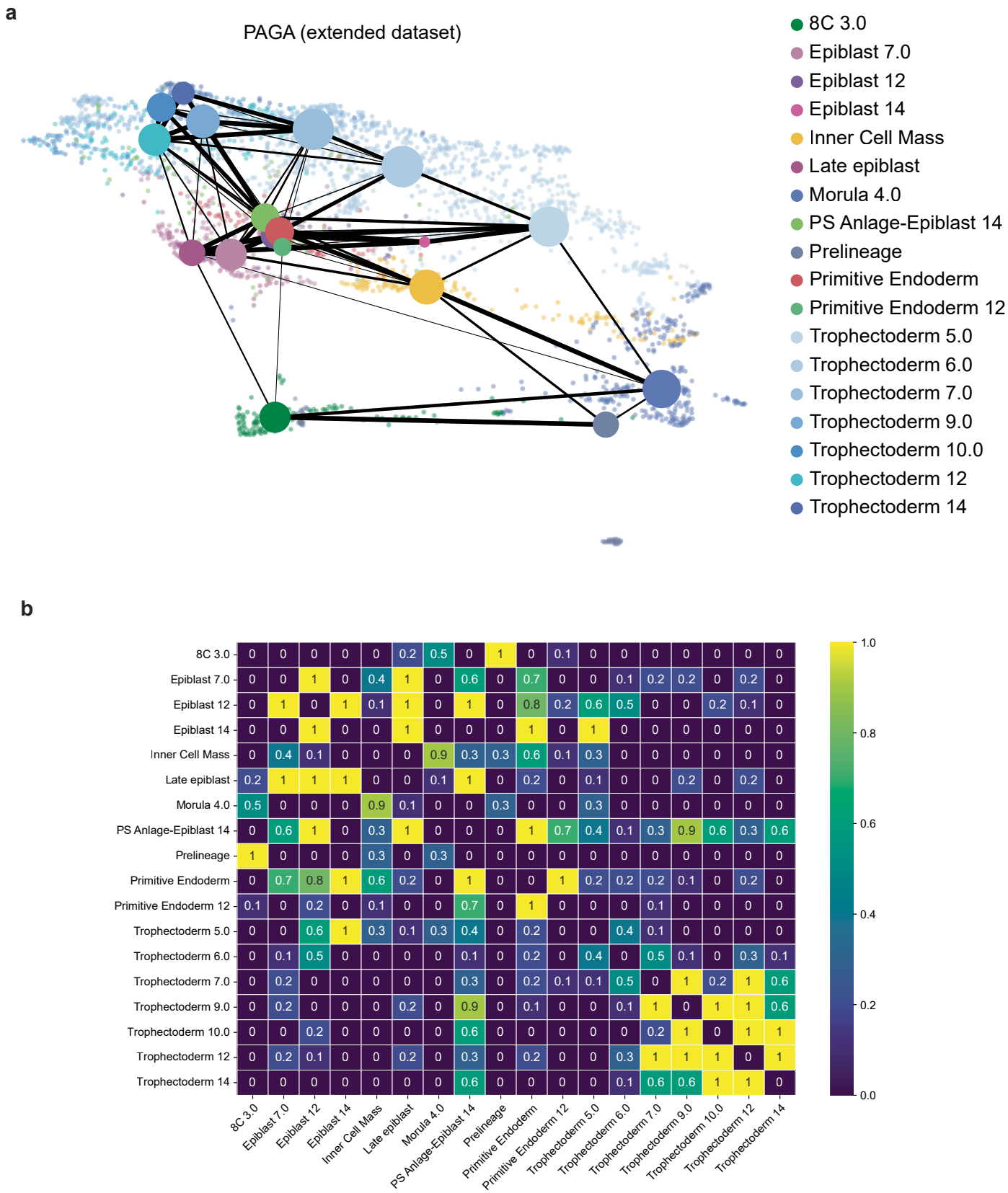
Extended Data Fig. 2 | Human integration. a) Number of cell types per sequencing technology. **b**) Number of cell types per dataset (left) and lineage (right). **c**) Integration metrics for individual methods. **d**) Force directed graph of latent space inferred using scGEN. **e**) Cell type proportion per identified unsupervised clustering. **f**) Visualization of human reference dataset using UMAP (top row) and Force Directed Graph (bottom row) dimensional reduction techniques. **g**) Trajectory inference using PAGA. **h**) Pseudotime inference using scFates (left) and dpt (right) algorithms. **i**) Hierarchical clustering (left, middle) of segmentation (overlaid on Force directed graph representation, right) based on pseudotime from scFates.

