Identification and comparison of orthologous cell types from primate embryoid bodies shows limits of marker gene transferability

Jessica Jocher[1,*], Philipp Janssen[1,*], Beate Vieth[1], Fiona C. Edenhofer[1], Tamina Dietl[2], Anita Térmeg[1], Paulina Spurk[1], Johanna Geuder[1], Wolfgang Enard[1,**], Ines Hellmann[1,**]

[1]Anthropology and Human Genomics, Faculty of Biology, Ludwig-Maximilians-Universität München, Germany

[2]Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt: Munich, Germany

[*] contributed equally

[**] correspondence:

Dr. Ines Hellmann, Telefon +49 (0)89 2180-74336

hellmann@bio.lmu.de, www.anthropologie.bio.lmu.de

Keywords

single-cell RNA-sequencing, primates, early development, embryoid bodies, marker genes, cross-species comparison, orthologous cell types

Abstract

The identification of cell types remains a major challenge. Even after a decade of single-cell RNA sequencing (scRNA-seq), reasonable cell type annotations almost always include manual non-automated steps. The identification of orthologous cell types across species complicates matters even more, but at the same time strengthens the confidence in the assignment. Here, we generate and analyze a dataset consisting of embryoid bodies (EBs) derived from induced pluripotent stem cells (iPSCs) of four primate species: humans, orangutans, cynomolgus, and rhesus macaques. This kind of data includes a continuum of developmental cell types, multiple batch effects (i.e. species and individuals) and uneven cell type compositions and hence poses many challenges. We developed a semi-automated computational pipeline combining classification and marker based cluster annotation to identify orthologous cell types across primates. This approach enabled the investigation of cross-species conservation of gene expression. Consistent with previous studies, our data confirm that broadly expressed genes are more conserved than cell type-specific genes, raising the question how conserved - inherently cell type-specific - marker genes are. Our analyses reveal that human marker genes are less effective in macaques and vice versa, highlighting the limited transferability of markers across species. Overall, our study advances the identification of orthologous cell types across species, provides a well-curated cell type reference for future *in vitro* studies and informs the transferability of marker genes across species.

Background

Cell types are a central concept for biology, but are - as other concepts like species - practically difficult to identify. Theoretically, one would consider all stable, irreversible states on a directed developmental trajectory as cell types. In practice, we are limited by our experimental possibilities. Historically, cell type definitions hinged on observations of cell morphology in a tissue context, which was later combined with immunofluorescence analyses of marker genes [1]. A lot of the functional knowledge that we have about cell types today is based on such visual and marker-based cell type definitions. With single cell-sequencing our capabilities to characterize and identify new cell types have radically changed [2, 3]. Clustering cells by their expression profiles enables a more systematic and higher-resolution identification of groups of cells that are then interpreted as cell types. However, distinguishing them from cell states or technical artifacts is not straight forward. A key criterion for defining a true cell type is its reproducibility across experiments, individuals, or even species.

Hence, identifying the same, i.e. orthologous, cell types across individuals and species is crucial. There are three principal strategies to match cell types from scRNA-seq data. 1) One is to integrate all cells prior to performing a cell type assignment on a shared embedding [4]. 2) The second approach is to consider cell types from one species as the reference and transfer these annotations to the other species using classification methods [5]. 3) The third strategy is to assign clusters and match them across species, which has the advantage of not requiring data integration of multiple species or an annotated reference [6, 7, 8].

Furthermore, established marker genes are still heavily used to validate and interpret clusters identified by scRNA-seq data [9, 10, 11]. Together with newly identified transcriptomic markers for human and mouse they are collected in databases [12, 13] and provide the basis for follow-up studies using spatial transcriptomics and/or immunofluorescence approaches. However, previous studies have shown that the same cell types may be defined by different

marker genes in different species [14, 7]. For example, Krienen et al. [15] found that only a modest fraction of interneuron subtype-specific genes overlapped between primates and even less between primate and rodent species.

To better understand how gene expression in general and the expression of marker genes in particular evolves across closely related species, we used induced pluripotent stem cells (iPSCs) and their derived cell types from humans and non-human primates (NHP). One fairly straight forward way to obtain diverse cell types from iPSCs are embryoid bodies (EBs). EBs are the simplest type of iPSC-derived organoids, contain a dynamic mix of cell types from all three germ layers and result from spontaneous differentiation upon withdrawal of key pluripotency factors [16, 17, 18, 19, 20].

EBs and brain organoids from humans and chimpanzees have for example been used to infer human-specific gene regulation in brain organoids [21] or to investigate mechanisms of gene expression evolution [22].

Here we explore to what extent levels of cell type specificity of marker genes are conserved in primates. We generated scRNA-seq data of 8 and 16 day old EBs from human, orangutan (*Pongo abelii*), cynomolgus (*Macaca fascicularis*) and rhesus macaque (*Macaca mulatta*) iPSCs. Using this data, we established an analysis pipeline to identify and assign orthologous cell types. With this annotation we provide a well curated cell type reference for *in vitro* studies of early primate development. Moreover, it allowed us to asses the cell type-specificity and expression conservation of genes across species. We find that even though the cell type-specificity of a marker gene remains similar across species, its discriminatory power still decreases with phylogenetic distance.

Results

Generation of embryoid bodies from iPSCs of different primate species

We generated EBs from iPSCs across multiple primate species: two human iPSC clones (from two individuals), two orangutan clones (from one individual), three cynomolgus clones (from two individuals), and three rhesus clones (from one individual) [23, 24, 25]. To optimize conditions for generating a sufficient number of cells from all three germ layers across these four species, we tested combinations of two culturing media ("EB-medium" and "DFK20", see Methods) and two EB-differentiation conditions ("single-cell seeding" and "clump seeding", see Methods). After 7 days of differentiation, germ layer composition was analyzed by flow cytometry (Supplementary Figure S1A,B,C). Among the four tested protocols, culture in DFK20-medium with clump seeding resulted in the most balanced representation of all germ layers, yielding a substantial number of cells from each layer across all species (Supplementary Figure S1D).

Under these conditions, we established an EB formation protocol based on 8 days of floating culture in dishes, followed by 8 days of attached culture (Figure 1A). This results in the formation of cells from all three germ layers, as confirmed by immunofluorescence staining for AFP (endoderm), β-III-tubulin (ectoderm) and α-SMA (mesoderm) (Figure 1B). To generate scRNA-seq data, we dissociated 8 or 16 day old EBs into single cells and pooled cells from all four species to minimize batch effects (Figure 1C). We performed the experiment in three independent replicates, generating a total of four lanes and six lanes of 10x Genomics scRNA-seq at day 8 and day 16, respectively (Supplementary Figure S2A). This resulted in a dataset comprising over 85,000 cells after filtering and doublet removal, distributed fairly equally over time points, species and clones (Supplementary Figure S2B-D).

In agreement with the immunofluorescence staining, we detected well-established marker genes of pluripotent cells and of all three germ layers [26] in the scRNA-seq data: *SOX2*,

SOX10, and *STMN4* expression was used to label ectodermal cells, *APOA1* and *EPCAM* for endodermal cells, *COL1A1* and *ACTA2* (α-SMA) for mesodermal cells, and *POU5F1* and *NANOG* for pluripotent cells (Figure 1D). Expression of these marker genes corresponded well with a classification based on a published scRNA-seq dataset from 21 day old human EBs [18]. This initial, rough germ layer assignment shows that our differentiation protocol generates EBs with the expected germ layers and cell type diversity from all four species (Figure 1E,Supplementary Figure S3A).

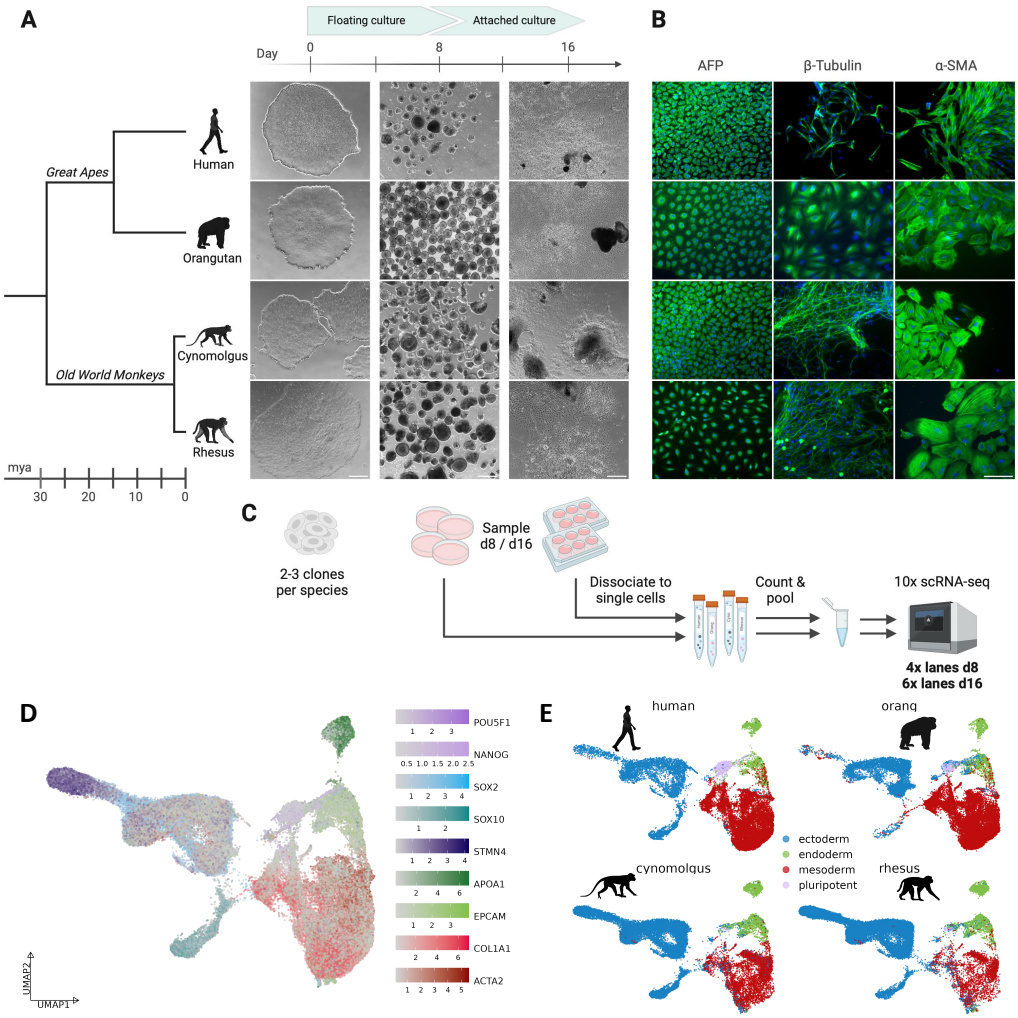


Figure 1. Generation of primate embryoid bodies. A) Overview about the EB differentiation workflow of the four primate species human (*Homo sapiens*), orangutan (*Pongo abelii*), cynomolgus (*Macaca fascicularis*) and rhesus (*Macaca mulatta*), including their phylogenetic relationship. Scale bar represents 500 µm. B) Immunofluorescence staining of day 16 EBs using α-fetoprotein (AFP), β-III-tubulin and α-smooth muscle actin (α-SMA). Scale bar represents 100 µm. C) Schematic overview of the sampling and processing steps prior to 10x scRNA-seq. D) UMAP representation of the whole scRNA-seq dataset, integrated across all four species with Harmony. Single cells are colored by the expression of known marker genes for the three germ layers and undifferentiated cells. E) UMAP representation, colored by assigned germ layers, split by species.

Assignment of orthologous cell types

Many integration methods encounter difficulties when they are applied to data from multiple species and uneven cell type compositions [4]. Indeed, when comparing clusters derived from an integrated embedding across all species [27, 28] to the aforementioned preliminary cell type assignments, we observed signs of overfitting. For instance, a cluster predominantly containing cells classified as neurons in humans, cynomolgus, and rhesus macaques consisted mainly of early ectoderm and mesoderm cells in orangutans (Supplementary Figure S3B,C). To address this issue, we developed an approach that assigns orthologous cell types without a common embedding space in an interactive shiny app (`https://shiny.bio.lmu.de/Cross_Species_CellType/`; Figure 2A, B):

First, we assign cells to clusters separately for each species. To avoid losing rare cell types, we aim to obtain at least double as many high resolution clusters (HRCs) per species as expected cell types. We then use the HRCs of one species as a reference to classify the cells of the other species using SingleR [29]. These pair-wise comparisons are done reciprocally for each species and via a cross-validation approach also within each species (see Methods). For each comparison, we average the two values for the fraction of cells annotated as the other HRC. For example, a perfect "reciprocal best-hit" between HRC-A in human and HRC-B in rhesus would have all cells of HRC-B assigned to HRC-A when using the human as a reference and reciprocally all cells in HRC-A assigned to HRC-B when using the rhesus as a reference. Next, we used the resulting distance matrix as input for hierarchical clustering to find orthologous clusters across species and merge similar clusters within species. Here, the user can choose and adjust the final cell type cluster number. This allows us to identify orthologous cell type clusters (OCCs) across all four species, while retaining species-specific clusters when no matching cluster was identified.

In the last steps, OCCs are manually further refined by merging neighboring OCCs with similar marker gene and transcriptome profiles (see Methods). To avoid bias, we first identify

marker genes independently for each species solely based on scRNA-seq expression data [30]. We then intersect those lists to identify the top ranking marker genes with consistently good specificity across all species. The final set of conserved marker genes then serves us to derive cell type labels by searching the literature as well as databases of known marker genes (Figure 2E). If the marker-gene based cell type assignment reveals cluster inconsistencies, they can be marked for further splitting. This feature is of particular importance for rare cell types. For example, we separated a cluster of early progenitor cells into iPSCs, cardiac progenitors, and early epithelial cells.

Suresh et al. [8] devised a conceptually similar approach to ours to identify orthologous cell types across species. The main difference is that they used scores from MetaNeighbor [6] where we use SingleR to measure distances between HRCs. However, in essence both scores are based on rank correlations and hence it may not be surprising that both scoring systems yield consistent cluster groupings that show high replicability across species. However, using our SingleR-based scores to compare OCCs across species may yield more clearly defined correspondences compared to MetaNeighbor scores (Supplementary Figures S5 and S4).

Overall, we are confident that our approach yields meaningful orthologous cell type assignments, without requiring a prior annotation per species or a reference dataset. Moreover, the necessary fine tuning of the cell type clusters by the expert user is facilitated by an interactive app.

Many cell types are shared between day 8 and day 16 EBs

Using the strategy described in the previous section, we detected a total of 15 reproducible cell types from the three germ layers, all of which were detected in at least 3 cell lines in 3 independent replicates. Among these, we identified four cell types that represent the latest time points along ectodermal developmental lineages (astrocyte progenitor, granule precursor, neurons, neural crest II), four that represent the latest time points along mesodermal lineages

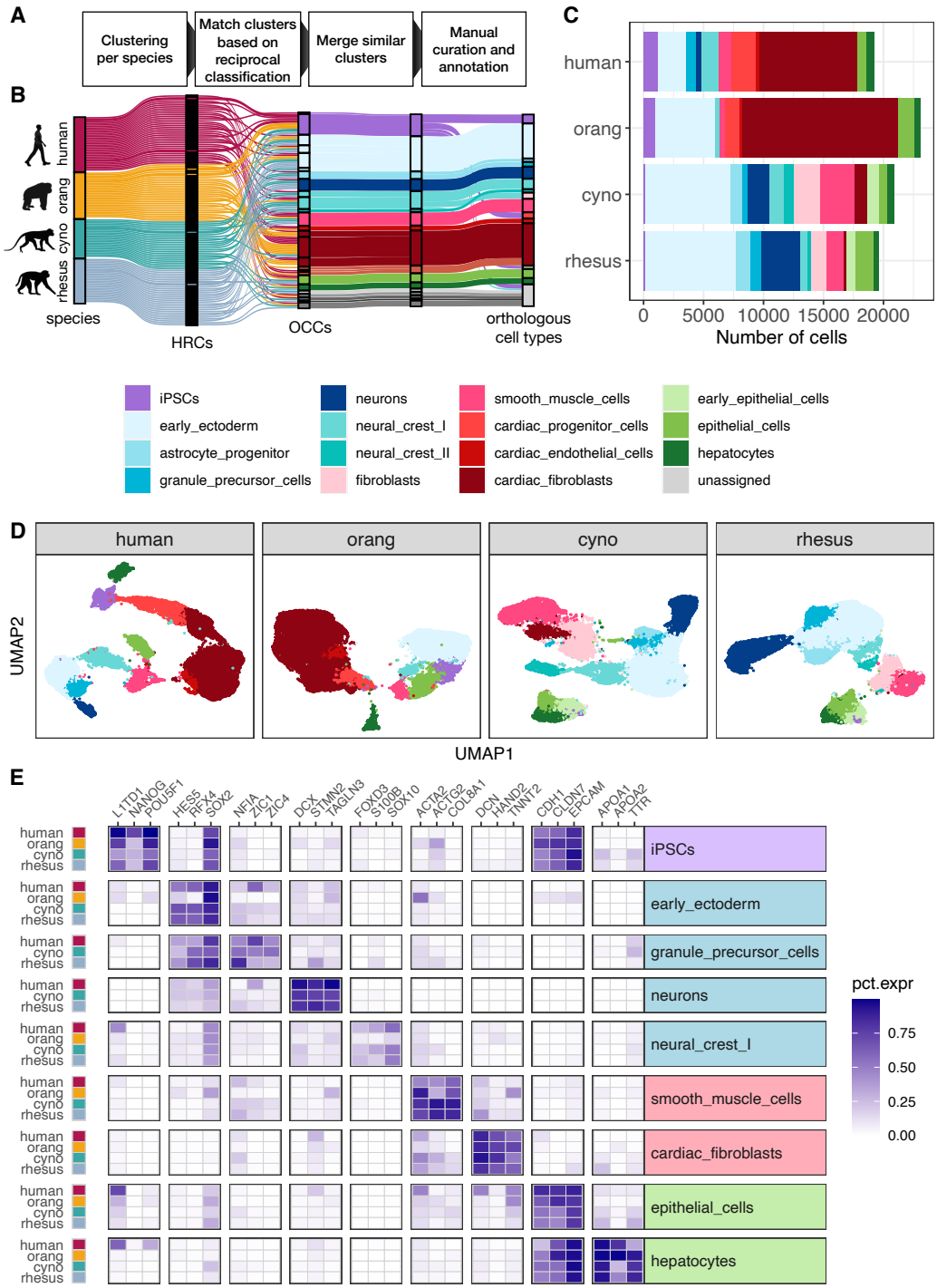


Figure 2. Assignment of orthologous cell types across species. A) Schematic overview of the pipeline to match clusters between species and assign orthologous cell types. B) Sankey plot visualizing the intermediate steps of the cell type assignment pipeline. Each line represents a cell which are colored by their species of origin on the left and by their current cell type assignment during the annotation procedure on the right. An initial set of 118 high resolution clusters (HRCs), 25-35 per species, was combined into 26 orthologous cell type clusters (OCCs). Similar cell type clusters were merged and after further manual refinement provided the basis for final orthologous cell type assignments. C) Fraction of annotated cell types per species. D) UMAPs for each species colored by cell type. E) To validate our cell type assignments, we selected three marker genes per cell type that exhibit a similar expression pattern across all four species and have been reported to be specific for this cell type in both human and mouse (Supplementary Table S1). The heatmap depicts the fraction of cells of a cell type in which the respective gene was detected for cell types present in at least three species.

(fibroblasts, smooth muscle cells, cardiac endothelial cells, cardiac fibroblasts), and two that represent the latest detected time points along endodermal lineages (epithelial cells, hepatocytes). Many of these cell types were present at both sampling times (Supplementary Figure S6C). The most notable exception is that orangutan EBs lost the majority of ectodermal cells at the later time point. Aside from this technical deviation—likely caused by the additional handling step (see previous chapter)—some more differentiated cell types only appear at day 16 at appreciable frequencies. This is most pronounced for smooth muscle cells in all species, but also holds for neuron-like cells in humans. Overall, this leads to an increase in the observed cell type diversity over time.

To further evaluate differences between the two sampling time points, we performed pseudotime analyses[31] on the experiments integrated per species and germ layer, defining iPSCs as the origin and the differentiated cell types listed above as the endpoints of the developmental trajectories (Supplementary Figures S7-S9). As expected, day 16 cells generally occupy later positions along the trajectories than day 8 cells, yet the distributions overlap: iPSCs and precursor states such as early ectoderm are still detectable, albeit at lower frequency, in the day 16 EBs. Still, the few states that are confined to one of the two time points improve cross-species comparability when both are considered jointly. Integrating day 8 and day 16 increased the overlap in detected cell types between species; for example, human neural cells were only observed at day 16, whereas they were already present at day 8 in macaques, and we therefore used the combined data from both time points for downstream analyses.

Overall, 9 of the 15 cell types were detected in at least 3 species, and 7 cell types were reproducibly detected in all four species (Figure 2C, D; Supplementary FigureS6). These 7 cell types consisted of iPSCs, two cell types representing ectoderm: early ectoderm and neural crest, two cell types of mesodermal origin: smooth muscle cells and cardiac fibroblasts and two endodermal cell types: epithelial cells and hepatocytes (Figure 2C,E) and are used

for the analysis of pleiotropy and marker genes in the remainder of this manuscript.

Cell type-specific genes have less conserved expression levels

Based on the premise that it is not necessarily the expression level, but rather the expression breadth that determines expression conservation [32], we developed a method to call a gene 'expressed' or not that considers the expression variance across the cells of one type, which we then used to score cell type-specificity and expression conservation (Figure 3B); see Methods).

For example, we find that the neural crest-marker *SOX10* [33] is cell type-specific and conserved, the lncRNA *ESRG* is iPSC- and human-specific, in contrast *RPL22*, a gene that encodes a protein of the large ribosomal subunit, is broadly expressed and conserved (Figure 3A). Overall we find on average ∼15% of genes to be cell type-specific, i.e. our score determined them to be expressed in only one cell type, while ∼40% of genes were found to be broadly expressed in all seven cell types (Supplementary Figure S10A).

Additionally, we obtained a measure of expression conservation, which quantifies the consistency of the cell type expression score across species. We found that broadly expressed genes present in all cell types exhibited high expression conservation, whereas cell type-specific genes tended to be more species-specific (Figure 3C; Supplementary Figure S10B).

Unsurprisingly, broadly expressed genes also showed higher average expression levels [34] (Supplementary Figure S10D). To ensure that the observed relationship between expression breadth and conservation in our data is not solely due to expression level differences, we sub-sampled genes from all cell type-specificity levels for comparable mean expression. This did not change the pattern: also broadly expressed genes with a low mean expression level are highly conserved across species (Supplementary Figure S10E,F). Moreover, also the coding sequences of broadly expressed genes show higher levels of constraint than more cell type-specific genes, thus supporting the notion that also the higher conservation of the

expression pattern that we observed here is due to evolutionary stable functional constraints on this set of genes (Figure 3D; Supplementary Figure S10C).

Marker gene conservation

Building on our previous observation that cell type-specific genes are less conserved across species, we investigated the conservation and transferability of marker genes, which are, by definition, cell type-specific, in greater detail. To this end, we call marker genes for all cell types and species, using a combination of differential expression analysis and a quantile rank-score based test for differential distribution detection[36]. Additionally, we define a good marker gene as one that is upregulated and expressed in a higher fraction of cells compared to the rest. To prioritize marker genes, we rank them based on the difference in the detection fraction: the proportion of cells of a given type in which a gene is detected compared to its detection rate in all other cells.

We found a low overlap of top marker genes among species, with a median of 15 of the top 100 ranked marker genes per cell type shared across all four species, while a larger proportion of markers was unique to individual species (Figure 4A). Notably, these species-specific markers often exhibited cell type-specific expression in only one species, with reduced or non-specific expression in others (Figure 4B; Supplementary Figure S11).

Given the special role of transcriptional regulators for the definition of a cell type [37] and the differences in conservation between protein-coding and non-coding RNAs [38], we analyzed the comparability of marker genes of different types. To this end, we assessed the concordance of the top 100 marker genes across species for protein-coding genes, lncRNAs, transcription factors (TFs) or all genes using rank biased overlap (RBO) scores [39]. We find that marker genes that are TFs have the highest concordance between species and that the two macaques species which are also phylogenetically most similar are also most similar in their ranked marker gene lists. In contrast, lncRNA markers show the lowest overlap

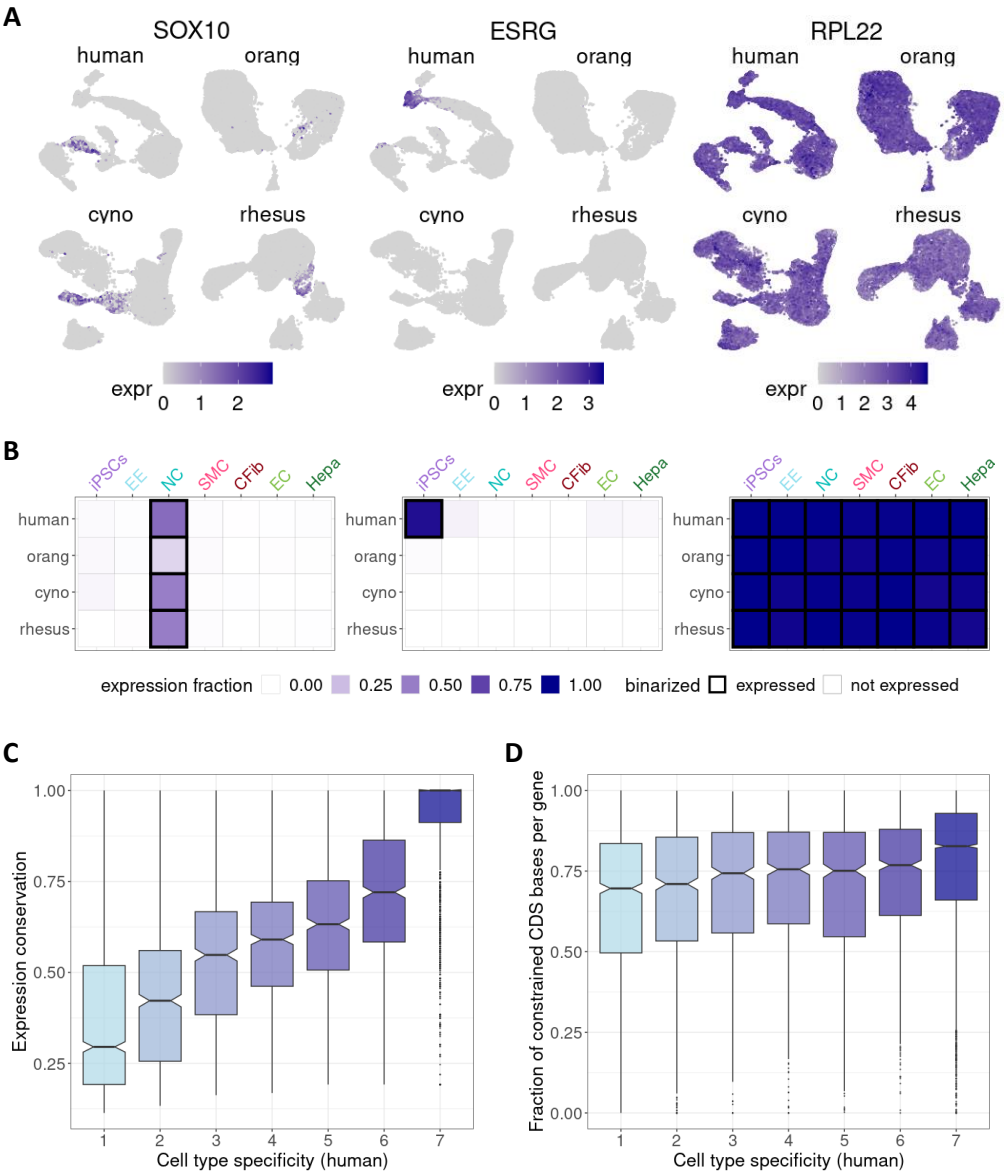


Figure 3. Effect of cell type specificity on expression conservation. A) UMAP visualizations depicting expression patterns of selected example genes: *SOX10* (conserved cell type-specific expression in neural crest cells), *ESRG* (species-specific and cell type-specific expression in human iPSCs), and *RPL22* (conserved, broad expression). B) For each gene, expression was summarized per species and cell type as the expression fraction and binarized into "not expressed"/"expressed" (black frame) based on cell type-specific thresholds. The same example genes as in A) are shown here. iPSCs: induced pluripotent stem cells, EE: early ectoderm, NC: neural crest, SMC: smooth muscle cells, CFib: cardiac fibroblasts, EC: epithelial cells, Hepa: hepatocytes. C) Boxplot of expression conservation of genes according to the number of different cell types in which a gene is expressed in humans (cell type specificity). D) Boxplot of the fraction of coding sequence sites that were found to evolve under constraint based on a 43 primate phylogeny [35], stratified by human cell type specificity.

between species. In fact, their cross-species conservation is so low that they also significantly reduce the performance if they are included together with protein-coding markers (Figure 4C).

To properly evaluate the performance of marker genes, it is essential to consider their ability to differentiate between cell types. This discriminatory power ultimately determines how well marker genes perform in cell type classification within and across species. To this end, we trained a k-nearest neighbors (kNN) classifier on varying numbers of marker genes per cell type in one human clone (29B5) and evaluated prediction performance using the average F1-score across cell types (Supplementary Figure S12). Again, we analyzed markers from a set of all protein-coding genes and TFs only and find that even though TFs appear to be more conserved across species, they do not discriminate cell types as well as the top protein-coding markers (Supplementary Figure S13). Using protein-coding marker genes only determined with 29B5 to classify the other human clone, we achieve good discriminatory power (F1 score > 0.9) with only 11 marker genes per cell type. In contrast, the classification performance for clones from the other species was substantially lower, failing to reach the performance levels observed in human clones even when using up to 30 marker genes (Figure 4D).

In summary, we find that lncRNA markers genes have low transferability between species, while protein-coding markers do reasonably well. However, the predictive value of marker genes decreases with increasing phylogenetic distance, requiring longer marker gene lists to achieve accurate cell type classification for more distantly related species.

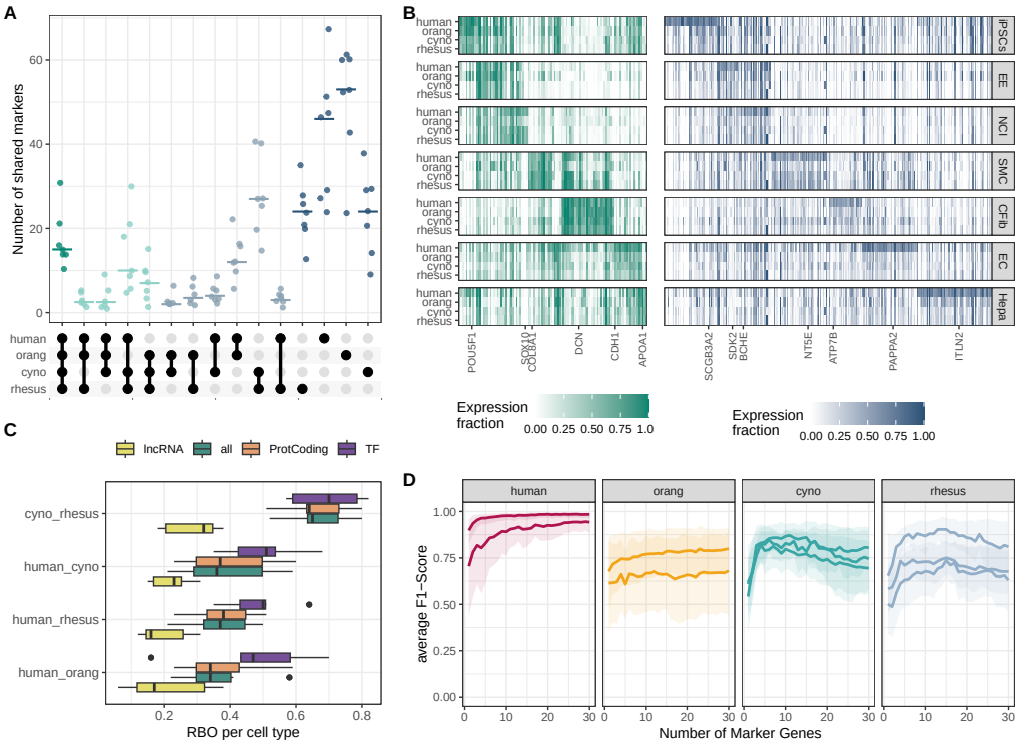


Figure 4. Evaluation of marker gene conservation. A) UpSet plot illustrating the overlap between species for the top 100 marker genes per cell type. B) Heatmap showing the expression fractions of marker genes: on the left, markers shared among all species, and on the right, markers unique to the human ranking. For each cell type, one representative gene is labeled and further detailed in Supplementary Figure S11. iPSCs: induced pluripotent stem cells, EE: early ectoderm, NC: neural crest, SMC: smooth muscle cells, CFib: cardiac fibroblasts, EC: epithelial cells, Hepa: hepatocytes. C) Rank-biased overlap (RBO) analysis comparing the concordance of gene rankings per cell type for lncRNAs, protein-coding genes and transcription factors. D) Average F1-score for a kNN-classifier trained in the human clone 29B5 to predict cell type identity based on the expression of 1-30 marker genes. Each line represents the performance in a different clone, with shaded areas indicating 95% bootstrap confidence intervals.

Discussion

An essential criterion for a true cell type is reproducibility across experiments, individuals, or even species. This raises the question of how to reliably identify reproducible cell types across species. When cell types are annotated separately for each species, their reproducibility can be evaluated based on transcriptomic similarity [6, 40]. If integration-based methods are used to accomplish this task [22, 7], reproducibility not only depends on the similarity of the expression profiles but also on cell type composition. Integration works best when the cell type compositions are as similar as possible across experiments. This however is not the case for organoids, which often have highly heterogeneous cell type compositions

[41] and our EB-data are no exception. Moreover, integration methods struggle with large and variable batch effects, which are expected due to the varying phylogenetic distances across species [4]. In contrast, classification methods such as SingleR [29] rely mainly on the similarity to a reference profile, which makes it less vulnerable to cell type composition and batch effects. Hence, in our pipeline to identify orthologous cell types we mainly rely on classification. We start with an unsupervised approach in that we identify cell clusters and then ensure reproducibility as well as comparability using a supervised approach with reciprocal classification of clusters across all species pairs.

Defining cell types in a developmental dataset is particularly challenging, and we do not believe that there is one perfect solution that would fit all cell types and samples. Therefore, we rely on an interactive approach that we implemented in a shiny app (`https://shiny.bio.lmu.de/Cross_Species_CellType/`) to facilitate the flexible choice of parameters for cluster matching, merging and inspection by visualizing marker genes. Suresh et al [8] employed a similar approach also requiring several manual parameter choices. This makes a formal comparison difficult. Generally both methods seem to agree well on the orthology assignments of cell type clusters (Supplementary Figures S5& S4). MetaNeighbor, as used by Suresh et al. [8], provides a more quantitative and potentially more sensitive framework for assessing cross-species cell type relationships. However, this higher sensitivity may also make it more affected by data with a lower signal-to-noise ratio, such as our developmental time series.

Hence, the carefully annotated dataset presented here can serve as a valuable resource for future research. Non-human primate iPSCs are central to many studies focusing on evolutionary comparisons, and the pool of iPSC lines for these purposes is expected to grow, incorporating more species and individuals. In this context, the transcriptomic data we generated offer a reference dataset that can be used to verify the pluripotency and differentiation potential of non-human primate iPSC lines by examining gene expression

during EB formation.

The set of shared cell types between all four primate species allowed us to evaluate the conservation and transferability of marker genes between species. To begin with, marker genes are by definition cell type-specific and also with this dataset, we can show that they are less conserved than broadly expressed genes. Expression breadth can be interpreted as a sign of pleiotropy and hence higher functional constraint [42, 32]. Conversely, we expect cell type-specific marker genes to be among the least conserved genes. Indeed, we and others find that the overlap of marker genes across species is limited [14, 15, 7, 43]. Moreover, conservation varies significantly across gene biotypes. On the one hand, lncRNAs, which are often highly cell type-specific, exhibit lower cross-species conservation. Their low sequence conservation further complicates their utility for comparative studies [38]. On the other hand, TFs, which have been proposed as central elements of a Core Regulatory Complex (CoRC) that defines cell type identity [37], are among the most conserved markers across species. However, the power to distinguish cell types based solely on the expression of TF markers remains lower than when markers are selected from the broader set of all protein-coding genes (Supplementary Figure S13). Even though within species already a handful of marker genes can achieve remarkable accuracy, their discriminatory power remains lower for other species. Thus, whole transcriptome profiles offer a more comprehensive approach to cross-species cell type classification for single cell data.

This said, marker genes remain fundamental to most current cell type annotations. Moreover, marker genes will continue to be used to match cell types across modalities, as for example to validate cell type properties in experiments that are often based on immunofluorescence of individual markers or gene panels as used for spatial transcriptomics [44, 45]. To this end, we have refined the ranking of marker genes beyond differential expression analysis to focus on consistent differences in detection rate. Markers identified in this way are bound to translate better into protein-based validations than markers defined

based on expression levels, due to the discrepancy of mRNA and protein expression [46]. Furthermore, the presence-absence signal is more robust against cross-species fluctuations in gene expression than measures based on expression level differences.

In conclusion, we present a robust reference dataset for early primate development alongside tools to identify and evaluate orthologous cell types. Our findings emphasize the need for caution when transferring marker genes for cell type annotation and characterization in cross-species studies.

Methods

EB differentiation method comparison

Four EB differentiation protocols are compared initially, which are combinations of two differentiation media (DFK20 and EB-medium) and two differentiation methods (dish and 96-well).

For single-cell differentiation in 96-well plates, primate iPSCs from one 80% confluent 6-well are washed with DPBS and incubated with Accumax (Sigma-Aldrich, SCR006) for 7 min at 37 °C. Afterwards, iPSCs are dissociated to single-cells, the enzymatic reaction is stopped by adding DPBS, and cells are counted and pelleted at 300 xg for 5 min. Single cells are resuspended in EB-medium consisting of StemFit Basic02 (Nippon Genetics, 3821.00) w/o bFGF or DFK20, both supplemented with 10 µM Y-27632 (Biozol, ESI-ST10019). The DFK20-medium consists of DMEM/F12 (Fisher Scientific, 15373541) with 20% KSR (Thermo Fisher Scientific, 10828-028), 1% MEM non-essential amino acids (Thermo Fisher Scientific, 11140-035), 1% Glutamax (Thermo Fisher Scientific, 35050038), 100 U/mL Penicillin, 100 µg/mL Streptomycin (Thermo Fisher Scientific, 15140122) and 0.1 mM 2-Mercaptoethanol (Thermo Fisher Scientific, M3148). Afterwards, 9,000 cells in 150 µl medium are seeded per well of a Nuclon Sphera 96-well plate (Fisher Scientific, 15396123) and cultured at 37 °C and 5% CO_2. A medium change with the corresponding EB differentiation medium w/o Rockinhibitor is performed every other day during the whole protocol. EBs are collected from the 96-well plate and subjected to flow cytometry after 7 days of differentiation.

For clump differentiation in culture dishes, primate iPSCs from one 80% confluent 12-well are washed with DPBS and incubated with 0.5 mM EDTA (Carl Roth, CN06.3) for 3-5 min at RT. The EDTA is removed, StemFit (Nippon Genetics, 3821.00) supplemented with 10 µM Y-27632 (Biozol, ESI-ST10019) is added and cells are dissociated to clumps of varying sizes. Subsequently, the clumps are transferred to sterile bacterial dishes with vents and

cultured at 37 °C and 5% CO_2. After 24 h, the medium is exchanged by either EB-medium or DFK20 supplemented with 10 µM Y-27632 for additional 24 h, before changing the medium to EB-medium or DFK20. A medium change is performed every other day during the protocol from day 4 on. EBs are collected from the dishes and subjected to flow cytometry after 7 days of differentiation.

Flow cytometry

Flow cytometry is performed on day 7 of the differentiation protocol. Therefore, 1/10 of the EBs are collected, washed with DPBS, incubated with Accumax (Sigma-Aldrich, SCR006) for 10 min at 37 °C and dissociated to single cells. After washing, cells are incubated with the Viability Dye eFluor 780 (Thermo Fisher Scientific, 65-0865-18) diluted 1/1000 in PBS for 30 min at 4°C in the dark. The live/dead stain is quenched by the addition of Cell Staining Buffer (CSB) consisting of DPBS with 0.5% BSA (Sigma-Aldrich, A3059), 0.01% NaN_3 (Sigma-Aldrich, S2002) and 2 mM EDTA (Carl Roth, CN06.3). Subsequently, cells are pelleted and incubated with a mixture of the following antibodies diluted 1/200 in CSB for 1h at 4°C in the dark. The antibodies used are anti-TRA-1-60-AF488 (STEMCELL Technologies, 60064AD.1), anti-CXCR4-PE (BioLegend, 306505), anti-NCAM1-PE/Cy7 (BioLegend, 318317) and anti-PDGFRα-APC (BioLegend, 323511). After centrifugation, cells are resuspended in PBS containing 0.5% BSA, 0.01% NaN_3 and 1 µg/ml DNase I (STEMCELL Technologies, 07469), filtered through a strainer and analyzed using the BD FACSCanto Flow Cytometry System. Flow cytometry data are analyzed using FlowJo (V10.8.2).

In-vitro embryoid body differentiation

Two human, two orangutan, three cynomolgus and three rhesus iPSC lines are used for EB differentiation. The human and orangutan iPSCs are reprogrammed from urinary cells,

while cynomolgus and rhesus iPSCs were reprogrammed from fibroblasts. All cell lines were characterized and validated previously and were tested negative for mycoplasma and SeV reprogramming vector integration [23, 24, 25].

For embryoid body formation prior to 10x scRNA-seq, the EB differentiation protocol using DFK20 medium in culture dishes is performed in duplicates for each clone. After 8 days of floating culture in dishes, EBs from both replicates are pooled and seeded into 6-wells coated with 0.2% gelatin (Sigma-Aldrich, G1890) for another 8 days of attached culture with subsequent medium changes every other day. In total, three replicates of EB formation are performed on different days, and each replicate includes cell lines from all four primate species.

scRNA-seq library generation and sequencing

EBs are sampled on day 8 and day 16 of the protocol. For dissociation, floating EBs are collected, while attached EBs are kept in their wells, washed with DPBS and incubated with Accumax (Sigma-Aldrich, SCR006) for 10-20 min at 37 °C. Afterwards, EBs are pipetted up and down with a p1000 pipette until they are completely dissociated. The enzymatic reaction is stopped by adding DFK20 medium, cells are pelleted at 300 xg for 5 min and resuspended in 1 mL DPBS. If cell clumps are observed, the liquid is filtered through a 40 µm strainer before counting them with a Countess II automated cell counter (Thermo Fisher Scientific, C10228). Equal cell numbers from each cell line are pooled, washed with DPBS + 0.04% BSA and resuspended in DPBS + 0.04% BSA aiming for a final concentration of 800 - 1000 cells/µL. scRNA-seq libraries are generated using the 10x Genomics Chromium Next GEM Single Cell 3´Kit V3.1 workflow in three replicates. Each time, evenly pooled single cells from the different cell lines are loaded on 2 to 6 lanes of a 10x chip, targeting 16,000 cells per lane. Libraries are sequenced on an Illumina NextSeq1000/1500 with an 100-cycle kit and the following sequencing setup: read 1 (28 bases), read 2 (10 bases), read

3 (10 bases) and read 4 (90 bases).

Alignment of scRNA-seq data

Reads are processed with Cell Ranger version 7.0.0. We map all reads to 4 reference genomes: *Homo sapiens* GRCh38 (GENCODE release 32), *Pongo abelii* Susie_PABv2/ponAbe3, *Macaca fascicularis* macFas6 and *Macaca mulatta* rheMac10. The orangutan, cynomolgus macaque and rhesus macaque GTF files are created by transferring the hg38 annotation to the corresponding primate genomes via the tool Liftoff [47], followed by removal of transcripts with partial mapping (<50%), low sequence identity (<50%) or excessive length (>100 bp difference and >2 length ratio) for all species.

Species and individual demultiplexing

Since we pool cells from multiple species on each 10x lane, we use cellsnp-lite [48] version 1.2.0 and vireo [49] version 0.5.7 to assign single cells to their respective species. Initially, we obtain a list of 51000 informative variants (referred to as 'species vcf file') from a bulk RNA-seq experiment involving samples from *Homo sapiens*, *Pongo abelii* and *Macaca fascicularis*, mapped to the GRCh38 reference genome. We run cellsnp-lite in mode 2b for whole-chromosome pileup and filter for high-coverage homozygous variants to identify informative variants.

For the demultiplexing of species in the scRNA-seq data we employ a two step strategy:

1) Initial species assignment: Using the Cell Ranger output aligned to GRCh38, we genotype each single cell with cellsnp-lite providing the species vcf file as candidate SNPs and setting a minimum UMI count filter of 10. Subsequently we assign single cells to human, orangutan or macaque identity with vireo using again the species vcf file as the donor file.

2) Distinguishing macaque species: To differentiate between the two macaque species, *Macaca fascicularis* and *Macaca mulatta*, we use the Cell Ranger output aligned to rheMac10.

After genotyping with cellsnp-lite we demultiplex with vireo specifying the number of donors to two, without providing a donor vcf file in this case. We assign the donor, for which the majority of distinguishing variants agreed with the rheMac10 reference alleles, to *Macaca mulatta* and the other donor to *Macaca fascicularis*.

To distinguish different human individuals pooled in the same experiment, we genotype single cells with cellsnp-lite with a candidate vcf file of 7.4 million common variants from the 1000 Genomes Project, demultiplexed with vireo specifying two donors and assign donors to individuals based on the intersection with variants from bulk RNA-seq data of the same individuals. To distinguish different cynomolgus individuals, we use a reference vcf with informative variants obtained from bulk RNA-seq data to genotype single cells and demultiplex the individuals.

Processing of scRNA-seq data

We remove background RNA with CellBender version 0.2.0 [50] at a false positive rate (FPR) of 0.01. After quality control we retain cells with more than 1000 detected genes and a mitochondrial fraction below 8%. We remove cross-species doublets based on the vireo assignments and intra-species doublets using scDblFinder version 1.6.0 [51], specifying the expected doublet rate based on the cross-species doublet fraction. For each species, we normalize the counts with scran version 1.28.2 [52] and integrated data from different experiments with scanorama [27]. UMAP dimensionality reductions are created with Seurat version 4.3.0 on the first 30 components of the scanorama corrected embedding per species.

For a preliminary cross-species integration, we merged all four species in Seurat, log-normalized and scaled the data and applied Harmony version 0.1.1 [28] on the first 50 PCs to regress out species and experiment effects, using default theta values of 2 for both variables. We identify clusters on the first 20 Harmony-integrated PCs with Seurat at a resolution of 0.1, resulting in a number of clusters similar to the broad cell types described in a human

EB dataset [18].

Reference based classification

To get an initial cell type annotation, we download a reference dataset of day 21 human EBs [18]. We normalize the count matrix with scran and intersect the genes between reference and our scRNA-seq dataset. Next, we train a SingleR version 2.0.0 [29] classifier for the broad cell type classes defined in Figure 1G of the original publication [18] using *trainSingleR* with pseudo-bulk aggregation. Cell type labels are transferred to cells of each species with *classifySingleR*.

Orthologous cell type annotation

To annotate orthologous cell types, we first perform high resolution clustering of the scRNA-seq data for each species separately. For this we take the first 20 components of the scanorama corrected embedding as input to perform clustering in Seurat with *FindNeighbors* and *FindClusters* at a resolution of 2 to obtain the initial high resolution clusters (HRCs).

Next, we score the similarity of all HRCs with an approach based on reciprocal classification. For each species, we train a SingleR classifier on all HRCs of a species. We then classify the cells of all other species with *classifySingleR*. In this way, we can calculate the similarity of each HRC in the target species to each HRC in the reference species as the fraction of cells of the target HRC classified as the reference HRC. To also obtain similarity scores between HRCs within a species, we split the data of each species into a reference set with 80% of cells and a test set with 20% of cells. Analogous to the cross-species classification scheme, we transfer HRC labels from the reference set to the test set and score the overlap of target and reference HRCs.

In the next step, we combine HRCs based on pairwise similarity scores. We average the bidirectional similarity scores for each HRC pair and construct a distance matrix with all

HRCs. Subsequently, based on hierarchical clustering (hclust, average method) we define 26 initial orthologous cell type clusters (OCCs) based on the visual inspection of the distance matrix. In this way, we merge similar HRCs within species and match HRCs across species to obtain a set of OCCs.

OCCs with very similar expression and marker profiles can be further merged. Therefore, we create pseudobulk profiles for each OCC and calculate Spearman's ρ for all pair-wise comparisons within a species (s) based on the 2,000 most variable genes. We perform hierarchical clustering on $1 - \bar{\rho}_s$ and merge orthogolous clusters at a cut height of 0.1, that was interactively determined by also inspecting the similarity of the top marker genes as found by Seurat's *FindMarkers*. In the shiny app, we provide a list of OCC markers for each species separately, but also the intersection of conserved markers. Based on those marker combinations the user can then assign the cell types. If the marker gene distribution as visualized in UMAPs reveals overmerged OCCs, the user can split them interactively. Specifically, we separate merged OCC 4 into iPSCs, cardiac progenitor cells and early epithelial cells for the final assignment. We assign merged OCC 5 as neural crest I, but re-annotate a subcluster present only in cynomolgus and rhesus macaques as fibroblasts. Similarly, we re-annotate a subcluster of merged OCC 12 (granule precursor cells) as astrocyte progenitors in cynomolgus and rhesus macaque. Finally, we exclude OCCs with less than 800 cells that are only present in 1 or 2 species.

We assess the correspondence of the final cell type assignments across species with two approaches. For the scores shown in Supplementary Figure S4 we apply the same reciprocal classification approach as described above providing cell type labels instead of HRCs as initial clusters. For the scores shown in Supplementary Figure S5 we use the function *MetaNeighborUS* of MetaNeighbor version 1.18.0 to compare cell type labels across species.

Pseudotime analysis

Pseudotime trajectories were inferred separately for ectodermal, mesodermal and endodermal lineages in each species using slingshot (version 2.12.0) [31]. For each germ layer, cells were filtered to include iPSCs and cell types belonging to the respective germ layer. The analysis was based on Scanorama-integrated PCA embeddings [27], with iPSCs defined as the starting cluster and germ layer-specific differentiated cell types as endpoints (ectoderm: astrocyte progenitors, granule precursor cells, neurons, and neural crest cells; mesoderm: fibroblasts, smooth muscle cells, cardiac endothelial cells, and cardiac fibroblasts; endoderm: epithelial cells and hepatocytes). If neural_crest_II was absent, neural_crest_I was used as an alternative endpoint. PHATE embeddings (phateR, version 1.0.7)[53] were computed from the Scanorama PCA space to visualize the inferred lineages in two dimensions.

Presence-absence scoring of expression

To determine when to define a gene as expressed in a certain cell type, we derive a lower limit of gene detection per cell type and species while accounting for noise and differences in power to detect expression. We first filter the count matrices for each clone, keeping only genes with at least 1% nonzero counts and cells within 3 median absolute deviations for number of UMIs and the number of genes with counts > 0 per cell type and species. These filtered matrices are then downsampled so that we keep the same number of cells in each species (n=18,800), while keeping the original cell type proportion. Next, per species, we estimate the following distributional characteristics per gene (i) across cell types (j): 1) the fraction of nonzero counts (f_{ij}), 2) the mean ($\mu_ij \pm s.e.(\mu_{ij})$) and dispersion ($\theta_i$) of the negative binomial distribution using glmgampoi v1.10.2 [54]. In the next step, we define a putative expression status per gene per cell type. 1) genes are detectable if their log mean expression $log(\mu_{ij})$ is above the fifth quantile of the $log(\mu)$ value distribution across all genes per cell type. 2) genes are reliably estimable if the ratio $log(\frac{s.e.(\mu_{ij})}{\mu_{ij}})$ is below the

90th quantile of $log(\frac{s.e.(\mu)}{\mu})$ value distribution. Only when both conditions are met is the expression status set to 1, otherwise 0. A binomial logistic regression model using Firth's bias reduction method as implemented in R package logistf (version 1.26.0) is then applied to derive the minimal gene detection needed to call a gene expressed, i.e. when P(Y=1) solve $log(\frac{p}{1-p}) = a + b * f_{ij}$ towards f_{ij}. To ensure consistency between species, we set the detection threshold for each cell type to the maximum threshold among all species.

Cell type specificity and expression conservation scores

To assess cell type specificity and expression conservation of genes across species, we first determine in which cell types a gene is expressed in a species, using the thresholds defined in the previous section. Thus we determine cell type specificity as the number of cell types in which a gene was found to be expressed. Here this score can be maximally 7, i.e. the gene is detected in all cell types that were found in all four species.

To evaluate expression conservation, we develop a phylogenetically weighted conservation score for each gene, reflecting the number of species in which the gene is expressed, weighted by the scaled phylogenetic distance as estimated in Bininda-Edmonds et al. [55]. For each gene, we calculate the expression conservation score as follows:

$$Expression\ conservation = \frac{1}{N_{ct}} \sum_{ct} \sum_{b \in detected} bl \tag{1}$$

where N_{ct} is the number of cell types in which the gene is detected. We then simply sum the scaled branch lengths bl across all cell types (ct) and branches (b) on which we infer the gene to be expressed. Because we only have 4 species, we only have one internal branch, for which we infer expression if at least one great ape and one macaque species show expression in that cell type. The score ranges from 0.075 (detected only in cynomolgus or rhesus macaque) to 1 (detected in the same cell types in all 4 species).

Furthermore, we extract measures of sequence conservation for protein-coding genes from

Supplementary Data S14 in the study by Sullivan et al. [35]. Here, we use the fraction of CDS bases with primate phastCons ¿= 0.96 as a gene-based measure of constraint.

Marker gene detection

We filter the count matrices for each clone to retain only genes with nonzero counts in one of the 7 cell types that were detected in all species. We then downsample these filtered matrices to equalize the number of cells across species, leaving us with ∼11,600 cells per species. Furthermore, to mitigate differences in statistical power due to varying numbers of cells per cell type, we perform testing on cell types with a minimum of 10 and a maximum of 250 cells for each pairwise comparison of 'self' versus 'other'. The maximum of 250 cells ensures that the cell type composition of the 'other' is comparable across species. We identify marker genes using the p-values ($p_{adj} < 0.1$) determined by ZIQ-Rank [36] and use Seurat *FindMarkers* with logistic regression to identify the cell types for which the gene is a marker. Furthermore, the marker gene needs to be above the cell type's detection threshold (see above) and needs to be up-regulated in the cell type for which it is a marker (log fold change > 0.01). Finally, a marker gene must be detected in a larger proportion of cells for which it is a marker than in other cell types ($p_j - \bar{p}_{other} = \Delta > 0.01$). The detection proportion Δ is also used as to sort the lists of marker genes, deeming the genes with the largest Δ as the best marker genes. In order to also gauge within species variation in marker gene detection, we conducted the same analysis across clones instead of species. In order to compare cross-species reproducibility of different types of marker genes, i.e protein-coding, lncRNAs and transcriptional regulators, we wanted to compare the ranked lists of marker genes across species. To this end, we perform a concordance analysis using rank biased overlap (RBO) [39] on the top 100 marker genes (rbo R package version 0.0.1). For this part, a list of transcription factors were created by selecting genes with at least one annotated motif in the motif databases JASPAR 2022 vertebrate core [56], JASPAR 2022 vertebrate

unvalidated [56] and IMAGE [57]. Annotations for protein-coding and lncRNA genes were extracted from the Ensembl GTF file provided with the human Cell Ranger reference dataset (GRCh38-2020-A). To assess the predictive performance of marker genes, we conduct a kNN classification (FNN R package version 1.1.4.1). We train a kNN classifier (k=3) on the log-normalized counts of the top 1-30 human markers per cell type in the human clone 29B5. We then predict the cell type identity in all clones and summarize classification performance per cell type with F1-scores, as well as the average F1-score across all seven cell types.

Abbreviations

CDS	coding sequence
DEA	Differential Expression Analysis
EBs	embryoid bodies
ESCs	embryonic stem cells
HRC	high resolution cluster
iPSCs	induced pluripotent stem cells
kNN	k-nearest neighbors
NHPs	non-human primates
OCC	orthologous cell type cluster
RBO	rank biased overlap
scRNA-seq	single cell RNA-sequencing
SeV	Sendai virus
SNP	single nucleotide polymorphism
TF	transcription factor
UMI	unique molecular identifier
UMAP	Uniform Manifold Approximation and Projection
VCF	Variant Call Format

Declarations

Ethics approval and consent to participate

All procedures performed are approved by the responsible ethic committee on human experimentation (20-122, Ethikkommission LMU München). All experiments were performed in accordance with relevant guidelines and regulations.

Consent for publication

Not applicable.

Availability of data and materials

Code for analysis and figures is available on GitHub (`https://github.com/Hellmann-Lab/EB-analyses`), and accompanying files are deposited in Zenodo (`https://doi.org/10.5281/zenodo.14198850`). All sequencing files were deposited in GEO (GSE280441).

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG): PJ and JJ as well as the majority of the projects cost were funded by a grant to IH and WE (458247426). BV was funded by the grant to IH (407541155) and FE by a grant to WE (458888224).

Author's contributions

WE and IH conceived the study. JJ optimized and conducted EB differentiation experiments and performed 10x scRNA-seq data generation with support of FCE. JG generated and provided human and orangutan iPSCs and supported optimization of EB differentiation

protocols. PS established FACS analyses of EBs. PJ and JJ did primary data analysis. PJ did the pre-processing of the data, developed the pipeline for orthologous cell type assignment, and created the Shiny app. PJ and BV performed the cell type specificity and marker gene conservation analysis. AT prepared reference genomes for non-human primates. TD supported cell type annotation. PJ, JJ and IH wrote the manuscript. All authors reviewed and edited the manuscript.

Acknowledgements

We thank all members of the Enard/Hellmann group for valuable input and discussions. We are grateful to Stefanie Färberböck for her expert technical assistance and help in cell culture. We acknowledge the Core Facility Flow Cytometry at the Biomedical Center, Ludwig-Maximilians-Universität München for providing equipment and services. We thank Dr. Stefan Krebs and the staff of LAFUGA and the NGS Competence Center Tübingen (NCCT) for sequencing services. Schemes were created with BioRender.com.

References

[1] Bakken, T., Cowell, L., Aevermann, B.D., Novotny, M., Hodge, R., Miller, J.A., Lee, A., Chang, I., McCorrison, J., Pulendran, B., Qian, Y., Schork, N.J., Lasken, R.S., Lein, E.S., Scheuermann, R.H.: Cell type discovery and representation in the era of high-content single cell phenotyping. BMC Bioinformatics **18**(Suppl 17), 559 (2017). doi:10.1186/s12859-017-1977-1

[2] Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, Principal investigators: Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature **562**(7727), 367–372 (2018). doi:10.1038/s41586-018-0590-4

[3] Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J.C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C.P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T.N., Shalek, A., Shapiro, E., Sharma, P., Shin, J.W., Stegle, O., Stratton, M., Stubbington, M.J.T., Theis, F.J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., Human Cell Atlas Meeting Participants: The Human Cell Atlas. Elife **6** (2017). doi:10.7554/eLife.27041

[4] Song, Y., Miao, Z., Brazma, A., Papatheodorou, I.: Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. Nat. Commun. **14**(1), 6495 (2023). doi:10.1038/s41467-023-41855-w

[5] Liu, X., Shen, Q., Zhang, S.: Cross-species cell-type assignment from single-cell RNA-seq data by a heterogeneous graph neural network. Genome Res. **33**(1), 96–111 (2023). doi:10.1101/gr.276868.122

[6] Crow, M., Paul, A., Ballouz, S., Huang, Z.J., Gillis, J.: Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. Nat. Commun. **9**(1), 884 (2018). doi:10.1038/s41467-018-03282-0

[7] Bakken, T.E., Jorstad, N.L., Hu, Q., Lake, B.B., Tian, W., Kalmbach, B.E., Crow, M., Hodge, R.D., Krienen, F.M., Sorensen, S.A., Eggermont, J., Yao, Z., Aevermann, B.D., Aldridge, A.I., Bartlett, A., Bertagnolli, D., Casper, T., Castanon, R.G., Crichton, K., Daigle, T.L., Dalley, R., Dee, N., Dembrow, N., Diep, D., Ding, S.-L., Dong, W., Fang, R., Fischer, S., Goldman, M., Goldy, J., Graybuck, L.T., Herb, B.R., Hou, X.,

Kancherla, J., Kroll, M., Lathia, K., van Lew, B., Li, Y.E., Liu, C.S., Liu, H., Lucero, J.D., Mahurkar, A., McMillen, D., Miller, J.A., Moussa, M., Nery, J.R., Nicovich, P.R., Niu, S.-Y., Orvis, J., Osteen, J.K., Owen, S., Palmer, C.R., Pham, T., Plongthongkum, N., Poirion, O., Reed, N.M., Rimorin, C., Rivkin, A., Romanow, W.J., Sedeño-Cortés, A.E., Siletti, K., Somasundaram, S., Sulc, J., Tieu, M., Torkelson, A., Tung, H., Wang, X., Xie, F., Yanny, A.M., Zhang, R., Ament, S.A., Behrens, M.M., Bravo, H.C., Chun, J., Dobin, A., Gillis, J., Hertzano, R., Hof, P.R., Höllt, T., Horwitz, G.D., Keene, C.D., Kharchenko, P.V., Ko, A.L., Lelieveldt, B.P., Luo, C., Mukamel, E.A., Pinto-Duarte, A., Preissl, S., Regev, A., Ren, B., Scheuermann, R.H., Smith, K., Spain, W.J., White, O.R., Koch, C., Hawrylycz, M., Tasic, B., Macosko, E.Z., McCarroll, S.A., Ting, J.T., Zeng, H., Zhang, K., Feng, G., Ecker, J.R., Linnarsson, S., Lein, E.S.: Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature **598**(7879), 111–119 (2021). doi:10.1038/s41586-021-03465-8

[8] Suresh, H., Crow, M., Jorstad, N., Hodge, R., Lein, E., Dobin, A., Bakken, T., Gillis, J.: Comparative single-cell transcriptomic analysis of primate brains highlights human-specific regulatory evolution. Nat Ecol Evol **7**(11), 1930–1943 (2023). doi:10.1038/s41559-023-02186-7

[9] Zhang, Z., Luo, D., Zhong, X., Choi, J.H., Ma, Y., Wang, S., Mahrt, E., Guo, W., Stawiski, E.W., Modrusan, Z., Seshagiri, S., Kapur, P., Hon, G.C., Brugarolas, J., Wang, T.: SCINA: A semi-supervised subtyping algorithm of single cells and bulk samples. Genes (Basel) **10**(7), 531 (2019). doi:10.3390/genes10070531

[10] Guo, H., Li, J.: scSorter: assigning cells to known cell types according to marker genes. Genome Biol. **22**(1), 69 (2021). doi:10.1186/s13059-021-02281-7

[11] Ianevski, A., Giri, A.K., Aittokallio, T.: Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data.

Nat. Commun. **13**(1), 1246 (2022). doi:10.1038/s41467-022-28803-w

[12] Franzén, O., Gan, L.-M., Björkegren, J.L.M.: PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database **2019** (2019). doi:10.1093/database/baz046

[13] Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., Ping, Y., Li, F., Shi, A., Bai, J., Zhao, T., Li, X., Xiao, Y.: CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res. **47**(D1), 721–728 (2019). doi:10.1093/nar/gky900

[14] Hodge, R.D., Bakken, T.E., Miller, J.A., Smith, K.A., Barkan, E.R., Graybuck, L.T., Close, J.L., Long, B., Johansen, N., Penn, O., Yao, Z., Eggermont, J., Höllt, T., Levi, B.P., Shehata, S.I., Aevermann, B., Beller, A., Bertagnolli, D., Brouner, K., Casper, T., Cobbs, C., Dalley, R., Dee, N., Ding, S.-L., Ellenbogen, R.G., Fong, O., Garren, E., Goldy, J., Gwinn, R.P., Hirschstein, D., Keene, C.D., Keshk, M., Ko, A.L., Lathia, K., Mahfouz, A., Maltzer, Z., McGraw, M., Nguyen, T.N., Nyhus, J., Ojemann, J.G., Oldre, A., Parry, S., Reynolds, S., Rimorin, C., Shapovalova, N.V., Somasundaram, S., Szafer, A., Thomsen, E.R., Tieu, M., Quon, G., Scheuermann, R.H., Yuste, R., Sunkin, S.M., Lelieveldt, B., Feng, D., Ng, L., Bernard, A., Hawrylycz, M., Phillips, J.W., Tasic, B., Zeng, H., Jones, A.R., Koch, C., Lein, E.S.: Conserved cell types with divergent features in human versus mouse cortex. Nature **573**(7772), 61–68 (2019). doi:10.1038/s41586-019-1506-7

[15] Krienen, F.M., Goldman, M., Zhang, Q., C H Del Rosario, R., Florio, M., Machold, R., Saunders, A., Levandowski, K., Zaniewski, H., Schuman, B., Wu, C., Lutservitz, A., Mullally, C.D., Reed, N., Bien, E., Bortolin, L., Fernandez-Otero, M., Lin, J.D., Wysoker, A., Nemesh, J., Kulp, D., Burns, M., Tkachev, V., Smith, R., Walsh, C.A., Dimidschstein, J., Rudy, B., S Kean, L., Berretta, S., Fishell, G., Feng, G., McCarroll,

S.A.: Innovations present in the primate interneuron repertoire. Nature **586**(7828), 262–269 (2020). doi:10.1038/s41586-020-2781-z

[16] Brickman, J.M., Serup, P.: Properties of embryoid bodies. Wiley Interdiscip. Rev. Dev. Biol. **6**(2) (2017). doi:10.1002/wdev.259

[17] Itskovitz-Eldor, J., Schuldiner, M., Karsenti, D., Eden, A., Yanuka, O., Amit, M., Soreq, H., Benvenisty, N.: Differentiation of Human Embryonic Stem Cells into Embryoid Bodies Comprising the Three Embryonic Germ Layers. Molecular Medicine **6**(2), 88–95 (2000). doi:10.1007/BF03401776

[18] Rhodes, K., Barr, K.A., Popp, J.M., Strober, B.J., Battle, A., Gilad, Y.: Human embryoid bodies as a novel system for genomic studies of functionally diverse cell types. Elife **11**, 71361 (2022). doi:10.7554/eLife.71361

[19] Guo, H., Tian, L., Zhang, J.Z., Kitani, T., Paik, D.T., Lee, W.H., Wu, J.C.: Single-Cell RNA Sequencing of Human Embryonic Stem Cell Differentiation Delineates Adverse Effects of Nicotine on Embryonic Development. Stem Cell Reports **12**(4), 772–786 (2019). doi:10.1016/j.stemcr.2019.01.022

[20] Han, X., Chen, H., Huang, D., Chen, H., Fei, L., Cheng, C., Huang, H., Yuan, G.-C., Guo, G.: Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. Genome Biol. **19**(1), 47 (2018). doi:10.1186/s13059-018-1426-0

[21] Kanton, S., Boyle, M.J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow, L., Fleck, J.S., Han, D., Qian, Z., Heide, M., Huttner, W.B., Khaitovich, P., Pääbo, S., Treutlein, B., Camp, J.G.: Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature **574**, 418–422 (2019). doi:10.1038/s41586-019-1654-9

[22] Barr, K.A., Rhodes, K.L., Gilad, Y.: The relationship between regulatory changes in cis and trans and the evolution of gene expression in humans and chimpanzees. Genome Biol. **24**(1), 207 (2023). doi:10.1186/s13059-023-03019-3

[23] Geuder, J., Wange, L.E., Janjic, A., Radmer, J., Janssen, P., Bagnoli, J.W., Müller, S., Kaul, A., Ohnuki, M., Enard, W.: A non-invasive method to generate induced pluripotent stem cells from primate urine. Scientific Reports 2021 11:1 **11**, 1–13 (2021). doi:10.1038/s41598-021-82883-0

[24] Jocher, J., Edenhofer, F.C., Janssen, P., Müller, S., Lopez-Parra, D.C., Geuder, J., Enard, W.: Generation and characterization of three fibroblast-derived Rhesus Macaque induced pluripotent stem cell lines. Stem Cell Res. **74**, 103277 (2023). doi:10.1016/j.scr.2023.103277

[25] Edenhofer, F.C., Térmeg, A., Ohnuki, M., Jocher, J., Kliesmete, Z., Briem, E., Hellmann, I., Enard, W.: Generation and characterization of inducible KRAB-dCas9 iPSCs from primates for cross-species CRISPRi. iScience **27**(6), 110090 (2024). doi:10.1016/j.isci.2024.110090

[26] Ludwig, T.E., Andrews, P.W., Barbaric, I., Benvenisty, N., Bhattacharyya, A., Crook, J.M., Daheron, L.M., Draper, J.S., Healy, L.E., Huch, M., Inamdar, M.S., Jensen, K.B., Kurtz, A., Lancaster, M.A., Liberali, P., Lutolf, M.P., Mummery, C.L., Pera, M.F., Sato, Y., Shimasaki, N., Smith, A.G., Song, J., Spits, C., Stacey, G., Wells, C.A., Zhao, T., Mosher, J.T.: ISSCR standards for the use of human stem cells in basic research. Stem Cell Reports **18**(9), 1744–1752 (2023). doi:10.1016/j.stemcr.2023.08.003

[27] Hie, B., Bryson, B.D., Berger, B.: Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. **37**, 685–691 (2019). doi:10.1038/s41587-019-0113-3

[28] Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., Raychaudhuri, S.: Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods **16**(12), 1289–1296 (2019). doi:10.1038/s41592-019-0619-0

[29] Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., Butte, A.J., Bhattacharya, M.: Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. **20**(2), 163–172 (2019). doi:10.1038/s41590-018-0276-y

[30] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M. 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E.P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P., Satija, R.: Integrated analysis of multimodal single-cell data. Cell **184**(13), 3573–358729 (2021). doi:10.1016/j.cell.2021.04.048

[31] Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., Dudoit, S.: Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics **19**(1) (2018). doi:10.1186/s12864-018-4772-0

[32] Duret, L., Mouchiroud, D.: Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**(1), 68–74 (2000). doi:10.1093/oxfordjournals.molbev.a026239

[33] Mollaaghababa, R., Pavan, W.J.: The importance of having your SOX on: role of SOX10 in the development of neural crest-derived melanocytes and glia. Oncogene **22**(20), 3024–3034 (2003). doi:10.1038/sj.onc.1206442

[34] Kliesmete, Z., Orchard, P., Lee, V.Y.K., Geuder, J., Krauß, S.M., Ohnuki, M., Jocher, J., Vieth, B., Enard, W., Hellmann, I.: Evidence for compensatory evolution within pleiotropic regulatory elements. Genome Res. **34**(10), 1528–1539 (2024). doi:10.1101/gr.279001.124

[35] Sullivan, P.F., Meadows, J.R.S., Gazal, S., Phan, B.N., Li, X., Genereux, D.P., Dong, M.X., Bianchi, M., Andrews, G., Sakthikumar, S., Nordin, J., Roy, A., Christmas, M.J., Marinescu, V.D., Wang, C., Wallerman, O., Xue, J., Yao, S., Sun, Q., Szatkiewicz, J., Wen, J., Huckins, L.M., Lawler, A., Keough, K.C., Zheng, Z., Zeng, J., Wray, N.R., Li, Y., Johnson, J., Chen, J., Zoonomia Consortium§, Paten, B., Reilly, S.K., Hughes, G.M., Weng, Z., Pollard, K.S., Pfenning, A.R., Forsberg-Nilsson, K., Karlsson, E.K., Lindblad-Toh, K., Andrews, G., Armstrong, J.C., Bianchi, M., Birren, B.W., Bredemeyer, K.R., Breit, A.M., Christmas, M.J., Clawson, H., Damas, J., Di Palma, F., Diekhans, M., Dong, M.X., Eizirik, E., Fan, K., Fanter, C., Foley, N.M., Forsberg-Nilsson, K., Garcia, C.J., Gatesy, J., Gazal, S., Genereux, D.P., Goodman, L., Grimshaw, J., Halsey, M.K., Harris, A.J., Hickey, G., Hiller, M., Hindle, A.G., Hubley, R.M., Hughes, G.M., Johnson, J., Juan, D., Kaplow, I.M., Karlsson, E.K., Keough, K.C., Kirilenko, B., Koepfli, K.-P., Korstian, J.M., Kowalczyk, A., Kozyrev, S.V., Lawler, A.J., Lawless, C., Lehmann, T., Levesque, D.L., Lewin, H.A., Li, X., Lind, A., Lindblad-Toh, K., Mackay-Smith, A., Marinescu, V.D., Marques-Bonet, T., Mason, V.C., Meadows, J.R.S., Meyer, W.K., Moore, J.E., Moreira, L.R., Moreno-Santillan, D.D., Morrill, K.M., Muntané, G., Murphy, W.J., Navarro, A., Nweeia, M., Ortmann, S., Osmanski, A., Paten, B., Paulat, N.S., Pfenning, A.R., Phan, B.N., Pollard, K.S., Pratt, H.E., Ray, D.A., Reilly, S.K., Rosen, J.R., Ruf, I., Ryan, L., Ryder, O.A., Sabeti, P.C., Schäffer, D.E., Serres, A., Shapiro, B., Smit, A.F.A., Springer, M., Srinivasan, C., Steiner, C., Storer, J.M., Sullivan, K.A.M., Sullivan, P.F., Sundström, E., Supple, M.A., Swofford, R., Talbot, J.-E., Teeling, E., Turner-Maier, J., Valenzuela, A., Wagner, F., Wallerman,

O., Wang, C., Wang, J., Weng, Z., Wilder, A.P., Wirthlin, M.E., Xue, J.R., Zhang, X.: Leveraging base-pair mammalian constraint to understand genetic variation and human disease. Science **380**(6643), 2937 (2023). doi:10.1126/science.abn2937

[36] Ling, W., Zhang, W., Cheng, B., Wei, Y.: ZERO-INFLATED QUANTILE RANK-SCORE BASED TEST (ZIQRANK) WITH APPLICATION TO SCRNA-SEQ DIFFERENTIAL GENE EXPRESSION ANALYSIS. Ann. Appl. Stat. **15**(4), 1673–1696 (2021). doi:10.1214/21-aoas1442

[37] Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., Cepko, C., Erwin, D.H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M.D., Wagner, G.P.: The origin and evolution of cell types. Nat. Rev. Genet. **17**(12), 744–757 (2016). doi:10.1038/nrg.2016.127

[38] Johnsson, P., Lipovich, L., Grandér, D., Morris, K.V.: Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim. Biophys. Acta **1840**(3), 1063–1071 (2014). doi:10.1016/j.bbagen.2013.10.035

[39] Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. **28**(4), 1–38 (2010). doi:10.1145/1852102.1852106

[40] Wang, J., Sun, H., Jiang, M., Li, J., Zhang, P., Chen, H., Mei, Y., Fei, L., Lai, S., Han, X., Song, X., Xu, S., Chen, M., Ouyang, H., Zhang, D., Yuan, G.-C., Guo, G.: Tracing cell-type evolution by cross-species comparison of cell atlases. Cell Rep. **34**(9), 108803 (2021). doi:10.1016/j.celrep.2021.108803

[41] He, Z., Dony, L., Fleck, J.S., Szałata, A., Li, K.X., Slišković, I., Lin, H.-C., Santel, M., Atamian, A., Quadrato, G., Sun, J., Paşca, S.P., Camp, J.G., Theis, F., Treutlein, B.: An integrated transcriptomic cell atlas of human neural organoids. bioRxiv (2023). doi:10.1101/2023.10.05.561097

[42] Hastings, K.E.: Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J. Mol. Evol. **42**(6), 631–640 (1996). doi:10.1007/BF02338796

[43] Feng, M., Swevers, L., Sun, J.: Hemocyte clusters defined by scRNA-seq in Bombyx mori: In silico analysis of predicted marker genes and implications for potential functional roles. Front. Immunol. **13**, 852702 (2022). doi:10.3389/fimmu.2022.852702

[44] Benito-Kwiecinski, S., Giandomenico, S.L., Sutcliffe, M., Riis, E.S., Freire-Pritchett, P., Kelava, I., Wunderlich, S., Martin, U., Wray, G.A., McDole, K., Lancaster, M.A.: An early cell shape transition drives evolutionary expansion of the human forebrain. Cell **184**(8), 2084–210219 (2021). doi:10.1016/j.cell.2021.02.050

[45] Gulati, G.S., D'Silva, J.P., Liu, Y., Wang, L., Newman, A.M.: Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. Nat. Rev. Mol. Cell Biol., 1–21 (2024). doi:10.1038/s41580-024-00768-2

[46] Pascal, L.E., True, L.D., Campbell, D.S., Deutsch, E.W., Risk, M., Coleman, I.M., Eichner, L.J., Nelson, P.S., Liu, A.Y.: Correlation of mRNA and protein levels: cell type-specific gene expression of cluster designation antigens in the prostate. BMC Genomics **9**(1), 246 (2008). doi:10.1186/1471-2164-9-246

[47] Shumate, A., Salzberg, S.L.: Liftoff: accurate mapping of gene annotations. Bioinformatics **37**(12), 1639–1643 (2021). doi:10.1093/bioinformatics/btaa1016

[48] Huang, X., Huang, Y.: Cellsnp-lite: an efficient tool for genotyping single cells. Bioinformatics **37**(23), 4569–4571 (2021). doi:10.1093/bioinformatics/btab358

[49] Huang, Y., McCarthy, D.J., Stegle, O.: Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biol. **20**(1), 273 (2019). doi:10.1186/s13059-019-1865-2

[50] Fleming, S.J., Chaffin, M.D., Arduini, A., Akkad, A.-D., Banks, E., Marioni, J.C., Philippakis, A.A., Ellinor, P.T., Babadi, M.: Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. Nat. Methods **20**(9), 1323–1335 (2023). doi:10.1038/s41592-023-01943-7

[51] Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W., Robinson, M.D.: Doublet identification in single-cell sequencing data using scDblFinder. F1000Res. **10**, 979 (2021). doi:10.12688/f1000research.73600.2

[52] 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. **17**, 1–14 (2016). doi:10.1186/S13059-016-0947-7/TABLES/2

[53] Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., van den Elzen, A., Hirn, M.J., Coifman, R.R., Ivanova, N.B., Wolf, G., Krishnaswamy, S.: Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. **37**(12), 1482–1492 (2019). doi:10.1038/s41587-019-0336-3

[54] Ahlmann-Eltze, C., Huber, W.: glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. Bioinformatics **36**(24), 5701–5702 (2021). doi:10.1093/bioinformatics/btaa1009

[55] Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L., Purvis, A.: The delayed rise of present-day mammals. Nature **446**(7135), 507–512 (2007). doi:10.1038/nature05634

[56] Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R.B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Leung, T.Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W.W., Parcy, F., Mathelier,

A.: JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. **50**(D1), 165–173 (2022). doi:10.1093/nar/gkab1113

[57] Madsen, J.G.S., Rauch, A., Van Hauwaert, E.L., Schmidt, S.F., Winnefeld, M., Mandrup, S.: Integrated analysis of motif activity and gene expression changes of transcription factors. Genome Res. **28**(2), 243–255 (2018). doi:10.1101/gr.227231.117

[58] Nguyen, Q.H., Lukowski, S.W., Chiu, H.S., Senabouth, A., Bruxner, T.J.C., Christ, A.N., Palpant, N.J., Powell, J.E.: Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. Genome Res. **28**(7), 1053–1066 (2018). doi:10.1101/gr.223925.117

[59] Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., Wong, K.-Y., Sung, K.W., Lee, C.W.H., Zhao, X.-D., Chiu, K.-P., Lipovich, L., Kuznetsov, V.A., Robson, P., Stanton, L.W., Wei, C.-L., Ruan, Y., Lim, B., Ng, H.-H.: The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat. Genet. **38**(4), 431–440 (2006). doi:10.1038/ng1760

[60] Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., Borowsky, M.L., Kharchenko, P.V., Park, P.J., Hochedlinger, K.: Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. Cell Stem Cell **12**(6), 699–712 (2013). doi:10.1016/j.stem.2013.04.013

[61] Närvä, E., Rahkonen, N., Emani, M.R., Lund, R., Pursiheimo, J.-P., Nästi, J., Autio, R., Rasool, O., Denessiouk, K., Lähdesmäki, H., Rao, A., Lahesmaa, R.: RNA-binding protein L1TD1 interacts with LIN28 via RNA and is required for human embryonic stem cell self-renewal and cancer cell proliferation. Stem Cells **30**(3), 452–460 (2012). doi:10.1002/stem.1013

[62] Graham, V., Khudyakov, J., Ellis, P., Pevny, L.: SOX2 functions to maintain neural progenitor identity. Neuron **39**(5), 749–765 (2003). doi:10.1016/s0896-6273(03)00497-5

[63] Lodato, M.A., Ng, C.W., Wamstad, J.A., Cheng, A.W., Thai, K.K., Fraenkel, E., Jaenisch, R., Boyer, L.A.: SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state. PLoS Genet. **9**(2), 1003288 (2013). doi:10.1371/journal.pgen.1003288

[64] Ziller, M.J., Edri, R., Yaffe, Y., Donaghey, J., Pop, R., Mallard, W., Issner, R., Gifford, C.A., Goren, A., Xing, J., Gu, H., Cachiarelli, D., Tsankov, A., Epstein, C., Rinn, J.R., Mikkelsen, T.S., Kohlbacher, O., Gnirke, A., Bernstein, B.E., Elkabetz, Y., Meissner, A.: Dissecting neural differentiation regulatory networks through epigenetic footprinting. Nature **518**(7539), 355–359 (2015). doi:10.1038/nature13990

[65] Harada, Y., Yamada, M., Imayoshi, I., Kageyama, R., Suzuki, Y., Kuniya, T., Furutachi, S., Kawaguchi, D., Gotoh, Y.: Cell cycle arrest determines adult neural stem cell ontogeny by an embryonic Notch-nonoscillatory Hey1 module. Nat. Commun. **12**(1), 6562 (2021). doi:10.1038/s41467-021-26605-0

[66] Kawase, S., Kuwako, K., Imai, T., Renault-Mihara, F., Yaguchi, K., Itohara, S., Okano, H.: Regulatory factor X transcription factors control Musashi1 transcription in mouse neural stem/progenitor cells. Stem Cells Dev. **23**(18), 2250–2261 (2014). doi:10.1089/scd.2014.0219

[67] Tan, L., Shi, J., Moghadami, S., Wright, C.P., Parasar, B., Seo, Y., Vallejo, K., Cobos, I., Duncan, L., Chen, R., Deisseroth, K.: Cerebellar granule cells develop non-neuronal 3D genome architecture over the lifespan. bioRxivorg (2023). doi:10.1101/2023.02.25.530020

[68] Fraser, J., Essebier, A., Brown, A.S., Davila, R.A., Harkins, D., Zalucki, O., Shapiro, L.P., Penzes, P., Wainwright, B.J., Scott, M.P., Gronostajski, R.M., Bodén, M., Piper, M., Harvey, T.J.: Common regulatory targets of NFIA, NFIX and NFIB during postnatal cerebellar development. Cerebellum **19**(1), 89–101 (2020). doi:10.1007/s12311-019-01089-3

[69] Aruga, J., Minowa, O., Yaginuma, H., Kuno, J., Nagai, T., Noda, T., Mikoshiba, K.: Mouse Zic1 is involved in cerebellar development. J. Neurosci. **18**(1), 284–293 (1998). doi:10.1523/jneurosci.18-01-00284.1998

[70] Schüller, U., Kho, A.T., Zhao, Q., Ma, Q., Rowitch, D.H.: Cerebellar 'transcriptome' reveals cell-type and stage-specific expression during postnatal development and tumorigenesis. Mol. Cell. Neurosci. **33**(3), 247–259 (2006). doi:10.1016/j.mcn.2006.07.010

[71] Blank, M.C., Grinberg, I., Aryee, E., Laliberte, C., Chizhikov, V.V., Henkelman, R.M., Millen, K.J.: Multiple developmental programs are altered by loss of Zic1 and Zic4 to cause Dandy-Walker malformation cerebellar pathogenesis. Development **138**(6), 1207–1216 (2011). doi:10.1242/dev.054114

[72] Kim, J., Lo, L., Dormand, E., Anderson, D.J.: SOX10 maintains multipotency and inhibits neuronal differentiation of neural crest stem cells. Neuron **38**(1), 17–31 (2003). doi:10.1016/s0896-6273(03)00163-6

[73] Tseng, T.-C., Hsieh, F.-Y., Dai, N.-T., Hsu, S.-H.: Substrate-mediated reprogramming of human fibroblasts into neural crest stem-like cells and their applications in neural repair. Biomaterials **102**, 148–161 (2016). doi:10.1016/j.biomaterials.2016.06.020

[74] Dottori, M., Gross, M.K., Labosky, P., Goulding, M.: The winged-helix transcription factor Foxd3 suppresses interneuron differentiation and promotes neural crest cell fate. Development **128**(21), 4127–4138 (2001). doi:10.1242/dev.128.21.4127

[75] Hackland, J.O.S., Frith, T.J.R., Thompson, O., Marin Navarro, A., Garcia-Castro, M.I., Unger, C., Andrews, P.W.: Top-Down Inhibition of BMP Signaling Enables Robust Induction of hPSCs Into Neural Crest in Fully Defined, Xeno-free Conditions. Stem Cell Reports **9**(4), 1043–1052 (2017). doi:10.1016/j.stemcr.2017.08.008

[76] Murphy, M., Bernard, O., Reid, K., Bartlett, P.F.: Cell lines derived from mouse neural crest are representative of cells at various stages of differentiation. J. Neurobiol. **22**(5), 522–535 (1991). doi:10.1002/neu.480220508

[77] Klim, J.R., Williams, L.A., Limone, F., Guerra San Juan, I., Davis-Dusenbery, B.N., Mordes, D.A., Burberry, A., Steinbaugh, M.J., Gamage, K.K., Kirchner, R., Moccia, R., Cassel, S.H., Chen, K., Wainger, B.J., Woolf, C.J., Eggan, K.: ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. Nat. Neurosci. **22**(2), 167–179 (2019). doi:10.1038/s41593-018-0300-4

[78] Guerra San Juan, I., Nash, L.A., Smith, K.S., Leyton-Jaimes, M.F., Qian, M., Klim, J.R., Limone, F., Dorr, A.B., Couto, A., Pintacuda, G., Joseph, B.J., Whisenant, D.E., Noble, C., Melnik, V., Potter, D., Holmes, A., Burberry, A., Verhage, M., Eggan, K.: Loss of mouse Stmn2 function causes motor neuropathy. Neuron **110**(10), 1671–16886 (2022). doi:10.1016/j.neuron.2022.02.011

[79] Ware, M., Hamdi-Rozé, H., Le Friec, J., David, V., Dupé, V.: Regulation of downstream neuronal genes by proneural transcription factors during initial neurogenesis in the vertebrate brain. Neural Dev. **11**(1), 22 (2016). doi:10.1186/s13064-016-0077-7

[80] Mori, K., Muto, Y., Kokuzawa, J., Yoshioka, T., Yoshimura, S., Iwama, T., Okano, Y., Sakai, N.: Neuronal protein NP25 interacts with F-actin. Neurosci. Res. **48**(4), 439–446 (2004). doi:10.1016/j.neures.2003.12.012

[81] Gleeson, J.G., Lin, P.T., Flanagan, L.A., Walsh, C.A.: Doublecortin is a microtubule-associated protein and is expressed widely by migrating neurons. Neuron **23**(2), 257–271 (1999). doi:10.1016/s0896-6273(00)80778-3

[82] Rojas, M.G., Pereira-Simon, S., Zigmond, Z.M., Varona Santos, J., Perla, M., Santos Falcon, N., Stoyell-Conti, F.F., Salama, A., Yang, X., Long, X., Duque, J.C., Salman, L.H., Tabbara, M., Martinez, L., Vazquez-Padron, R.I.: Single-cell analyses offer insights into the different remodeling programs of arteries and veins. Cells **13**(10), 793 (2024). doi:10.3390/cells13100793

[83] Muhl, L., Mocci, G., Pietilä, R., Liu, J., He, L., Genové, G., Leptidis, S., Gustafsson, S., Buyandelger, B., Raschperger, E., Hansson, E.M., Björkegren, J.L.M., Vanlandewijck, M., Lendahl, U., Betsholtz, C.: A single-cell transcriptomic inventory of murine smooth muscle cells. Dev. Cell **57**(20), 2426–24436 (2022). doi:10.1016/j.devcel.2022.09.015

[84] Hashmi, S.K., Barka, V., Yang, C., Schneider, S., Svitkina, T.M., Heuckeroth, R.O.: Pseudo-obstruction-inducing ACTG2R257C alters actin organization and function. JCI Insight **5**(16) (2020). doi:10.1172/jci.insight.140604

[85] Mononen, M.M., Leung, C.Y., Xu, J., Chien, K.R.: Trajectory mapping of human embryonic stem cell cardiogenesis reveals lineage branch points and an ISL1 progenitor-derived cardiac fibroblast lineage. Stem Cells **38**(10), 1267–1278 (2020). doi:10.1002/stem.3236

[86] Tachampa, K., Wongtawan, T.: Unique patterns of cardiogenic and fibrotic gene expression in rat cardiac fibroblasts. Vet. World **13**(8), 1697–1708 (2020). doi:10.14202/vetworld.2020.1697-1708

[87] Floy, M.E., Givens, S.E., Matthys, O.B., Mateyka, T.D., Kerr, C.M., Steinberg, A.B., Silva, A.C., Zhang, J., Mei, Y., Ogle, B.M., McDevitt, T.C., Kamp, T.J., Palecek, S.P.:

Developmental lineage of human pluripotent stem cell-derived cardiac fibroblasts affects their functional phenotype. FASEB J. **35**(9), 21799 (2021). doi:10.1096/fj.202100523R

[88] Ko, T., Nomura, S., Yamada, S., Fujita, K., Fujita, T., Satoh, M., Oka, C., Katoh, M., Ito, M., Katagiri, M., Sassa, T., Zhang, B., Hatsuse, S., Yamada, T., Harada, M., Toko, H., Amiya, E., Hatano, M., Kinoshita, O., Nawata, K., Abe, H., Ushiku, T., Ono, M., Ikeuchi, M., Morita, H., Aburatani, H., Komuro, I.: Cardiac fibroblasts regulate the development of heart failure via Htra3-TGF-β-IGFBP7 axis. Nat. Commun. **13**(1), 3275 (2022). doi:10.1038/s41467-022-30630-y

[89] Furtado, M.B., Costa, M.W., Pranoto, E.A., Salimova, E., Pinto, A.R., Lam, N.T., Park, A., Snider, P., Chandran, A., Harvey, R.P., Boyd, R., Conway, S.J., Pearson, J., Kaye, D.M., Rosenthal, N.A.: Cardiogenic genes expressed in cardiac fibroblasts contribute to heart development and repair. Circ. Res. **114**(9), 1422–1434 (2014). doi:10.1161/CIRCRESAHA.114.302530

[90] Oikawa, T., Otsuka, Y., Onodera, Y., Horikawa, M., Handa, H., Hashimoto, S., Suzuki, Y., Sabe, H.: Necessity of p53-binding to the CDH1 locus for its expression defines two epithelial cell types differing in their integrity. Sci. Rep. **8**(1), 1595 (2018). doi:10.1038/s41598-018-20043-7

[91] Bondow, B.J., Faber, M.L., Wojta, K.J., Walker, E.M., Battle, M.A.: E-cadherin is required for intestinal morphogenesis in the mouse. Dev. Biol. **371**(1), 1–12 (2012). doi:10.1016/j.ydbio.2012.06.005

[92] Martowicz, A., Seeber, A., Untergasser, G.: The role of EpCAM in physiology and pathology of the epithelium. Histol. Histopathol. **31**(4), 349–355 (2016). doi:10.14670/HH-11-678

[93] Huang, L., Yang, Y., Yang, F., Liu, S., Zhu, Z., Lei, Z., Guo, J.: Functions of EpCAM in physiological processes and diseases (Review). Int. J. Mol. Med. **42**(4), 1771–1785 (2018). doi:10.3892/ijmm.2018.3764

[94] Farkas, A.E., Hilgarth, R.S., Capaldo, C.T., Gerner-Smidt, C., Powell, D.R., Vertino, P.M., Koval, M., Parkos, C.A., Nusrat, A.: HNF4α regulates claudin-7 protein expression during intestinal epithelial differentiation. Am. J. Pathol. **185**(8), 2206–2218 (2015). doi:10.1016/j.ajpath.2015.04.023

[95] Xing, T., Benderman, L.J., Sabu, S., Parker, J., Yang, J., Lu, Q., Ding, L., Chen, Y.-H.: Tight junction protein claudin-7 is essential for intestinal epithelial stem cell self-renewal and differentiation. Cell. Mol. Gastroenterol. Hepatol. **9**(4), 641–659 (2020). doi:10.1016/j.jcmgh.2019.12.005

[96] Banas, A., Teratani, T., Yamamoto, Y., Tokuhara, M., Takeshita, F., Quinn, G., Okochi, H., Ochiya, T.: Adipose tissue-derived mesenchymal stem cells as a source of human hepatocytes. Hepatology **46**(1), 219–228 (2007). doi:10.1002/hep.21704

[97] Lavon, N., Benvenisty, N.: Study of hepatocyte differentiation using embryonic stem cells. J. Cell. Biochem. **96**(6), 1193–1202 (2005). doi:10.1002/jcb.20590

[98] Krueger, W.H., Tanasijevic, B., Barber, V., Flamier, A., Gu, X., Manautou, J., Rasmussen, T.P.: Cholesterol-secreting and statin-responsive hepatocytes from human ES and iPS cells to model hepatic involvement in cardiovascular health. PLoS One **8**(7), 67296 (2013). doi:10.1371/journal.pone.0067296

[99] De Giorgi, M., Li, A., Hurley, A., Barzi, M., Doerfler, A.M., Cherayil, N.A., Smith, H.E., Brown, J.D., Lin, C.Y., Bissig, K.-D., Bao, G., Lagor, W.R.: Targeting the Apoa1 locus for liver-directed gene therapy. Mol. Ther. Methods Clin. Dev. **21**, 656–669 (2021). doi:10.1016/j.omtm.2021.04.011

[100] Peng, W.C., Logan, C.Y., Fish, M., Anbarchian, T., Aguisanda, F., Álvarez-Varela, A., Wu, P., Jin, Y., Zhu, J., Li, B., Grompe, M., Wang, B., Nusse, R.: Inflammatory cytokine TNFα promotes the long-term expansion of primary hepatocytes in 3D culture. Cell **175**(6), 1607–161915 (2018). doi:10.1016/j.cell.2018.11.012

Supplementary Information

Identification and comparison of orthologous cell types from primate embryoid bodies shows limits of marker gene transferability

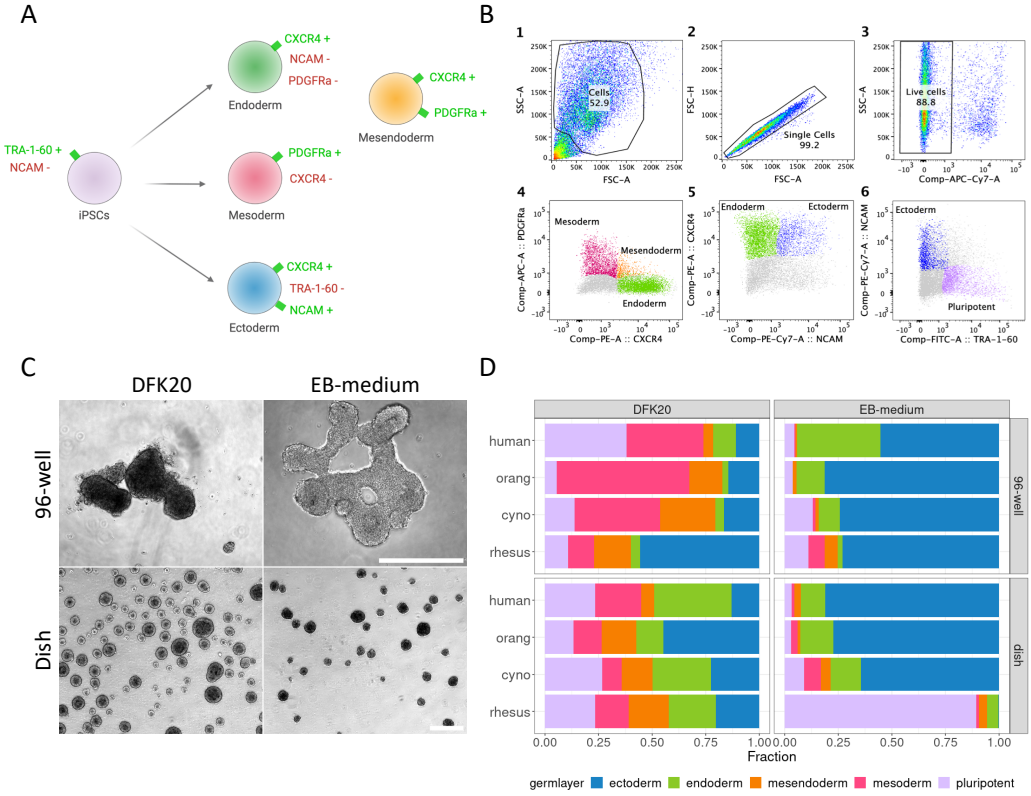
Jessica Jocher[1,*], Philipp Janssen[1,*], Beate Vieth[1], Fiona C. Edenhofer[1], Tamina Dietl[2], Anita Térmeg[1], Johanna Geuder[1], Wolfgang Enard[1,**], Ines Hellmann[1,**]

[1]Anthropology and Human Genomics, Faculty of Biology, Ludwig-Maximilians-Universität München, Germany

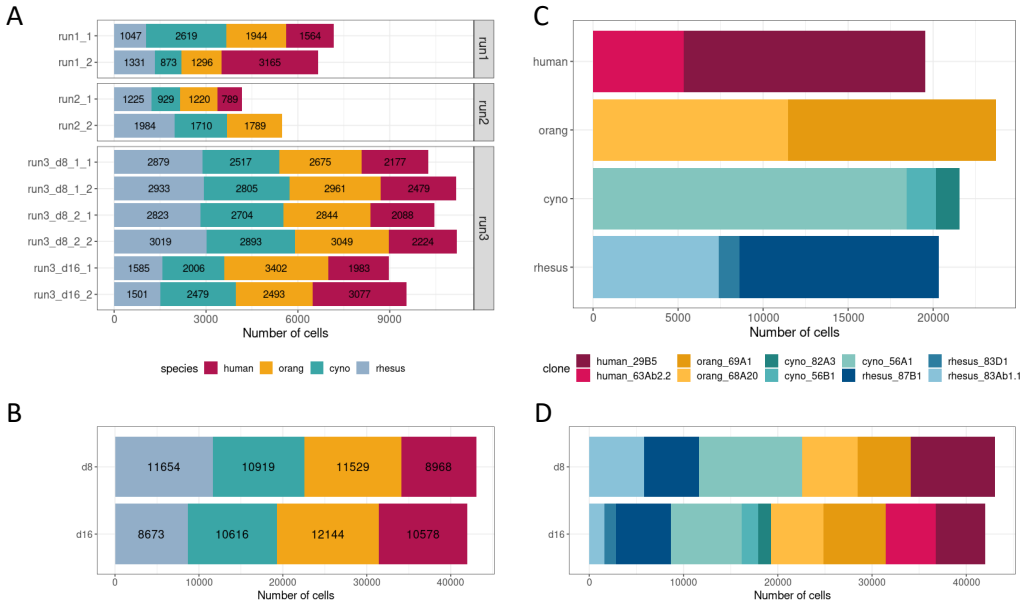
[2]Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt: Munich, Germany

[*] contributed equally

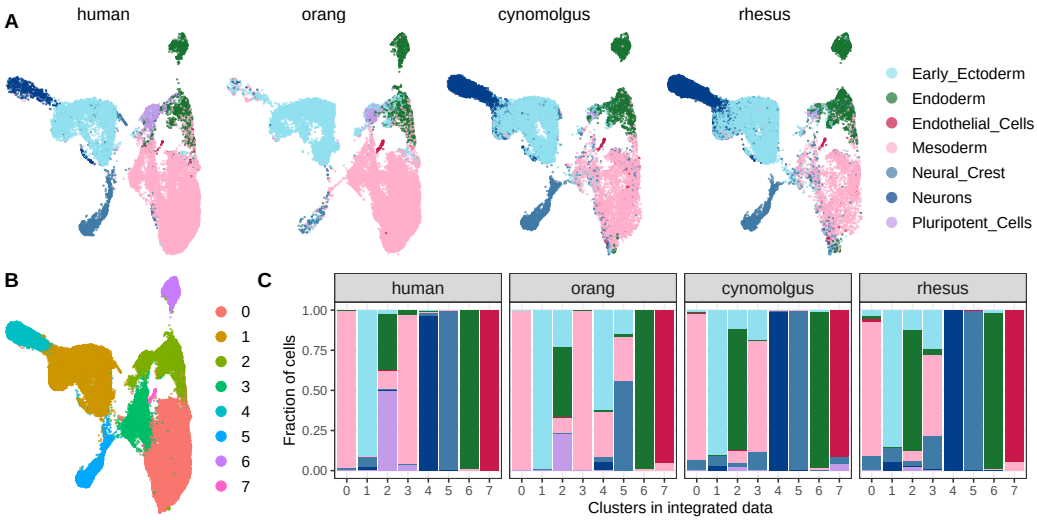
[**] correspondence hellmann@bio.lmu.de



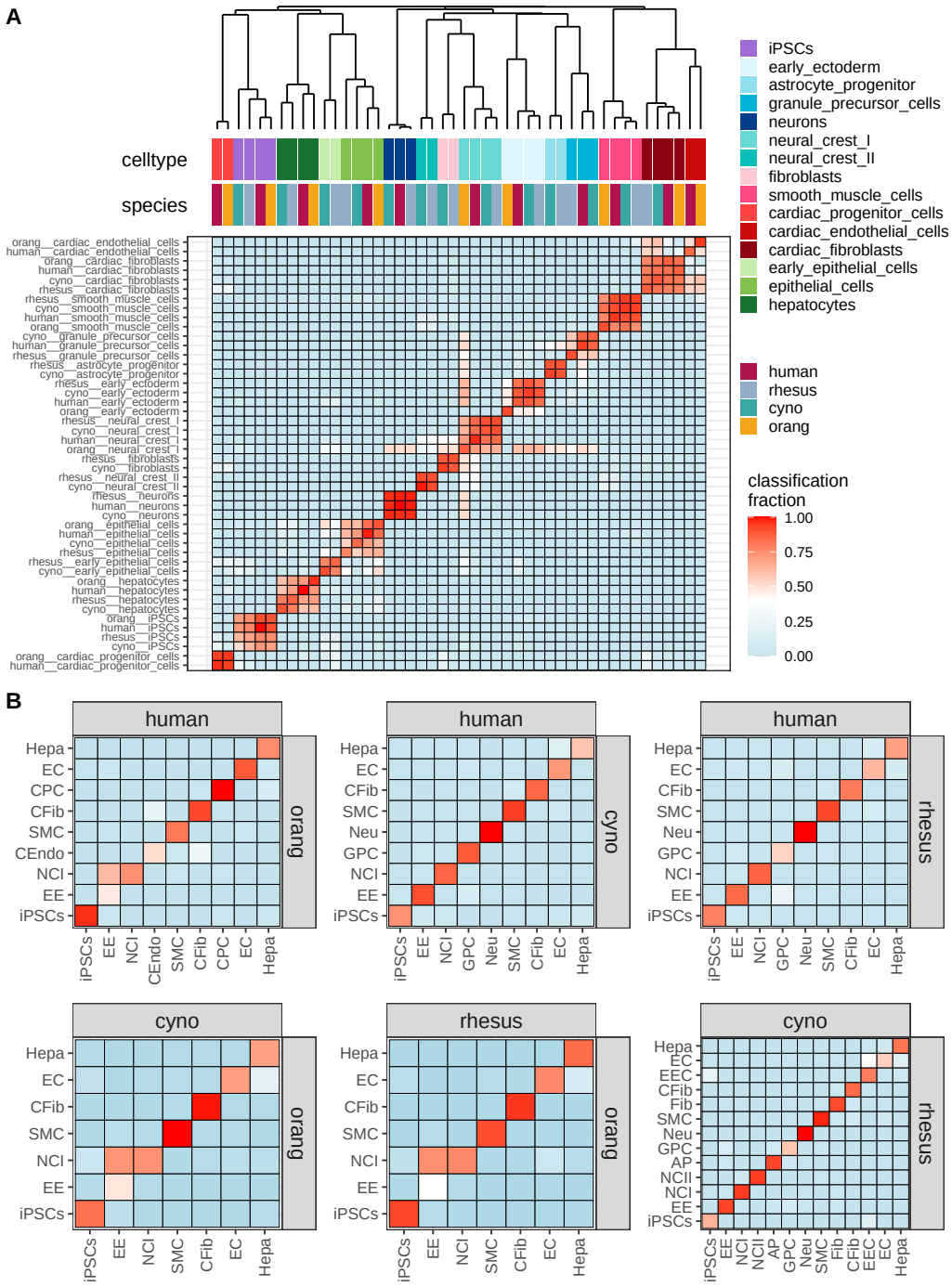
Supplementary Figure S1. Comparison of EB differentiation protocols using flow cytometry. A) Antibody combination to analyze iPSCs and cells of the three primary germ layers in a single sample. B) Flow cytometry gating overview using human EBs at day 7 of differentiation. 1. Gating of cell population. 2. Gating of single cell population. 3. Gating of live cell population. 4.-6. Gating of cells belonging to pluripotent or germ layer populations based on the antibody combination shown in S1A). C) Phase contrast images of orangutan EBs on day 6 of differentiation in 4 different culture conditions. Scale bar represents 250 µm. D) Barplot of pluripotency and germ layer proportions of day 7 EBs from human, orangutan, cynomolgus and rhesus in the 4 different culture conditions.



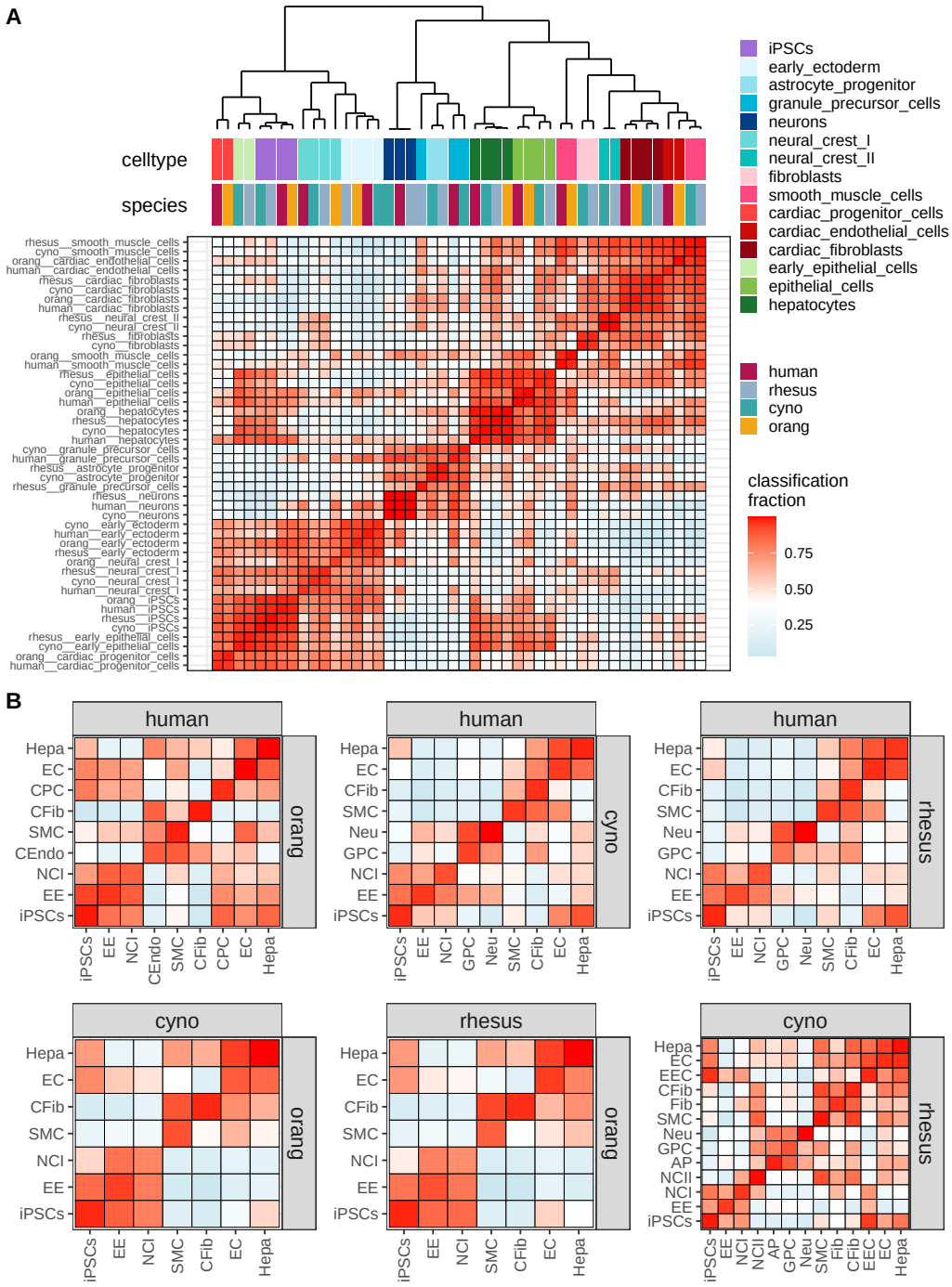
Supplementary Figure S2. Total number of recovered cells. A) Barplot of cell numbers per species and experimental batch and 10x lane. B) Barplot of cell numbers per species and day of differentiation. C) Barplot of cell numbers per clone. D) Barplot of cell numbers per clone and day of differentiation.



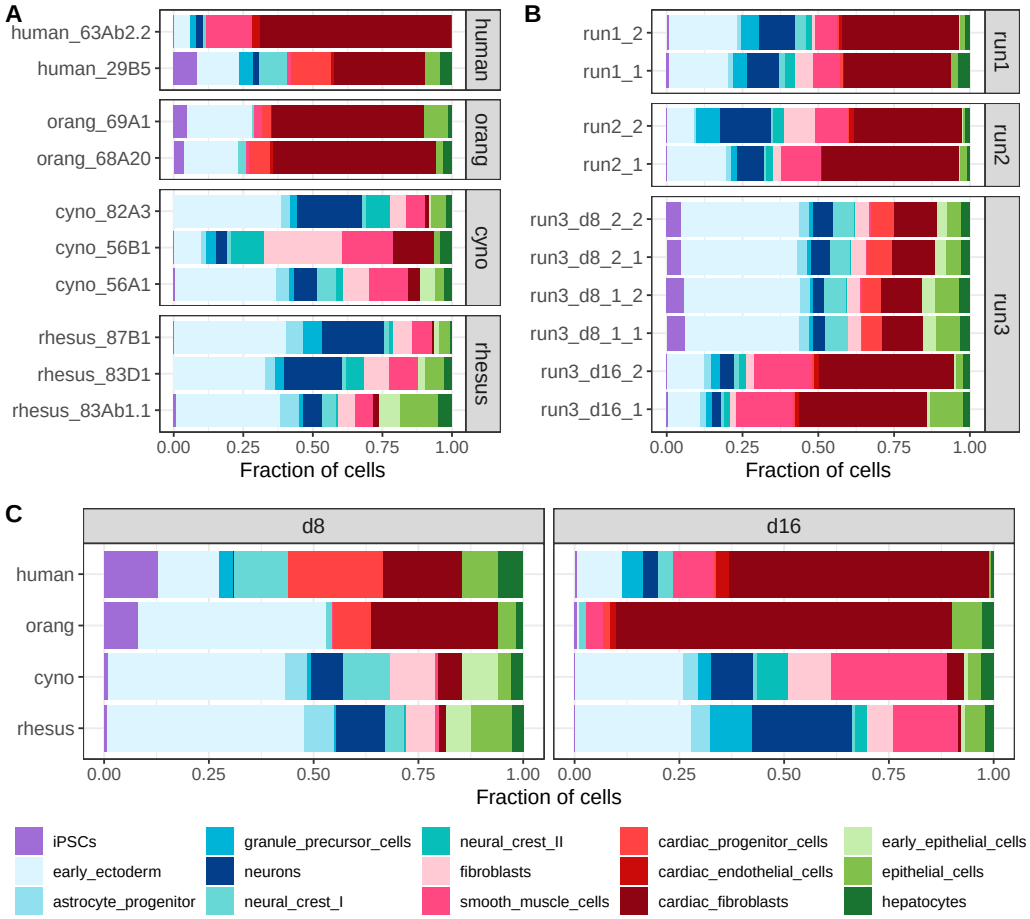
Supplementary Figure S3. Reference based cell type classification. A) UMAP representations colored by labels from a classification with a reference dataset of day 21 human embryoid bodies [18]. B) Single cell clusters in integrated data from all 4 species. C) Stacked bar plot of the proportions of predicted labels across clusters obtained in the integrated dataset.



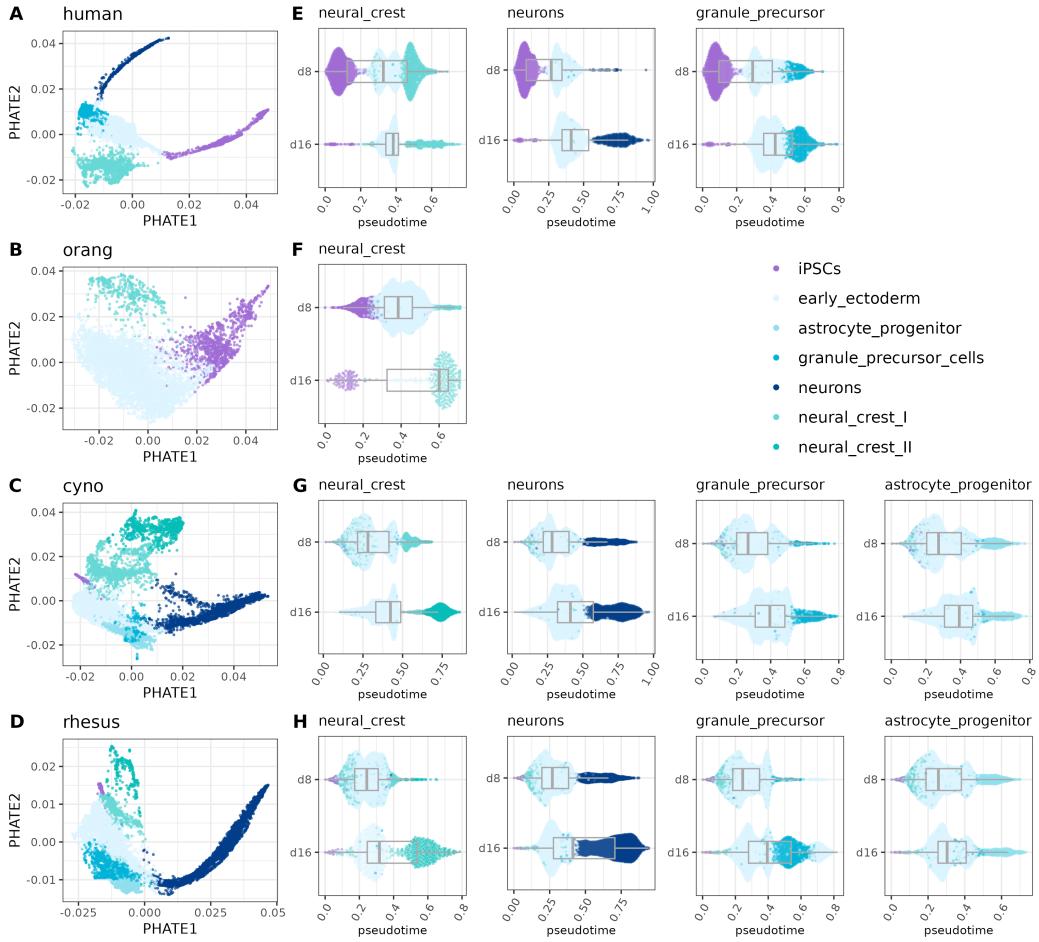
Supplementary Figure S4. Replicability of cell types across species measured by reciprocal classification. A) Heatmap illustrating 'all vs all' similarities of cell types from all four species. For each cell type pair the similarity represents the average classification fraction obtained through reciprocal classification between each species pair. B) Average classification fractions for cell types that are shared among each species pair. AP: astrocyte progenitor, CFib: cardiac fibroblasts, CEndo: cardiac endothelial cells, CPC: cardiac progenitor cells, EEC: early epithelial cells, EE: early ectoderm, EC: epithelial cells, Fib: fibroblasts, GPC: granule precursor cells, Hepa: hepatocytes, NCI: neural crest I, NCII: neural crest II, Neu: neurons, SMC: smooth muscle cells.



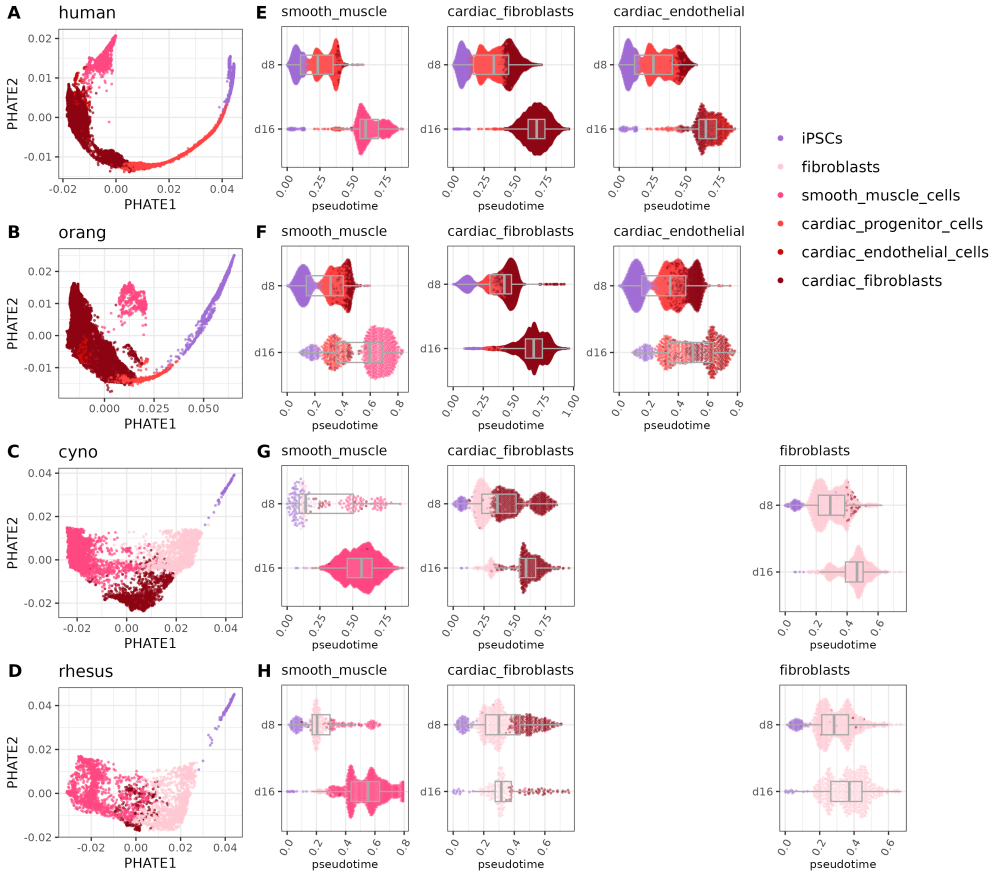
Supplementary Figure S5. Replicability of cell types across species measured with MetaNeighbor. A) Heatmap illustrating 'all vs all' similarities of cell types from all four species. For each cell type pair the similarity represents area under the receiver operator characteristic curve (AUROC) scores obtained with MetaNeighbor [6] in unsupervised mode. B) AUROC scores for cell types that are shared among each species pair. AP: astrocyte progenitor, CFib: cardiac fibroblasts, CEndo: cardiac endothelial cells, CPC: cardiac progenitor cells, EEC: early epithelial cells, EE: early ectoderm, EC: epithelial cells, Fib: fibroblasts, GPC: granule precursor cells, Hepa: hepatocytes, NCI: neural crest I, NCII: neural crest II, Neu: neurons, SMC: smooth muscle cells.



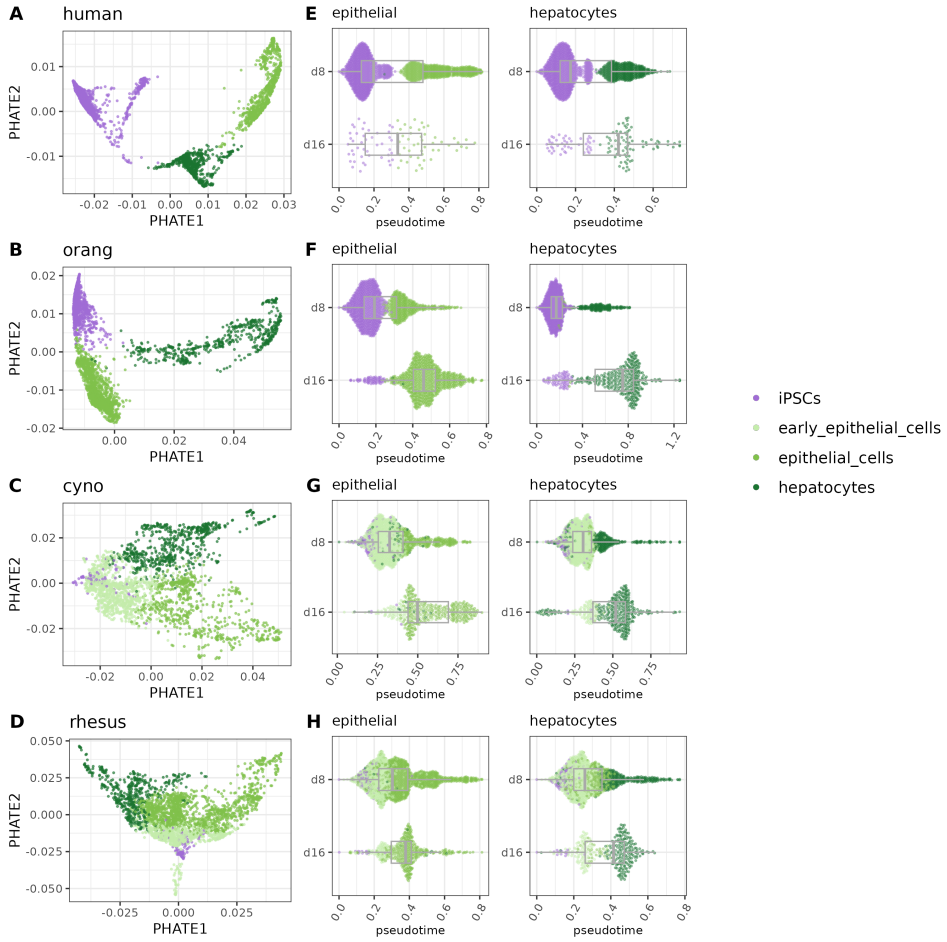
Supplementary Figure S6. Cell type annotation. A) Barplot of cell type fractions per species and clone. B) Barplot of cell type fractions per experimental batch and 10x lane. C) Barplot of cell type fractions per day of differentiation across different species.



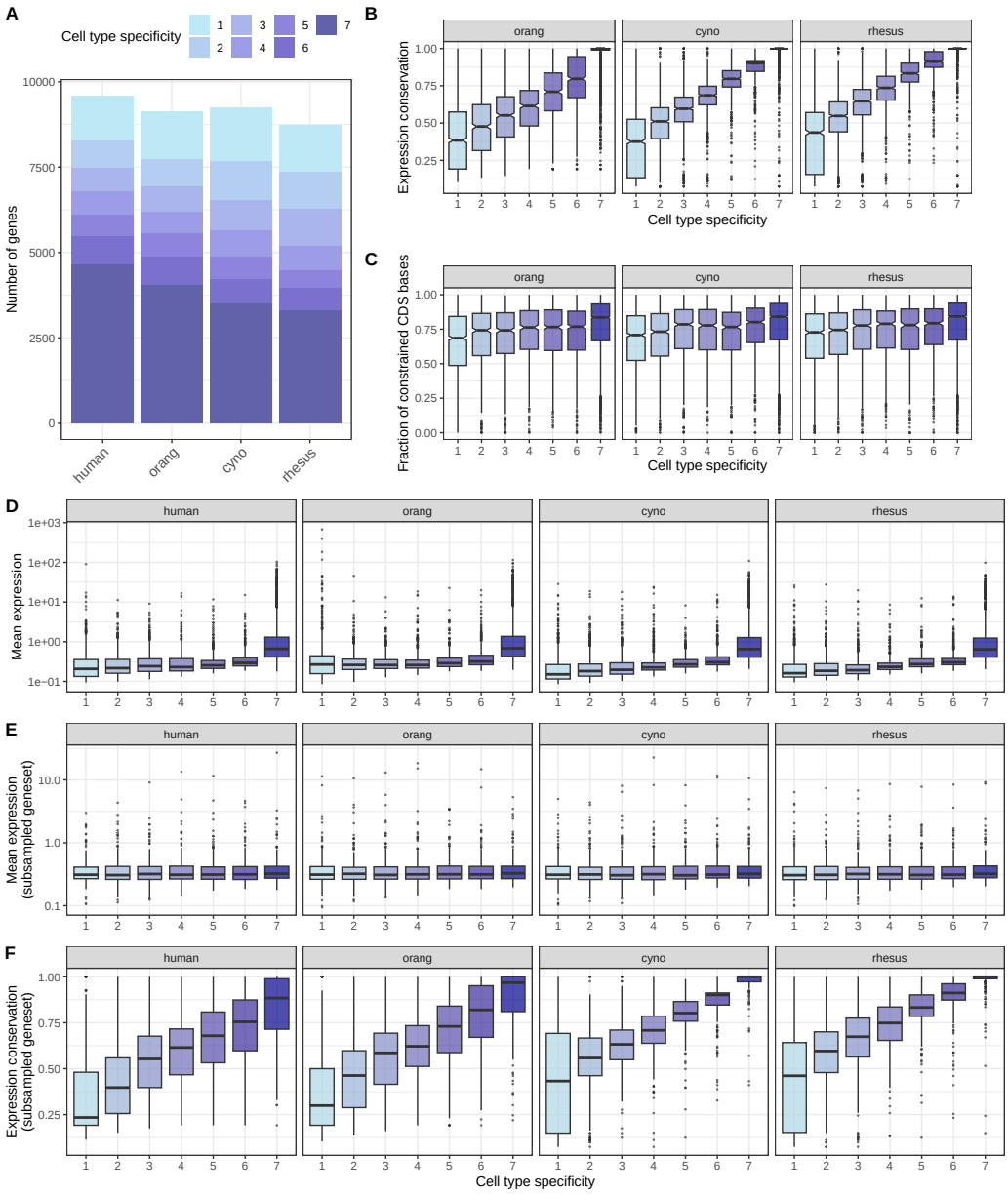
Supplementary Figure S7. Pseudotime analysis of ectoderm differentiation trajectories. A-D) PHATE embeddings of iPSCs and ectodermal cells from human (A), orangutan (B), cynomolgus (C) and rhesus macaque (D). E-F) Distribution of pseudotime values across major ectodermal lineages for each species. Boxplots illustrate shifts in pseudotime distributions between day 8 (d8) and day 16 (d16) of differentiation.



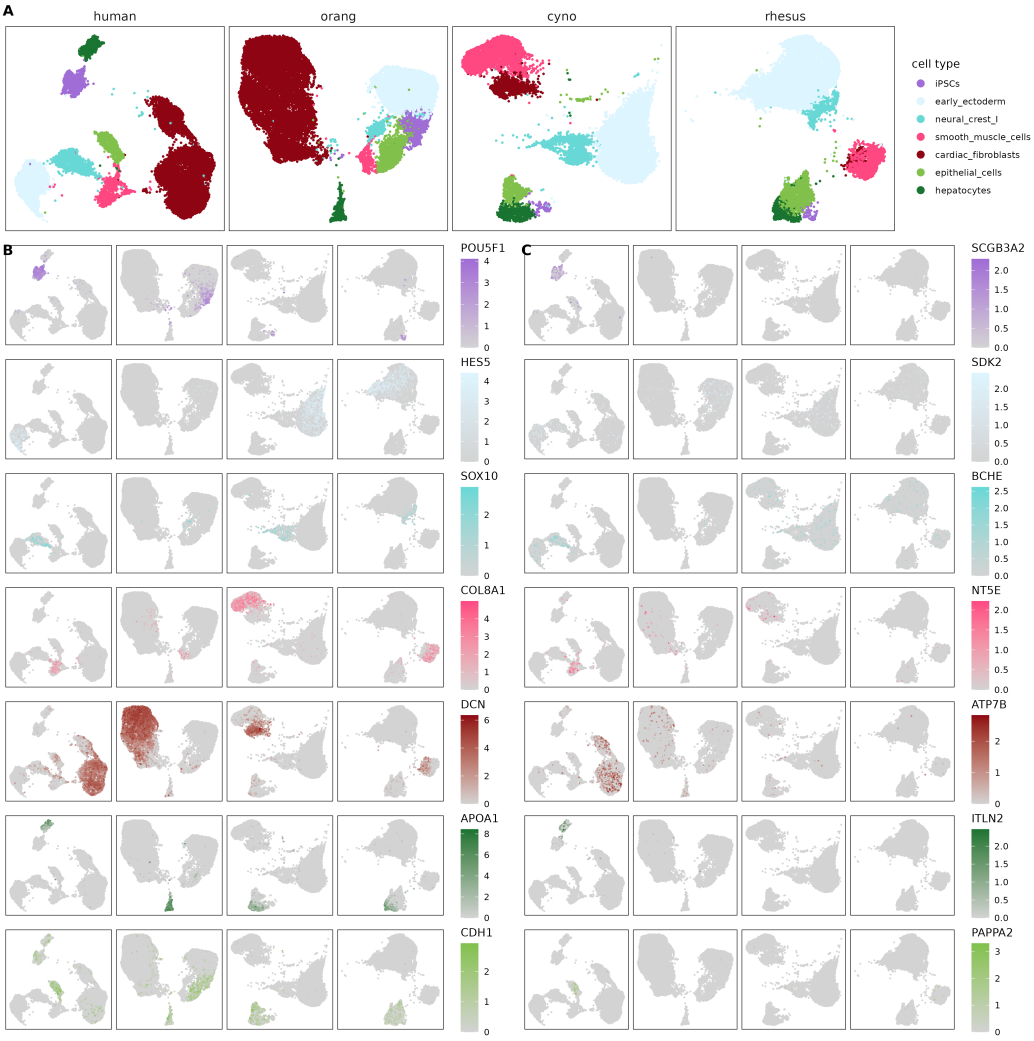
Supplementary Figure S8. Pseudotime analysis of mesoderm differentiation trajectories. A-D) PHATE embeddings of iPSCs and mesodermal cells from human (A), orangutan (B), cynomolgus (C) and rhesus macaque (D). E-F) Distribution of pseudotime values across major mesodermal lineages for each species. Boxplots illustrate shifts in pseudotime distributions between day 8 (d8) and day 16 (d16) of differentiation.



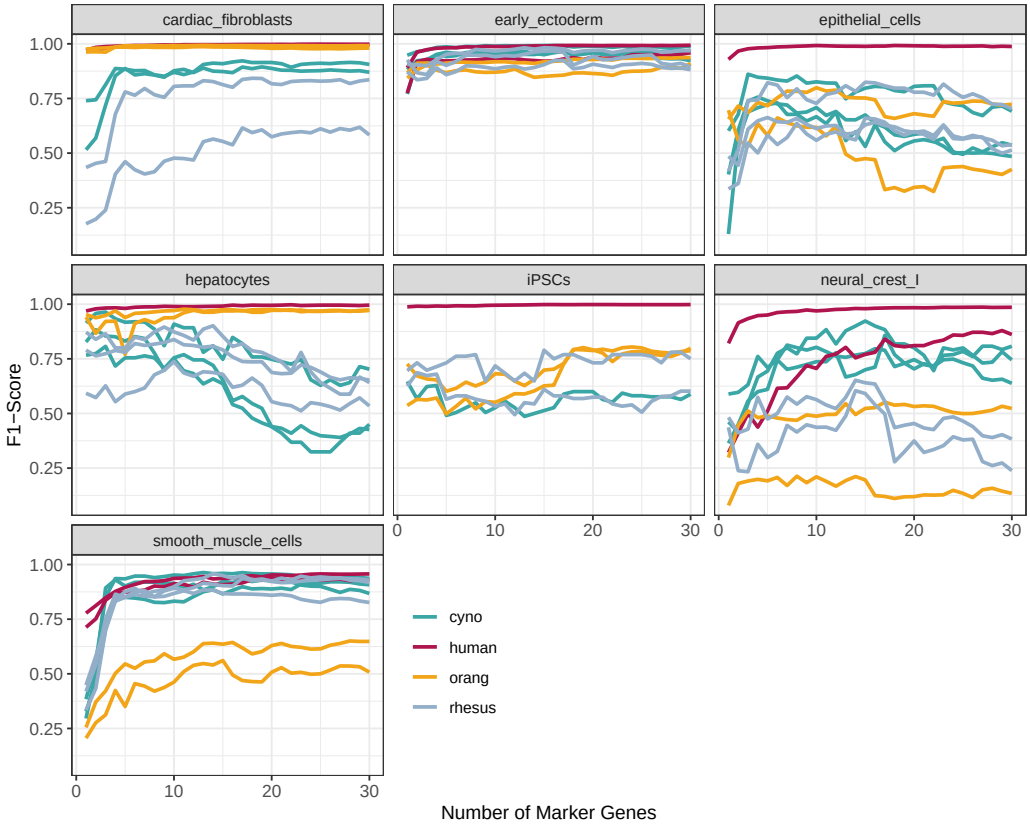
Supplementary Figure S9. Pseudotime analysis of endoderm differentiation trajectories. A-D) PHATE embeddings of iPSCs and endodermal cells from human (A), orangutan (B), cynomolgus (C) and rhesus macaque (D). E-F) Distribution of pseudotime values across major endodermal lineages for each species. Boxplots illustrate shifts in pseudotime distributions between day 8 (d8) and day 16 (d16) of differentiation.



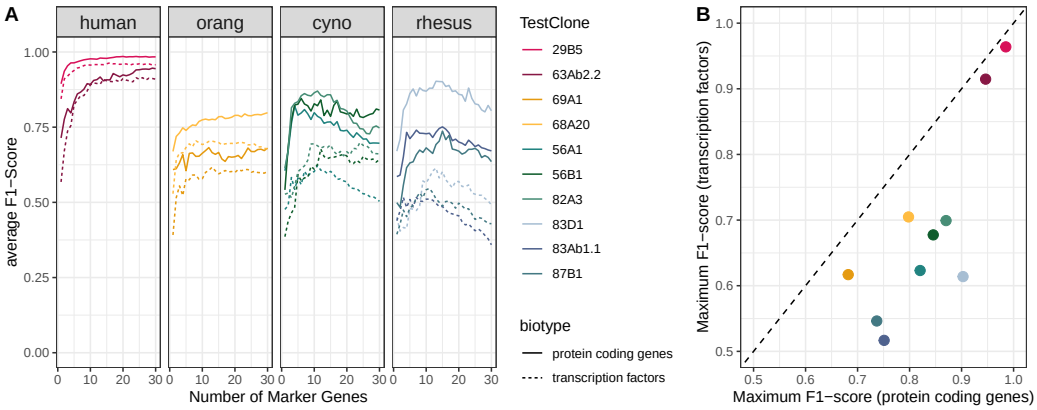
Supplementary Figure S10. Characteristics of genes with different levels of cell type-specific expression. A) Stacked bar plot of the number of genes per cell type specificity level for different species. B) Boxplot of expression conservation of genes with different levels of cell type specificity in orangutan, cynomolgus and rhesus. C) Boxplot of gene-level constraint based on primate phastCons scores [35] for protein-coding genes. D) Boxplot of mean expression per cell type for genes with different levels of cell type specificity. E) Boxplot of mean expression per cell type for a subset of 236 genes per cell type specificity and species that were sampled to have a similar distribution of mean expression. F) Boxplot of expression conservation of the same subsampled genesets as in E).



Supplementary Figure S11. Expression patterns of shared and human specific marker genes. A) UMAP representation per species filtered for the 7 cell types that are present in all 4 species. B) UMAP representations colored by the log-normalized expression of 7 representative marker genes that are shared among the top100 marker genes per cell type in all 4 species. C) UMAP representations colored by the log-normalized expression of 7 representative marker genes that are only present in the human top100 marker gene ranking per cell type.



Supplementary Figure S12. kNN classification performance per cell type. F1-score per cell type for a kNN-classifier trained in the human clone 29B5 to predict cell type identity based on the expression of 1-30 protein-coding marker genes. Each line represents the performance in a different clone, colored by species identity.



Supplementary Figure S13. kNN classification performance for transcription factors and protein coding marker genes. A) Average F1-score for a kNN-classifier trained in the human clone 29B5 to predict cell type identity in the other clones. The classifier is trained on the expression of the top 1-30 protein coding markers (solid lines) or transcription factor markers (dashed lines). B) Comparison of the maximum average F1-score between transcription factors and protein coding markers for the classifications depicted in A).

Supplementary Table S1. Marker genes. Literature review for marker genes used in human and mouse / rodents to determine a specific cell type.

Cell type	**Marker gene**	**used in human**	**used in mouse**
iPSCs	POU5F1	[58]	[59]
iPSCs	NANOG	[58]	[60]
iPSCs	L1TD1	[61]	[61]
early ectoderm	SOX2	[62]	[63]
early ectoderm	HES5	[64]	[65]
early ectoderm	RFX4	[64]	[66]
granule precursor cells	NFIA	[67]	[68]
granule precursor cells	ZIC1	[69]	[70]
granule precursor cells	ZIC4	[69]	[71]
neural crest	SOX10	[33]	[33, 72]
neural crest	FOXD3	[73]	[74]
neural crest	S100B	[75]	[76]
neurons	STMN2	[77]	[78, 79]
neurons	TAGLN3 (NP25)	[80]	[79]
neurons	DCX	[81]	[81]
smooth muscle cells	COL8A1	[82]	[83]
smooth muscle cells	ACTG2	[84]	[83]
smooth muscle cells	ACTA2	[82]	[83]
cardiac fibroblasts	TNNT2	[85]	[86]
cardiac fibroblasts	DCN	[87]	[88]
cardiac fibroblasts	HAND2	[85]	[89]
epithelial cells	CDH1	[90]	[91]
epithelial cells	EPCAM	[92]	[93]
epithelial cells	CLDN7	[94]	[95]
hepatocytes	TTR	[96]	[97]
hepatocytes	APOA1	[98]	[99]
hepatocytes	APOA2	[98]	[100]