https://doi.org/10.1038/s41592-024-02511-3



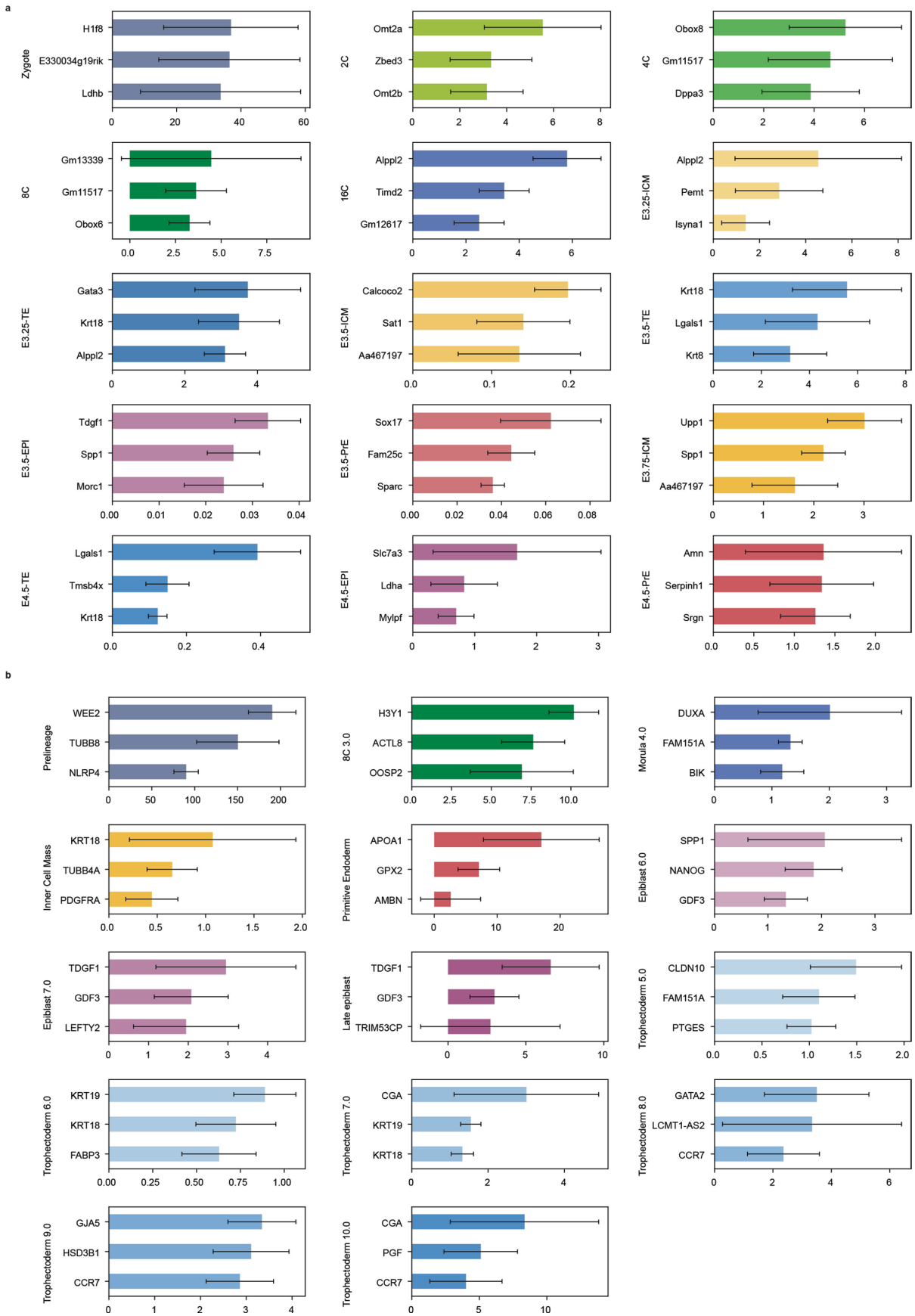
Extended Data Fig. 3 | Classification metrics. a) Accuracy metrics for individual classification algorithms. **b**) Confusion matrix of XGBoost predictions based on denoised expression from either scVI (left) or scGEN (right) models, where y-axis are original annotations and x-axis are predictions. **c**) Maximum certainty of predictions of cells originally annotated as E3.5 ICM in the mouse dataset. **d**) Expression of SOX2, KLF17 and NANOG in human ICM cells reannotated to trophectoderm using the scANVI classifier. **e**) Force directed graph visualisation of the human dataset after reannotation.

https://doi.org/10.1038/s41592-024-02511-3



Extended Data Fig. 4 | Extension of the human model using scRNA-seq data from Days 12 and 14 of development. a) Trajectory inference using PAGA showcasing the correct positioning of the Primitive Streak Anlage cells, emerging from Late Epiblast and Epiblast day 12, based on transcriptome. **b**) Strength of connected link for better inspection of PAGA graph in a).

https://doi.org/10.1038/s41592-024-02511-3



Extended Data Fig. 5 | Feature importance for scANVI classifiers. a) Importance of genes in classification accuracy of individual cell types for the mouse scANVI classifier. **b**) Importance of genes in classification accuracy of individual cell types for the human scANVI classifier. Error bars represent mean ± standard deviation after ten randomised runs.

nature portfolio

Corresponding author(s): Joshua M. Brickman

Last updated by author(s): Aug 30, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	https://github.com/nf-core/fetchngs https://github.com/brickmanlab/scrnaseq/tree/feature/smartseq
Data analysis	https://github.com/brickmanlab/proks-salehin-et-al/tree/master/notebooks cellrank==2.0.1 forceatlas2 gtftools==0.9.0 huggingface-hub==0.20.3 jax==0.4.13 jaxlib==0.4.13+cuda11.cudnn86 natsort==8.4.0 networkx==2.8.8 phate==1.0.11 PhenoGraph==1.5.7 pybiomart==0.2.0 ray==2.5.1 scanpy==1.9.3 scFates==1.0.2 scgen git@06084773e56cad0dec340138441dee47a39af752 scib-metrics==0.3.3 scvelo==0.2.5

scvi-tools==1.0.0
shap==0.41.0
xgboost==1.7.6
XlsxWriter==3.1.9

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

GSE57249
GSE80810
E-MTAB-2958
GSE74155
GSE45719
GSE53386
E-MTAB-3321
GSE100597
GSE123046
GSE84892
GSE44183
GSE159030
GSE148462
PRJEB30442
E-MTAB-3929
GSE136447
GSE36552
GSE171820
GSE44183
Human 8CLC (Mazid et al, 2022) https://figshare.com/s/34110eebb58462a79dd5
Mouse In vitro PrE differentiation (Perera et al, 2022, Proks et al, 2023) GSE200534

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample sizes.
Data exclusions	Data was subset for early preimplantation development, excluding experimental conditions, in vitro stem cells and cells published as

Data exclusions	ambiguous within the parent publications
Replication	Collection of published datasets from unique, sequenced embryos precluded replication.
Randomization	Randomisation was performed by bootstrapping and standard randomised 80:20 splits with replacement by the Python library scikit-learn or custom scripts.
Blinding	Blinding was not required as all data points are used for atlas analysis

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging