Improving the methodological basis of cross-species scRNA-seq analysis

Dissertation an der Fakultät für Biologie

der Ludwig-Maximilians-Universität München

Philipp Janßen

München 2025

Improving the methodological basis of cross-species scRNA-seq analysis

Dissertation an der Fakultät für Biologie

der Ludwig-Maximilians-Universität München

Philipp Janßen

München 2025

Diese Dissertation wurde angefertigt
unter der Leitung von PD Dr. Ines Hellmann
an der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

Erstgutachter:	PD Dr. Ines Hellmann
Zweitgutachter:	Prof. Dr. Korbinian Schneeberger
Tag der Abgabe:	11.04.2025
Tag der mündlichen Prüfung:	09.10.2025

Eigenständigkeitserklärung und Erklärung

Eigenständigkeitserklärung

Hiermit versichere ich an Eides statt, dass die vorliegende Dissertation von mir selbstständig verfasst wurde und dass keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden. Die Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinne nach entnommen sind, wurden in jedem Fall unter Angabe der Quellen (einschließlich des World Wide Web und anderer elektronischer Text- und Datensammlungen) kenntlich gemacht. Weiterhin wurden alle Teile der Arbeit, die mit Hilfe von Werkzeugen der künstlichen Intelligenz de novo generiert wurden, durch Fußnote/Anmerkung an den entsprechenden Stellen kenntlich gemacht und die verwendeten Werkzeuge der künstlichen Intelligenz gelistet. Die genutzten Prompts befinden sich im Anhang. Diese Erklärung gilt für alle in der Arbeit enthaltenen Texte, Graphiken, Zeichnungen, Kartenskizzen und bildliche Darstellungen.

München, den 11.04.2025

Philipp Janßen

Erklärung

Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich anderweitig einer Doktorprüfung ohne Erfolg **nicht** unterzogen habe.

München, den 11.04.2025

Philipp Janßen

Contents

Abbreviations **xi**

Publications **xii**

Declaration **xvi**

Summary **1**

1 Introduction **3**

1.1 Cell type characterization and identification using single-cell transcriptomics 4

1.1.1 Measuring gene expression 4

1.1.2 The rise of single-cell RNA sequencing 5

1.1.3 Computational analysis of scRNA-seq data 6

1.1.4 Approaches for cell type annotation 9

1.1.5 Relevance of marker genes 11

1.2 Evaluate technical artifacts with species-mixing experiments . . 14

1.2.1 Technological confounders of cell type identification . . . 14

1.2.2 Species-mixing experiments 15

1.3 Comparative single-cell studies and cross-species analysis 18

1.3.1 Evolutionary cell type definition 18
1.3.2 Cross-species comparisons in primates 19
1.3.3 Generation and characterization of primate iPSCs 20
1.3.4 iPSC-derived organoid systems 22
1.3.5 Computational challenges of cross-species analysis 23
1.4 Aims of the thesis 28

2 Results **29**

2.1 The effect of background noise and its removal on the analysis of single-cell expression data 31
2.2 A non-invasive method to generate induced pluripotent stem cells from primate urine 69
2.3 Generation and characterization of three fibroblast-derived Rhesus Macaque induced pluripotent stem cells 85
2.4 Generation and characterization of two Vervet monkey induced pluripotent stem cell lines derived from fibroblasts 93
2.5 Generation and characterization of two fibroblast-derived Baboon induced pluripotent stem cell lines 101
2.6 Identification and comparison of orthologous cell types from primate embryoid bodies shows limits of marker gene transferability 109

3 Discussion **169**

3.1 The power of genetic variants in transcriptomic experiments . . 170
3.1.1 Enhancing multiplexing in cross-species studies 170
3.1.2 Authentication of cell lines 172

3.1.3 Genetic variants as natural barcodes in cell-mixing experiments 173
3.2 Marker genes - fragile cornerstones of scRNA-seq analysis 175
3.2.1 Susceptibility to background noise 176
3.2.2 Limited transferability across species 177
3.3 Cell type assignment across species 178
3.3.1 Classification of bulk RNA-seq data 179
3.3.2 Orthologous cell type assignment from scRNA-seq data . 179

4 Conclusion and Outlook **181**

Bibliography **185**

List of Figures **208**

Acknowledgements **211**

Abbreviations

Abbreviation	Definition
AUC	area under the curve
CD	Cluster of Differentiation
cDNA	complementary DNA
CoRC	core regulatory complex
DE	differential expression
DNA	deoxyribonucleic acid
EB	embryoid body
ESC	embryonic stem cell
GSEA	Gene Set Enrichment Analysis
iPSC	induced pluripotent stem cell
lncRNA	long non-coding RNA
mRNA	messenger RNA
NHP	non-human primate
PBMC	peripheral blood mononuclear cell
PCA	principal component analysis
RNA	ribonucleic acid
RNA-seq	RNA-sequencing
RT-qPCR	Real-time quantitative PCR
scRNA-seq	single-cell RNA sequencing
SNP	single nucleotide polymorphism
snRNA-seq	single-nucleus RNA sequencing
STR	Short Tandem Repeat
TF	transcription factor
UMAP	Uniform Manifold Approximation and Projection
UMI	unique molecular identifier
WES	whole-exome sequencing
WGS	whole-genome sequencing

Chronological List of Publications

I. Geuder J, Wange LE, Janjic A, Radmer J, **Janssen P**, Bagnoli JW, Müller S, Kaul A, Ohnuki M, Enard W:
"A non-invasive method to generate induced pluripotent stem cells from primate urine."
Scientific Reports 11, 3516 (2021). doi: 10.1038/s41598-021-82883-0

II. **Janssen P**, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, McCabe C, Heyn H, Levin JZ, Enard W, Hellmann I:
"The effect of background noise and its removal on the analysis of single-cell expression data."
Genome Biology 24, 140 (2023). doi: 10.1186/s13059-023-02978-x

III. Jocher J, Edenhofer FC, **Janssen P**, Müller S, Lopez-Para DC, Geuder J, Enard W:
"Generation and characterization of three fibroblast-derived Rhesus Macaque induced pluripotent stem cell lines."
Stem Cell Research 74, 103277 (2023). doi: 10.1016/j.scr.2023.103277

IV. Jocher J, Edenhofer FC, Müller S, **Janssen P**, Briem E, Geuder J, Enard W:
"Generation and characterization of two Vervet monkey induced pluripotent stem cell lines derived from fibroblasts."
Stem Cell Research 75, 103315 (2024). doi: 10.1016/j.scr.2024.103315

V. Jocher J, Edenhofer FC, Müller S, **Janssen P**, Briem E, Geuder J, Enard W:
"Generation and characterization of two fibroblast-derived Baboon induced pluripotent stem cell lines."
Stem Cell Research 75, 103316 (2024). doi: 10.1016/j.scr.2024.103316

VI. Jocher J and **Janssen P**, Vieth B, Edenhofer FC, Dietl T, Térmeg A, Spurk P, Geuder J, Enard W, Hellmann I:
"Identification and comparison of orthologous cell types from primate embryoid bodies shows limits of marker gene transferability."
Reviewed Preprint at *eLife* 14:RP105398 (2025). doi: 10.7554/eLife.105398.1

Other Publications

VII. Kälin RE, Cai L, Li Y, Zhao D, Zhang H, Cheng J, Zhang W, Wu Y, Eisenhut K, **Janssen P**, Schmitt L, Enard W, Michels F, Flüh C, Hou M, Kirchleitner SV, Siller S, Schiemann M, Andrä I, Montanez E, Giachino C, Taylor V, Synowitz M, Tonn JC, von Baumgarten L, Schulz C, Hellmann I, Glass R:
"TAMEP are brain tumor parenchymal cells controlling neoplastic angiogenesis and progression."
Cell systems 12, 248-262 (2021). doi: 10.1016/j.cels.2021.01.002

Declarations of contribution as a first-author

The effect of background noise and its removal on the analysis of single-cell expression data

Ines Hellmann, Wolfgang Enard, and I conceptualized this study. Ines Hellmann and I wrote the original draft. I, Beate Vieth, and Zane Kliesmete conducted the formal analysis. Sean Simmons did the data curation. Xian Adiconis, Jamie Marshall, and Cristin McCabe performed the experiments. Joshua Z. Levin supervised the experiments. Wolfgang Enard, Holger Heyn, and Joshua Z. Levin acquired funding.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the above contributions of Philipp Janßen to these publications.

Ines Hellmann

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

Wolfgang Enard and Ines Hellmann conceived the study. Jessica Jocher optimized and conducted EB differentiation experiments and performed 10x scRNA-seq data generation with support of Fiona C. Edenhofer. Johanna Geuder generated and provided human and orangutan iPSCs and supported optimization of EB differentiation protocols. Paulina Spurk established FACS analyses of EBs. I and Jessica Jocher did primary data analysis. I did the pre-processing of the data, developed the pipeline for orthologous cell type assignment, and created the Shiny app. I and Beate Vieth performed the cell type specificity and marker gene conservation analysis. Anita Térmeg prepared reference genomes for non-human primates. Tamina Dietl supported cell type annotation. I, Jessica Jocher and Ines Hellmann wrote the manuscript.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the above contributions of Philipp Janßen to these publications.

Jessica Jocher

Ines Hellmann

Wolfgang Enard

Declarations of contribution as a co-author

A non-invasive method to generate induced pluripotent stem cells from primate urine

Johanna Geuder, Mari Ohnuki and Wofgang Enard conceived the study. I helped with analysis of bulk RNA-seq data. Johanna Geuder and Wolfgang Enard wrote the manuscript.

Generation and characterization of three fibroblast-derived Rhesus Macaque induced pluripotent stem cell lines

Wolfgang Enard and Jessica Jocher conceived the study. I and Jessica Jocher analyzed the scRNA-seq data. Dana C. Lopez-Parra and I performed variant calling for authentication. Wolfgang Enard and Jessica Jocher wrote the manuscript.

Generation and characterization of two Vervet monkey induced pluripotent stem cell lines derived from fibroblasts

The study was conceived by Wolfgang Enard and Jessica Jocher. I performed variant calling from bulk RNA-seq data. Wolfgang Enard and Jessica Jocher wrote the manuscript.

Generation and characterization of two fibroblast-derived Baboon induced pluripotent stem cell lines

Wolfgang Enard and Jessica Jocher conceived the study. I performed variant calling from bulk RNA-seq data. Wolfgang Enard and Jessica Jocher wrote the manuscript.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the above contributions of Philipp Janßen to these publications.

Ines Hellmann

Summary

Single-cell RNA sequencing (scRNA-seq) has become a powerful method to explore cell type diversity and gene expression at unprecedented resolution. Extending this approach across species not only enables the identification of conserved and species-specific cell types, but also provides insight into how cellular programs evolve. Comparative single-cell studies in primates are especially valuable for understanding the molecular changes that underlie human-specific traits within an evolutionary framework. However, meaningful cross-species comparisons rely not only on the availability of single-cell data from different organisms, but also on robust data quality, well-matched cellular systems and appropriate computational frameworks for integration. This thesis addresses key challenges in cross-species single-cell transcriptomics, with a focus on improving the methodological foundation for comparative studies in primates.

Ensuring good data quality is essential for all single-cell studies and becomes even more important when comparing data across species. Yet technical artifacts are not uncommon and can obscure biological signal and complicate data interpretation. One such artifact is background noise, which originates from cell-free ambient RNA or barcode swapping events. To evaluate the extent and impact of background noise in 10x Genomics data, I established a benchmarking dataset generated from pooled kidney cells of two mouse subspecies. I used naturally occurring genetic variants to determine the origin of individual reads and identify transcripts that were incorrectly assigned to a cell barcode to quantify background noise. I found that background levels varied substantially between cells and replicates, with ambient RNA identified as the primary source. This noise particularly compromises the detection

of marker genes, reducing their specificity. Furthermore, I evaluated several computational methods for noise correction and found that most approaches improved marker detection, with CellBender showing the strongest performance. These findings help characterize the nature of background noise and provide practical guidance for its mitigation in future single-cell studies.

Besides accurate measurements, cross-species single-cell studies also rely on access to comparable cellular material. For primates in particular, obtaining such material remains a challenge. In this context, induced pluripotent stem cells (iPSC) and their derivates offer a powerful resource for comparative studies. I contributed to the characterization of newly established iPSC lines from various non-human primates (NHP), including vervet monkeys, baboons, rhesus macaques, gorillas and orangutans. My contributions focused on validating the pluripotency and identity of these cell lines using bulk and single-cell RNA-seq data. On the one hand, I helped to classify primary cells, iPSCs and derived cell types based on their expression profiles. On the other hand, I called genetic variants from RNA-seq data for authentication of the cell lines.

Finally, I analysed a cross-species dataset of embryoid bodies (EB) derived from human and NHP iPSCs to enable comparative analyses of early primate development. This dataset includes four species and spans a wide range of different cell types. To identify orthologous cell types in this complex setting, I developed a semi-automated pipeline combining classification and manual annotation steps. Based on these annotations I investigated cross-species conservation of gene expression, with a particular focus on the transferability of marker genes. The results showed that while broadly expressed genes are relatively well conserved, many cell type-specific marker genes are less transferable across species. These findings underscore the challenges of cell type annotation in cross-species settings and provide a curated dataset and computational approach to support future comparative analyses in primates.

1 Introduction

Over the past decade, advances in transcriptomic technologies have dramatically increased the resolution and sensitivity with which cellular states can be measured. In particular, the ability to profile gene expression at single-cell resolution has transformed how we study cellular diversity and define cell types. Including multiple species in single-cell analyses not only places cell type characterizations in a broader biological context, but also enables direct comparisons across species to study evolutionary change. In this thesis, I explore how multi-species single-cell transcriptomic data can be used to improve cell type characterization and address some challenges associated with such analyses.

First, in the section **Cell type characterization and identification using single-cell transcriptomics**, I describe the basic principles and commonly used strategies for identifying and defining cell types based on gene expression profiles.

In the section **Evaluating technical artifacts with species-mixing experiments**, I focus on technical confounders such as background RNA contamination that can distort cell type assignments, and explain how controlled species-mixing experiments can be used to assess the magnitude and impact of these artifacts.

Finally, in the section **Comparative single-cell studies and cross-species analysis**, I explore how including multiple species - particularly non-human primates - can help refine our understanding of cell types, and discuss the conceptual and methodological challenges involved in comparing transcriptomic data across species.

1.1 Cell type characterization and identification using single-cell transcriptomics

1.1.1 Measuring gene expression

Cells are the fundamental structural and functional units of life. All cells in a multicellular organism carry essentially the same genome, yet they can develop into remarkably diverse cell types. This diversity is driven not by differences in genetic content, but by differential gene regulation, which determines which genes are active in a given cell type. Since active genes are transcribed into RNA, measuring RNA levels provides a direct insight into gene activity. Messenger RNA (mRNA), in particular, is the intermediary between genes and proteins, so its abundance offers an informative readout of gene expression at a given moment (Lowe et al. 2017). Before high-throughput genomics, gene expression was measured one gene at a time. One of the earliest approaches was Northern blotting, which detects specific RNA molecules by separating them via gel electrophoresis and hybridizing them to labeled probes (Alwine et al. 1977). Real-time quantitative PCR (RT-qPCR) later improved sensitivity by converting mRNA into complementary DNA (cDNA) and amplifying target sequences (Heid et al. 1996). To go beyond the analysis of single genes, the introduction of microarrays enabled the parallel profiling of hundreds to thousands of genes by hybridizing labeled cDNA to pre-spotted DNA sequences (Schena et al. 1995).

The introduction of RNA sequencing (RNA-seq) in the late 2000s transformed transcriptomics by enabling high-throughput, unbiased gene expression measurement (Mortazavi et al. 2008). The method involves converting RNA into cDNA, fragmenting it, and sequencing millions of reads, which are then mapped back to the genome to quantify transcript levels. Unlike microarrays, RNA-seq is not restricted to predefined probes, allowing the genome-wide quantification of RNA variants. In its standard form, bulk RNA-seq is performed on RNA extracted from a population of cells, generating a single, averaged gene expression profile. This approach is well-suited for identifying global expression patterns and comparing gene activity between conditions but does not distinguish contributions from individual cell types within a mixed sample (Trapnell 2015).

1.1.2 The rise of single-cell RNA sequencing

Motivated by the scarcity of biological material in contexts like embryonic development (Kolodziejczyk et al. 2015) and the limitations of bulk RNA-seq, which masks cellular heterogeneity through averaging (Trapnell 2015), researchers began developing RNA-seq protocols with single-cell resolution to enable more precise characterization of cell types and states.

The first whole transcriptome mRNA measurements from a single cell were achieved in 2009 by Tang et al. (2009). In the years that followed, advances in technology led to an almost exponential increase in the number of cells that could be profiled in a single experiment (Svensson et al. 2018). A key advance was the introduction of early barcoding strategies, which allowed multiple single cells to be processed simultaneously (Islam et al. 2011). This was soon complemented by improvements in cell isolation and capture techniques, paving the way for high-throughput protocols. Two main approaches emerged: plate-based methods, which capture individual cells in microwell plates (Picelli et al. 2014) and droplet-based methods, which encapsulate single cells in nanoliter emulsions (Macosko et al. 2015). The introduction of unique molecular identifiers (UMI) around the same time improved the accuracy of transcript quantification by correcting for amplification bias, making high-throughput protocols more reliable (Islam et al. 2014). Within a decade, further improvements in throughput and sensitivity, along with the decreasing cost of next-generation sequencing, have enabled a drastic scale-up from profiling tens or hundreds of cells to millions (Cao et al. 2019a).

Today, scRNA-seq technologies come in many variations, but they all follow the same core workflow (Kolodziejczyk et al. 2015; Ziegenhain et al. 2017): single cells are first isolated, followed by the reverse transcription of mRNA into cDNA. The cDNA is then amplified and finally prepared into libraries for next-generation sequencing to generate transcriptomic data. While most scRNA-seq methods rely on whole-cell capture, an alternative approach, single-nucleus RNA sequencing (snRNA-seq), isolates and profiles mRNA from cell nuclei instead. This method is particularly useful for studying frozen or hard-to-dissociate tissues, where intact cells are difficult to obtain (Lake et al. 2016).

With these technical foundations in place, scRNA-seq has rapidly become a key tool across diverse areas of biological and biomedical research. One of its most significant applications is cell atlas projects, which aim to catalog the diversity of cell types in different organisms. Large-scale efforts like the Human Cell Atlas (Regev et al. 2017) and Tabula Muris (Tabula Muris Consortium et al. 2018) have mapped gene expression across many tissues, providing a reference for understanding how cells function in health and disease. In biomedical research, scRNA-seq has been instrumental in studying diseases at the cellular level. In cancer research, for example, it has helped to dissect tumor heterogeneity (González-Silva et al. 2020; Wu et al. 2021) and highlight interactions within the tumor microenvironment (Ren et al. 2021; Bridges and Miller-Jensen 2022). In immunology, it has provided insights into immune cell states and responses to infection, including COVID-19 (Liao et al. 2020).

1.1.3 Computational analysis of scRNA-seq data

Alongside the rapid growth of scRNA-seq technologies, data and applications, there has been an equally fast expansion of computational tools for data analysis. As of 2025, more than 1800 methods had already been developed for various types of scRNA-seq analysis (scRNA-tools.org n.d.). In principle, the analysis workflow can be divided into three main stages (Figure 1): 1) raw data processing, 2) pre-processing of the count matrix and 3) downstream analysis at either the cell or gene level.

Firstly, raw sequencing data are processed to generate a count matrix (Figure 1A). This begins with quality control of the sequencing reads, where low-quality sequences and adapter contamination are removed. Next, the reads are aligned to a matching reference genome. For protocols which use UMIs to correct for PCR amplification bias, an additional step involves collapsing reads that share the same UMI. Finally, gene expression is quantified by counting the number of reads or UMIs assigned to each gene for each cell barcode. To automate these steps, analysis pipelines such as Cell Ranger (Zheng et al. 2017) and zUMIs (Parekh et al. 2018) are commonly used. The output is a gene-by-cell matrix, which forms the basis for all subsequent analysis.

Next, several pre-processing steps are required to prepare the count matrix for downstream

analysis (Figure 1B). This involves quality control and filtering at both the cell and gene level. Low quality cells are typically identified and removed based on thresholds on the number of counts and genes detected per cell, as well as the fraction of counts from mitochondrial genes (Ilicic et al. 2016). Furthermore, doublets - instances where more than one cell is captured and measured together - can be detected based on their expression profile (McGinnis et al. 2019a; DePasquale et al. 2019; Wolock et al. 2019). Gene-level filtering is also commonly applied to remove genes with very low or no expression, which can introduce noise and inflate data sparsity (Luecken and Theis 2019). In addition, ambient RNA removal methods perform quality control and correction directly on the count level to correct for contamination with cell-free or wrongly assigned RNA molecules (Fleming et al. 2023; Yang et al. 2020; Young and Behjati 2020). Following these filtering and correction steps, the count matrix then needs to be normalized to account for sampling differences between individual cells. This step has been shown to be particularly crucial for some downstream analysis (Vieth et al. 2019). If cells were handled in different groups or experiments, additional batch correction and data integration steps might be necessary (Haghverdi et al. 2018; Stuart et al. 2019; Korsunsky et al. 2019). Finally, for many downstream analyses it is helpful to reduce the dimensionality of the data. In this context, feature selection aims to reduce noise by keeping only a set of most informative genes. In contrast, dimensionality reduction condenses the expression space into a smaller set of components. Linear methods like principal component analysis (PCA) are commonly used for this purpose in order to summarize the data for other analysis steps, while non-linear approaches like Uniform Manifold Approximation and Projection (UMAP) are primarily for visualization (Luecken and Theis 2019).

Once the count matrix has been processed, a wide range of analytical approaches can be applied to explore patterns in the data (Jovic et al. 2022; Luecken and Theis 2019) (Figure 1C). A key step in most downstream workflows, however, is the grouping of cells into clusters or distinct cell types. Beyond this step, which will be discussed in detail in a later section, several other cell-level analyses can provide complementary insights. Compositional analysis evaluates shifts in cell type or state proportions across conditions (Cao et al. 2019b). Cell–cell communication tools infer signaling interactions based on ligand-receptor expression (Efremova et al. 2020). When discrete categorization of cells is not sufficient, trajectory

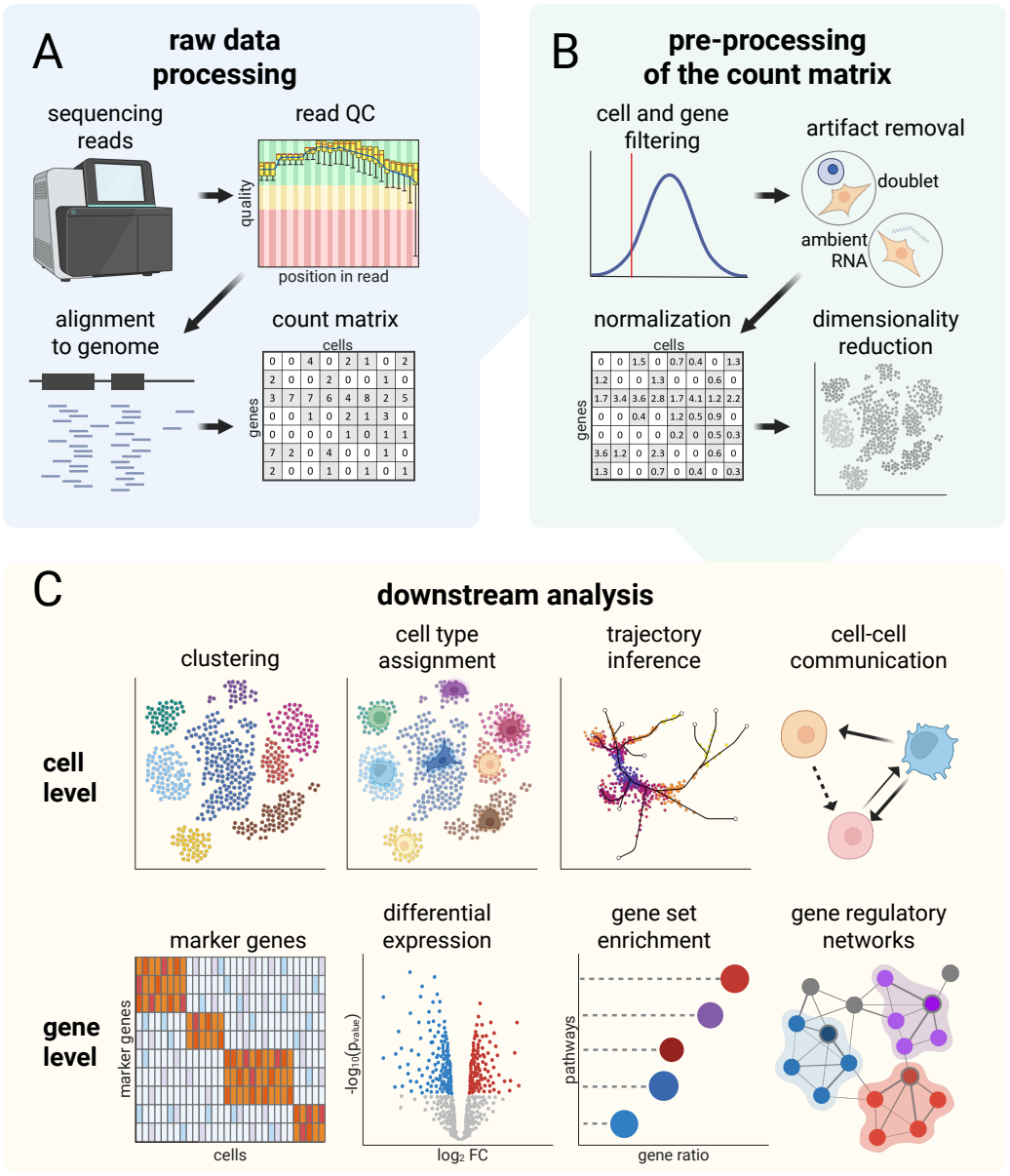


Figure 1. Computational workflow for scRNA-seq analysis. The analysis is broadly divided into three stages: (A) Raw data processing, including read quality control, alignment to a reference genome, and quantification to generate a count matrix; (B) Pre-processing of the count matrix, involving filtering of low-quality cells and genes, doublet detection, ambient RNA correction, normalization, and dimensionality reduction; (C) Downstream analysis, which includes both cell-level approaches such as clustering, cell type annotation, trajectory inference and cell-cell communication analysis, and gene-level analyses like marker gene identification, DE analysis, gene set enrichment analysis, and inference of gene regulatory networks. Created with BioRender.com

inference methods can reconstruct cellular transitions (Trapnell et al. 2014; Street et al. 2018).

Gene-level analyses are often performed in parallel. Marker gene identification is used to pinpoint genes that define specific populations (Pullin and McCarthy 2024), while differential expression (DE) analysis identifies genes that vary between conditions (Finak et al. 2015; Vieth et al. 2019). Gene set enrichment analysis (GSEA) can then be applied to identify overrepresented biological processes in long candidate gene lists (Kuleshov et al. 2016). In addition, gene regulatory network inference aims to uncover interactions between transcription factors and their downstream targets which helps characterize the regulatory programs underlying cellular identity (Aibar et al. 2017).

1.1.4 Approaches for cell type annotation

Despite the wide range of possible downstream analyses, most rely on a common foundation: the classification of cells into biologically meaningful groups. Accurate cell type annotation is thus a central task in single-cell analysis. There are two main strategies to achieve this task: unsupervised clustering followed by manual annotation and supervised classification with a reference dataset (Sun et al. 2022).

Unsupervised approaches start by grouping cells with a similar expression profile into discrete clusters. For this task, community-detection algorithms like Louvain are commonly employed (Luecken and Theis 2019). The resulting clusters are subsequently annotated based on the expression of individual genes. Candidate gene lists are generated by comparing expression patterns across clusters and are then matched to known marker genes from the literature or databases, which exhibit cell type-specific expression patterns (Wang et al. 2020). If a cluster shows high expression of a well-established marker gene or a characteristic set of genes associated with a particular cell type, it is assigned the corresponding label. This manual approach allows for flexible interpretation, enabling the identification of novel or unexpected cell types. However, it is time-consuming and relies on expert knowledge, making it inherently subjective and susceptible to bias.

In contrast, supervised methods assign cell type labels by comparing query cells to a

pre-annotated reference dataset. These approaches can be broadly sub-categorized into classification methods and integration-based label transfer. Classification methods use correlation measures or machine learning techniques to predict cell types for single cells in the query based on expression profiles in the reference (Pasquini et al. 2021). They offer a fast and automated solution, but their accuracy depends heavily on the quality and completeness of the reference dataset. When dealing with novel or rare cell types that are not well represented in the reference, classification methods may default to an "unassigned" label, but also risk incorrect assignments (Abdelaal et al. 2019). Integration-based label transfer methods, on the other hand, align the query data with a reference dataset before assigning labels to query cells based on their neighborhood in the integrated space (Stuart et al. 2019; Lotfollahi et al. 2022). They can be more robust than classification methods, as they also account for relationships between query cells rather than assigning labels independently. However, if the reference and query datasets are poorly matched, particularly when cell type compositions differ significantly, these methods may misalign cell populations, leading to inaccurate annotations. While both approaches enhance reproducibility compared to manual annotation, their effectiveness hinges on the availability of a well-matched, high-quality reference dataset.

Overall, cell type annotation is rarely a straightforward process, and relying on a single approach or method is unlikely to produce fully reliable results. When using reference-based supervised methods, combining annotations across multiple reference datasets (Yuan et al. 2022) and multiple computational tools (Ergen et al. 2024) can improve accuracy. This not only results in a more robust consensus annotation, but importantly also highlights areas of uncertainty that may require further investigation. Furthermore, annotations should not be taken at face value and should undergo careful validation using a combination of manual and supervised approaches. For instance, Clarke et al. (2021) recommend a three-step workflow: 1) automatic annotation to assign initial labels based on available reference data, 2) manual annotation to review and refine these labels by assessing marker gene expression and 3) verification through additional experiments, statistical analyses or expert consultation.

1.1.5 Relevance of marker genes

During cell type annotation, the examination of marker gene expression plays a key role at several stages. Whether used for manual cluster annotation, refining and verifying existing labels, or selecting candidate genes for experimental validation, marker genes remain essential throughout the process.

Here, I refer to marker genes as genes with specific expression patterns, allowing for clear distinction between different cell populations. In this role they have already been relevant in cellular biology long before the emergence of scRNA-seq. Historically, surface proteins such as Cluster of Differentiation (CD) markers have been widely used for cell classification via flow cytometry (Maecker et al. 2012) or immunohistochemistry (Lyck et al. 2008). With the advent of transcriptomics, marker discovery has expanded beyond surface proteins to include any gene exhibiting sufficiently discriminatory expression patterns. An ideal marker gene should meet several criteria: 1) specificity, meaning it is highly expressed in the target cell type while showing minimal expression elsewhere (Pullin and McCarthy 2024); 2) stability and replicability, ensuring consistent expression in this cell type across conditions and datasets (Fischer and Gillis 2021); and 3) detectability, meaning that, in the case of scRNA-seq, its mRNA levels are high enough at the time of measurement to be reliably detected despite the dropout events inherent to this technology (Hicks et al. 2018). In many cases, strong marker genes are also biologically relevant for the function or identity of the cell type, making them especially meaningful.

Given these properties, marker genes play an important role in characterizing the heterogeneity within a dataset. On the one hand, they serve as a reference point for known cell types. On the other hand, they are also essential for describing novel or poorly characterized cell types. To systematically identify marker genes, gene expression is compared across clusters or cell types to determine which genes exhibit discriminatory expression patterns. This is typically done using one-vs-rest comparisons, where each cluster is tested against all others, or alternatively pairwise comparisons that evaluate differences between individual cluster pairs (Pullin and McCarthy 2024). One of the most widely used strategies is DE-based analysis. Common statistical tests include the Wilcoxon rank-sum test, Student's

t-test or logistic regression. These tests are readily available in a one-vs-rest setting in standard analysis frameworks like Seurat and scanpy, making them widely popular. Despite their simplicity, they have proven to be highly effective in a recent benchmark (Pullin and McCarthy 2024). Pairwise testing, such as implemented in scran, aims to make the DE testing more independent of the overall cell type composition (Amezquita et al. 2020). Another intuitive strategy is presence/absence scoring, which identifies genes that are consistently detected in one population but absent in others. By focusing on detection alone rather than quantitative expression levels, this binary approach can be a simple yet effective alternative to DE methods. Beyond traditional DE testing and presence/absence scoring, feature selection and machine learning-based approaches provide an alternative way to define marker genes. Examples include NS-Forest (Liu et al. 2024), which uses Random Forest classifiers, SMaSH (Nelson et al. 2022), a deep learning-based selection method, and RankCorr (Vargo and Gilbert 2020), which ranks genes based on correlation patterns. While these methods can improve marker selection, they are computationally intensive and do not always outperform simpler DE-based approaches (Pullin and McCarthy 2024).

Regardless of the method used for identification, an additional challenge is determining how many markers should be selected to best define a cell type. Using p-value cut-offs can be problematic, as pseudo-replication can inflate false discovery rates (Squair et al. 2021). Instead, markers are often ranked based on effect sizes such as log fold change, Cohen's d or area under the curve (AUC), with a fixed number of top genes selected (Amezquita et al. 2020). Binary metrics that compare detection rates between the target cell type and others can also be useful for ranking markers. For example, Fischer and Gillis (2021) used a detection rate-based approach to evaluate the signal-to-noise ratio of marker genes in combination with measures for coverage and replicability and found that around 50 to 200 markers were optimal for defining cortical cell types.

After identifying marker genes within a dataset, the next step is to compare them to previously established markers to better understand their biological significance. To support this, thousands of marker genes have been compiled into publicly available databases like PanglaoDB (Franzén et al. 2019) and CellMarker (Zhang et al. 2019; Hu et al. 2023), which provide curated lists of established cell type-specific markers for human and mouse.

Collections of large-scale single-cell datasets also serve as valuable resources for marker-based annotation (CZI Cell Science Program et al. 2025). Beyond databases, a literature search for marker genes reported and validated in published studies can also yield useful references. However, caution is needed when using markers identified through different technologies or experimental approaches. For instance, protein markers do not always correlate well with RNA expression (Stoeckius et al. 2017). Similarly, signatures derived from bulk RNA-seq data may be difficult to detect in scRNA-seq (Noureen et al. 2022). Additionally, marker lists from different sources can be inconsistent, with significant variation between databases and studies (Franzén et al. 2019; Zhang et al. 2019; Clarke et al. 2021). Given these challenges, careful validation is essential when incorporating external marker genes into single-cell analyses.

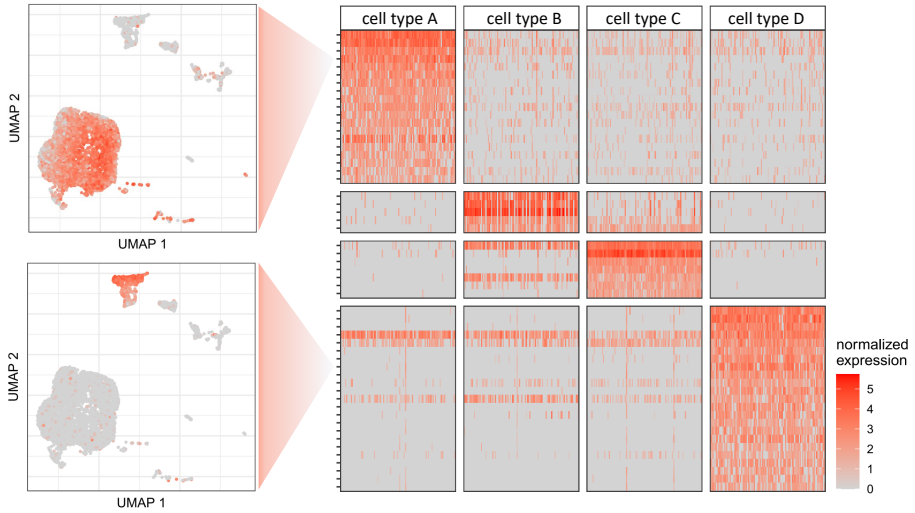


Figure 2. Visualization of marker genes to support cell type annotation. UMAPs show the expression of individual markers across cells (left), while heatmaps summarize multiple markers across cell types (right). Adapted from Janssen et al. (2023).

Finally, marker genes also serve a practical purpose: they help simplify and communicate the results of scRNA-seq analyses (Clarke et al. 2021). Given that single-cell datasets contain thousands of genes across thousands of cells, raw expression data can be overwhelming. Distilling this complexity into a set of key marker genes enhances interpretability. An important aspect of this simplification is visualization. Individual marker genes are often visualized by overlaying their expression on a 2D-embedding like UMAP, while heatmaps

and dot plots provide an intuitive summary for multiple marker genes across cell types (Figure 2). In this function, marker genes not only help in annotation, but also enhance reproducibility and clarity when presenting the results, ultimately concluding the cell type annotation process.

1.2 Evaluate technical artifacts with species-mixing experiments

1.2.1 Technological confounders of cell type identification

For the annotation of cell types, the base assumption is that reads that are associated with the same cell barcode originate from a single cell. However, certain technical artifacts introduced during the experimental workflow can challenge this assumption and impact downstream analysis. One notable artifact is the formation of doublets that occurs when two cells are captured within the same droplet. Due to the stochastic nature of cell encapsulation in droplet-based scRNA-seq technologies, a fraction of the droplets will contain two or more cells, proportional to the loaded cell density (Germain et al. 2021). For instance, in a typical 10x Genomics scRNA-seq experiment recovering 10,000 cells, this fraction is estimated to be approximately 8% (10xgenomics.com n.d.). Doublets can also arise from incomplete cell dissociation during sample preparation, where cells remain physically attached in the suspension (Schiebout et al. 2023). Doublets can be classified into two types: homotypic doublets, involving two cells of the same transcriptional state, and heterotypic doublets, involving two cells with distinct transcriptional states (Germain et al. 2021). While homotypic doublets are less disruptive as their combined transcriptional profiles may resemble that of a single cell, heterotypic doublets are particularly problematic. These hybrid profiles can introduce spurious cell clusters, creating the false appearance of novel or transitional cell types (Xi and Li 2021; Wolock et al. 2019; McGinnis et al. 2019a). To address this issue,

several computational tools have been developed to identify and remove doublets (Bais and Kostka 2020; DePasquale et al. 2019; McGinnis et al. 2019a; Wolock et al. 2019; Bernstein et al. 2020; Weber et al. 2021; Zhang et al. 2023a). These tools use various strategies, including artificial doublet simulation and gene expression similarity metrics, to flag and remove suspected doublets, thereby improving the reliability of downstream analyses.

In addition to doublets, even droplets containing a single cell can be affected by background noise of transcripts attributed to the wrong cell. This noise partly originates from cell-free ambient RNA that is released by ruptured or degraded cells into the suspension (Fleming et al. 2023; Young and Behjati 2020). Moreover, the formation of chimeric cDNA molecules during library amplification can lead to misassigned transcripts (Dixit 2016). This spillover of transcripts reduces the specificity of cell-type marker genes and can lead to the creation of artificial marker gene combinations, ultimately causing misannotation of cell types and masking distinctions between rare populations. Several computational tools (Fleming et al. 2023; Young and Behjati 2020; Yang et al. 2020) have been developed to address these issues by estimating and subtracting the contribution of ambient RNA and have proven effective in various scenarios such as for example removing counts from neuronal ambient RNA in non-neuronal cell populations (Caglayan et al. 2022; Zhang et al. 2023b).

Together, the impacts of doublets and background noise highlight the critical importance of rigorous quality control and artifact correction in scRNA-seq workflows. Characterizing how these artifacts influence analyses, as well as evaluating the performance of computational methods for their mitigation, remains a key challenge.

1.2.2 Species-mixing experiments

To characterize the extent and impact of these artifacts, it is essential to work with datasets where the distinction between true biological signal and technical noise is possible. One powerful approach to generating such ground-truth data involves species-mixing experiments, where cells from different species are pooled together before sequencing. These experiments make use of the genetic variation between species as natural barcodes, making it possible to assign individual transcripts to their species of origin.

The core idea behind these mixing experiments is straightforward: cells from two or more species—or alternatively other genetically distinct sources like subspecies or even individuals—are pooled into the same suspension and processed together through the scRNA-seq workflow (Figure 3). After sequencing, individual reads can be assigned to their genetic origin based on sequence identity. For species with sufficiently divergent genomes, like in the commonly used mixtures of human and mouse cell lines, this is typically done by aligning reads to both reference genomes and assigning them based on alignment quality. For more closely related species, subspecies or individuals of the same species, where a shared reference genome is used, single-nucleotide variants can be used to make this distinction (Kang et al. 2018; Huang et al. 2019; Xu et al. 2019; Heaton et al. 2020).

In a perfect, noise-free mixing experiment with two species, all reads assigned to a given cell barcode would originate from only one of the two species. In reality, however, transcripts from both species can be detected within the same barcode. In this case, a small fraction of foreign transcripts typically indicates background noise, such as ambient RNA spillover from lysed cells (Fleming et al. 2023). If transcripts from both species appear in similar proportions, it suggests a doublet, where two cells were captured together.

In this way, cell-mixing experiments enable the quantification of background noise and doublet rates, but they only detect cross-species events and therefore provide a lower-bound estimate. By considering the mixing proportions, these estimates can be extrapolated to approximate the overall rate of these artifacts, including within-species effects that cannot be directly identified (Bloom 2018).

Cell-mixing experiments have played a key role in evaluating the technical performance of single-cell technologies. For example, they are frequently used to demonstrate the doublet rate of newly developed methods (Macosko et al. 2015; Goldstein et al. 2017; Zheng et al. 2017; Rosenberg et al. 2018). Additionally, they have been employed in systematic comparisons of different protocols, helping to highlight differences in doublet rates and levels of ambient RNA contamination (Ding et al. 2020).

Beyond benchmarking, cell-mixing approaches are also used in experimental designs that combine different species or individuals to increase throughput and reduce batch effects. In such cases, the principles outlined above can be taken advantage of for quality control and

filtering of the count matrix. Computational tools designed for demultiplexing individuals based on genotype such as demuxlet (Kang et al. 2018), vireo (Huang et al. 2019) or scSplit (Xu et al. 2019) include functionality for identifying and removing doublets, while souporcell (Heaton et al. 2020) in addition provides an estimate of ambient RNA.

Finally, cell-mixing datasets serve as a ground truth for benchmarking computational methods. A widely used example is the mixture of the human HEK293T and mouse NIH3T3 cell lines provided by 10x Genomics. This dataset has been used to assess the accuracy of doublet-detection methods (Wolock et al. 2019; DePasquale et al. 2019; Xi and Li 2021). It has also played an important role in the evaluation of computational approaches for background noise removal (Yang et al. 2020; Young and Behjati 2020; Fleming et al. 2023). Similarly, Tian et al. (2019) used controlled mixtures of human lung adenocarcinoma cell lines to create structured datasets with known proportions, providing a benchmark for evaluating computational methods across various stages of scRNA-seq analysis.

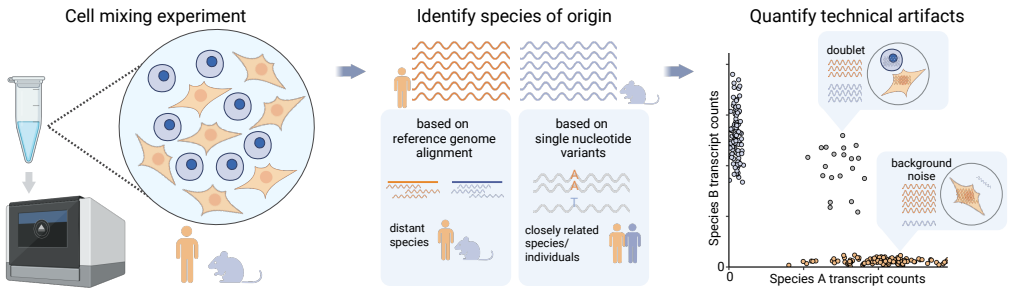


Figure 3. Species mixing experiments for scRNA-seq quality assessment. Cells from two or more species are pooled prior to performing scRNA-seq. During analysis, reads can be attributed back to their species of origin using genome alignment differences or single-nucleotide variants. Summarizing the contributions of different species per cell barcode helps to quantify technical artifacts: balanced contributions suggest doublets, while low-level signals from a second species indicate background RNA contamination. Created with BioRender.com

1.3 Comparative single-cell studies and cross-species analysis

Beyond the technology-focused application of species-mixing experiments, integrating scRNA-seq with cross-species analysis offers a powerful approach to gain biological insights, particularly in evolutionary research. In this section, I will explore how comparing cell types across species can deepen our understanding of cell type biology and evolution. Focusing on primate transcriptomics, I will discuss the experimental challenges that such studies face and highlight how in vitro cellular systems can help address them. Finally, I will outline the main analysis challenges associated with cross-species datasets and discuss strategies to overcome them.

1.3.1 Evolutionary cell type definition

While grouping transcriptomic profiles from a single species is sufficient for defining cell types (see 1.1), incorporating an evolutionary perspective adds depth to these classifications. It has been proposed that cell types can be defined as evolutionary units, maintained through conserved regulatory mechanisms (Arendt et al. 2016). Comparative single-cell transcriptomics is a powerful tool to identify these cell type-specific regulatory programs and gene expression profiles, providing insights into how cell types evolve (Arendt et al. 2019). This perspective also helps to distinguish between genuine cell types and transient cell states, which can be difficult to disentangle based on single-species data alone. While cell types are characterized by stable, hard-wired regulatory programs, cell states reflect reversible responses to environmental stimuli or physiological conditions (Tasic 2018). Cross-species comparisons make it possible to identify transcriptional programs that are consistently maintained—suggesting conserved cell types—as opposed to programs that vary flexibly within a type, indicating state-dependent changes (Arendt et al. 2019).

1.3.2 Cross-species comparisons in primates

Narrowing this perspective to comparisons within the primate clade offers a more targeted view on human biology. Primate studies have long been used to investigate evolutionary change at the DNA level, identifying instances of lineage-species sequence divergence and signatures of selection in both coding regions and regulatory elements (Rogers and Gibbs 2014; Chimpanzee Sequencing and Analysis Consortium 2005; Prabhakar et al. 2006; Lindblad-Toh et al. 2011). These efforts have provided important insights into the genetic basis of species differences, but they offer only indirect clues about how such changes translate into cellular or phenotypic effects. With transcriptomic data it is now possible to study gene expression directly, obtaining a functional readout that is closer to cellular phenotypes than genomic sequence alone (Khaitovich et al. 2006). Comparative single-cell analysis in particular has the potential to characterize evolutionary changes in gene expression and regulation at the cell type level.

In addition to revealing general patterns of gene regulation and expression, cross-species studies in primates can also highlight human-specific features. These include both novel cell types and regulatory changes within orthologous cell types—features that have only become accessible through single-cell transcriptomic approaches (Pollen et al. 2023; Juan et al. 2023). Beyond their evolutionary relevance, comparative primate studies also hold significant value for biomedical research. While the mouse remains the most widely used model organism, it often fails to replicate key aspects of human physiology. In this context, NHPs serve as an important intermediate model, helping to bridge the gap between rodent systems and humans (Enard 2012).

The brain's complexity and its link to human-specific traits make it a primary focus of comparative primate single-cell transcriptomic research so far and recent studies have provided examples of how this approach can uncover cellular and molecular differences. Several studies have identified human- or primate-specific shifts in cellular diversity. For instance, Krienen et al. (2020) compared single-nucleus RNA-seq from the brains of primates (human, macaque and marmoset), mice, and weasels and reported an expansion of inhibitory interneurons in primates. In addition, Ma et al. (2022) profiled the dorsolateral prefrontal

cortex in humans, chimpanzees, rhesus macaques, and marmosets, detecting some human-specific microglial and neuronal subtypes. Other studies have focused on species-specific gene expression and regulation differences within conserved cell types. Suresh et al. (2023) and Jorstad et al. (2023) analyzed single-cell transcriptomes from the middle temporal gyrus of humans, chimpanzees, gorillas, macaques, and marmosets, where the cell type composition is largely conserved across primates. They identified human-specific expression and co-expression changes in hundreds of genes (Suresh et al. 2023). Jorstad et al. (2023) further showed that gene expression evolved faster in human neurons compared to other primates, with human-specific changes enriched in genes related to synaptic function.

While these studies provide valuable insights into species differences in adult brain tissues, they capture only a limited view of evolutionary processes, missing the dynamic changes that occur during development. Comparisons of development across primates are of special interest, as this is when the foundation for many phenotype differences is laid. However, it is particularly difficult to obtain developmental tissues due to ethical and practical limitations (Pollen et al. 2023). Therefore, cellular systems that can model aspects of primate development in vitro are essential.

1.3.3 Generation and characterization of primate iPSCs

A widely used system for this purpose is induced pluripotent stem cells (iPSCs) (Wunderlich et al. 2014), which can be maintained in culture and differentiated in a wide range of cell types and organoid systems. In early approaches, pluripotent stem cells were obtained as embryonic stem cells from human and primate blastocysts (Thomson et al. 1995; Thomson et al. 1998). The possibility to generate iPSCs by reprogramming somatic cells revolutionized the field by providing an ethical and accessible alternative to embryonic stem cells. Soon after the initial experiments in mice (Takahashi and Yamanaka 2006), this technology was successfully adapted to generate the first human iPSC lines (Takahashi et al. 2007; Yu et al. 2007) as well as iPSCs from rhesus macaques, an NHP model organism (Liu et al. 2008). By 2023, researchers have derived over 100 iPSC lines from NHPs (Anwised et al. 2023).

An important consideration for the generation of primate iPSC lines is the availability

of primary cells. A majority of studies so far used skin fibroblasts or blood samples as source material (Juan et al. 2023). These cells are typically obtained post-mortem or during medical procedures and are a somatic cell type for which the reprogramming process is well established in humans (Raab et al. 2014). In a study by Geuder et al. (2021), urine from NHPs was explored as a non-invasive alternative. In this study, I contributed to the transcriptomic characterization of the primary urine cells and reprogrammed iPSCs.

Once a new stem cell line has been established, it is crucial to validate it. For human cell lines, some standardized protocols and recommendations have been defined to streamline the process (Ludwig et al. 2023). This includes the use of functional assays such as teratoma formation or in vitro differentiation into all three germ layers (ectoderm, mesoderm and endoderm) to confirm pluripotency. Additionally, the undifferentiated state can be validated by the presence of specific marker genes. Moreover, genome stability is a critical aspect of validation, with karyotyping routinely performed to detect chromosomal abnormalities. Importantly, authentication is necessary to ensure the identity of the cell line can be reliably verified in the future. For human cell lines, this is commonly achieved using short tandem repeat (STR) analysis.

While proper validation is equally important for NHP iPSC lines, the protocols are less standardized and approaches developed for human cell lines are not always directly applicable (Yang et al. 2018). For instance, established markers for humans may exhibit different cell type specificities in NHPs and antibodies are often susceptible to cross-reactivity (Bjornson-Hooper et al. 2022). Furthermore, STR panels have been developed or adapted from human panels for certain primate species, including rhesus macaques (Kanthaswamy et al. 2006), African green monkeys (Almeida et al. 2011) and chimpanzees (Singh et al. 2019). However, due to their species-specific nature, STR panels are unlikely to generalize across non-human primates, and comprehensive validation would be required for each species individually.

Bulk and single-cell RNA-seq offers a versatile readout for cell line characterization and authentication in NHPs. First of all, comparing whole transcriptomes to reference cell types allows for characterization independent of individual markers and antibody specificity. In this context, I contributed to the characterization of orangutan and gorilla iPSC lines, as

well as primary cells, by classifying bulk RNA-seq profiles with a reference dataset of human cell types (Geuder et al. 2021). Furthermore, sequencing data allows for the identification of single nucleotide polymorphisms (SNPs), providing an alternative to STR profiling for cell line authentication. To this end, I compiled a list of informative variants from bulk RNA-seq data for two baboon iPSC lines (Jocher et al. 2024b) and two vervet monkey iPSC lines (Jocher et al. 2024c), as well as from scRNA-seq data of three rhesus macaque iPSC lines (Jocher et al. 2024a). Additionally, scRNA-seq could confirm pluripotency for the rhesus macaque iPSC lines by demonstrating the presence of cell populations of all three germ layers in an embryoid body formation experiment (Jocher et al. 2024a).

1.3.4 iPSC-derived organoid systems

The main benefit of having iPSCs from multiple primate species and individuals is their potential for use in comparative differentiation protocols and organoid models. These systems not only overcome some limitations in tissue availability, but also enable conducting time-course measurements and provide a controlled environment for genetic modifications and other experimental manipulations (Pollen et al. 2023). Their integration with high-throughput approaches like scRNA-seq is especially powerful for detailed comparative analyses of gene expression and cellular dynamics across species. In a directed differentiation approach towards a specific cell type, single-cell resolution makes it possible to distinguish intermediate differentiation stages and pinpoint species-specific divergence points. For instance, Housman et al. (2022) used iPSC-derived mesenchymal stem cells from humans and chimpanzees to study osteogenic differentiation at the single-cell level. They found that while most gene expression patterns were conserved, hundreds of genes were differentially expressed between species and the biggest differences were observed in mineralizing osteoblasts, a transitional cell type.

In recent years, the emergence of organoid models has made it possible to recapitulate some aspects of tissue organization and development with 3D in vitro models (Clevers 2016). In this case, single-cell transcriptomics helps to dissect the cellular composition and regulatory programs. Comparative studies building on this synergy of organoid and

single-cell technologies improve the analysis of cross-species developmental variation. For instance, several studies have used scRNA-seq data of cerebral organoids to study differences in cortical development across humans, chimpanzees and other primates (Kanton et al. 2019; Pollen et al. 2019; Mora-Bermúdez et al. 2016; Fischer et al. 2022). These studies indicated differences in differentiation speed (Kanton et al. 2019), regulation of signaling pathways in radial glia cells (Pollen et al. 2019) and the function of human-specific genes (Fischer et al. 2022), all of which may contribute to unique features of human cortical development.

A simple organoid system that is well-suited to study early cell differentiation processes are embryoid bodies (EBs). They are three-dimensional structures that form by spontaneous and asynchronous differentiation of pluripotent stem cells (Han et al. 2018; Rhodes et al. 2022). As such, EBs contain a wide range of different cell types from all three germ layers, as well as from different stages of differentiation. This high variability is both the greatest strength and limitation of EBs. On the one hand, it allows to study gene regulation in a broad spectrum of differentiation processes and cell types. On the other hand, the unpredictable nature of EBs and variability across replicates and individuals (Rhodes et al. 2022) complicates comparative analyses. Nevertheless, several studies have used single-cell technologies to dissect the cell type composition of EBs and study early developmental trajectories in humans (Han et al. 2018; Moon et al. 2019; Rhodes et al. 2022), mouse (Spangler et al. 2018; Kim et al. 2020), as well as in a comparative setting of human and chimpanzee (Barr et al. 2023).

1.3.5 Computational challenges of cross-species analysis

Using comparable experimental systems is an important step toward meaningful cross-species comparisons, but comparability also needs to be ensured at every step of the data analysis. Cross-species analysis presents some unique computational challenges. Cross-species single-cell studies come with specific computational challenges that affect nearly every step of the workflow, from generation of the count matrix to cell type annotation. In this section, I will discuss the main obstacles and the strategies used to address them.

Comparable feature space

In scRNA-seq, genes serve as the fundamental features for profiling and representing cells. In cross-species studies, matching genes between species is therefore essential, not only to directly compare gene expression but also to obtain a shared feature space for cell-level analyses (Tanay and Sebé-Pedrós 2021). A common approach is to restrict comparisons to one-to-one orthologs to ensure direct gene correspondence across species. However, this can substantially reduce the feature space, especially when multiple species are included (Tarashansky et al. 2021). Alternative strategies aim to incorporate many-to-many gene relationships to preserve complex gene correspondences. For example, SAMap (Tarashansky et al. 2021) refines gene mappings based on both sequence and expression similarity and SATURN (Rosen et al. 2024) identifies functionally related genes using large protein language models. Besides orthology challenges, genome annotation quality further complicates gene comparability. While the human genome is well-curated, non-human primate annotations are often incomplete or inaccurate (Housman and Gilad 2020). These discrepancies can bias gene mappings and affect downstream analyses.

Batch effects

Technical batch effects are a common challenge in scRNA-seq and can complicate cross-species comparisons by introducing unwanted variation. These effects can arise at various experimental stages, potentially obscuring true biological differences or creating misleading artifacts (Hicks et al. 2018). Distinguishing real cross-species variation from batch effects is therefore crucial. Careful experimental design can help minimize these issues or allow them to be accounted for during data processing. One way to reduce batch effects is sample multiplexing, where cells from different samples are pooled before sequencing (Zhang et al. 2022). When working with samples from different individuals or species, naturally occurring genetic variants can be used to computationally assign cells back to their sample of origin (Kang et al. 2018; Xu et al. 2019; Huang et al. 2019; Heaton et al. 2020).

Homologous cell type assignment

As with most scRNA-seq analyses, organizing single cells into distinct cell types is essential for cross-species comparisons. However, achieving a consistent and reliable cell type assignment across species is not always straightforward. Various strategies have been proposed to address this challenge.

One option is to integrate data from different species and assign cell types on a combined representation (Figure 4A). In this approach, the species differences are treated as a batch effect and data integration tools are applied to correct for the species effect and combine the data in a shared embedding (Shafer 2019). Cell types are then assigned based on clustering within this shared space. Data integration is a central step in scRNA-seq analysis, and numerous tools have been developed and evaluated for this purpose. However, a benchmarking study of integration methods (Luecken et al. 2020) found that cross-species integration, specifically between human and mouse immune cells, was among the most challenging scenarios for all tested methods. The species effect is particularly strong compared to other batch effects and the harsh integration required across species often leads to a loss of biologically meaningful variation (Luecken et al. 2020). The challenge becomes even greater when comparing multiple species or when large phylogenetic distances make it difficult to establish accurate gene orthologies (Tanay and Sebé-Pedrós 2021). More specialized methods, such as SAMap (Tarashansky et al. 2021) and SATURN (Rosen et al. 2024), aim to improve cross-species integration by considering sequence similarity and functional relatedness of genes. SAMap refines gene relationships iteratively based on expression similarity, allowing for the detection of functional paralogs, while SATURN makes use of protein language models to align functionally similar genes across species. Indeed, benchmarks of cross-species integration have shown the effectiveness of these methods especially when combining distantly related species (Song et al. 2023; Zhong et al. 2025). However, integration can blur species-specific expression signatures, and overcorrection may lead to loss of cell type distinguishability, limiting the reliability of annotation in some cases Song et al. 2023.

An alternative approach is to use a well-annotated species as a reference and transfer cell type labels to the other species (Figure 4B). This effectively turns it into a classification task,

where the reference species serves as the training set and cell type identities are predicted in the query species. This method is particularly effective when one species has comprehensive annotations and the others are expected to have similar cellular compositions. Several single-cell classification tools, that were originally developed for within-species cell type prediction, have also been applied and tested for cross-species classification between human and mouse (Pliner et al. 2019; Tan and Cahan 2019; Abdelaal et al. 2019). Alternatively, specialized cross-species methods have been developed to improve annotation transfer, even across greater phylogenetic distances (Liu et al. 2023; Zhang et al. 2024; Park et al. 2024). However, this approach has certain limitations: it relies on a highly comprehensive reference dataset and since it assumes that cell types in the query species correspond to those in the reference, it does not allow for the identification of novel or species-specific cell types.

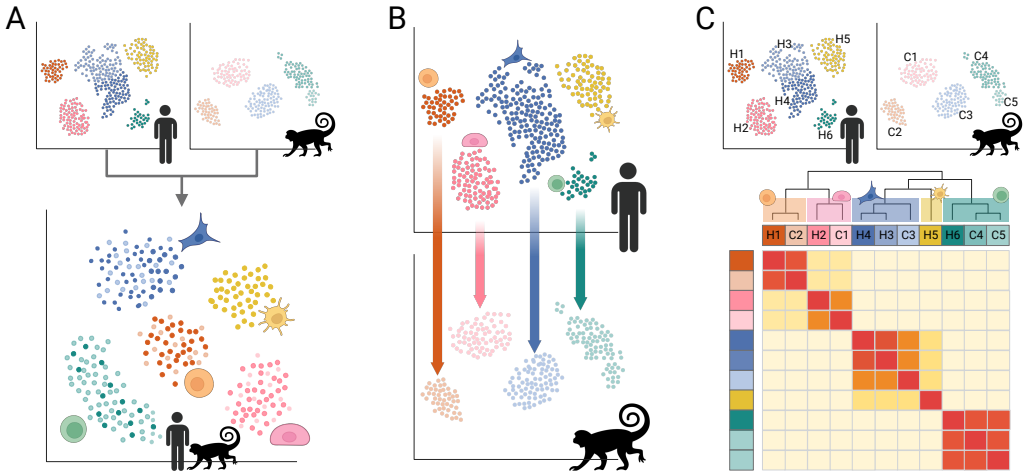


Figure 4. Main approaches for cross-species cell type assignment from scRNA-seq data. A) Integration across species and annotation on a shared embedding. B) Classification or label transfer from one annotated reference species to the other. C) Independent grouping of cells within each species, followed by identification of correspondences across species. Created with BioRender.com

Finally, a third approach involves grouping cells within each species and then matching these groups across species ((Figure 4C). The grouping can be based on preliminary cell type annotations or unsupervised clustering. Once clusters are established, they are compared using different similarity metrics. One popular tool, MetaNeighbor (Crow et al. 2018), applies a neighbor-voting algorithm to assess how well the transcriptional profile of a cluster is

retained across species. Other strategies include scoring the similarity of cluster-specific marker gene lists and fold changes (Gao et al. 2019; González-Velasco et al. 2024) or employing classification-based metrics that evaluate how well a cluster in one species can be classified using another species as a reference (Biharie et al. 2023). The final step consists of linking clusters across species by selecting the most similar pairs or groups of clusters based on predefined similarity metrics. The main advantage of this approach is that it can preserve species-specific differences and avoids the risk of overintegration. Nevertheless, it depends on accurate initial clustering or annotation, making it sensitive to errors in grouping. In addition, selecting an appropriate similarity metric and matching criteria is not straightforward.

The choice of the general approach and particular methods ultimately depends on features of the dataset like the number of species, phylogenetic distances and the complexity and overlap of the cell type compositions. There is no "one-for-all" solution to the problem of cell type matching across species and careful evaluation of the final assignments remains crucial.

1.4 Aims of the thesis

The aim of this thesis is to improve the methodological basis for cross-species scRNA-seq, with a focus on primates. To this end, the work addresses key technical and conceptual challenges that affect how reliably scRNA-seq data can be interpreted in a comparative context. The specific aims are as follows:

- To assess and quantify the impact of ambient RNA contamination in single-cell and single-nucleus RNA-seq data and to evaluate correction strategies using species-mixing experiments as a benchmarking framework.
- To support comparative primate studies by validating and characterizing iPSC lines from multiple non-human primates and to provide a reference dataset for early primate differentiation dynamics using an embryoid body model.
- To explore how reliably marker genes can be used across species and to better understand the challenges of assigning cell types in a cross-species setting.

Together, these studies aim to improve the quality and reliability of cross-species single-cell analyses and to support more accurate biological interpretation.

2 | Results

2.1 The effect of background noise and its removal on the analysis of single-cell expression data

Philipp Janssen, Zane Kliesmete, Beate Vieth, Xian Adiconis, Sean Simmons, Jamie Marshall, Cristin McCabe, Holger Heyn, Joshua Z. Levin, Wolfgang Enard and Ines Hellmann "The effect of background noise and its removal on the analysis of single-cell expression data" *Genome Biology* 24.1 (2023): 140. doi: 10.1186/s13059-023-02978-x

Janssen *et al. Genome Biology* (2023) 24:140
https://doi.org/10.1186/s13059-023-02978-x

Genome Biology

RESEARCH Open Access

The effect of background noise and its removal on the analysis of single-cell expression data

Philipp Janssen[1], Zane Kliesmete[1], Beate Vieth[1], Xian Adiconis[2,3], Sean Simmons[2,3], Jamie Marshall[4], Cristin McCabe[2], Holger Heyn[5], Joshua Z. Levin[2,3], Wolfgang Enard[1] and Ines Hellmann[1*]

*Correspondence: hellmann@bio.lmu.de

[1] Anthropology and Human Genomics, Faculty of Biology, Ludwig-Maximilians University, Munich, Germany
[2] Klarman Cell Observatory, Broad Institute of Harvard and MIT, Cambridge, USA
[3] Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, USA
[4] Broad Institute of Harvard and MIT, Cambridge, USA
[5] CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain

Abstract

Background: In droplet-based single-cell and single-nucleus RNA-seq experiments, not all reads associated with one cell barcode originate from the encapsulated cell. Such background noise is attributed to spillage from cell-free ambient RNA or barcode swapping events.

Results: Here, we characterize this background noise exemplified by three scRNA-seq and two snRNA-seq replicates of mouse kidneys. For each experiment, cells from two mouse subspecies are pooled, allowing to identify cross-genotype contaminating molecules and thus profile background noise. Background noise is highly variable across replicates and cells, making up on average 3–35% of the total counts (UMIs) per cell and we find that noise levels are directly proportional to the specificity and detectability of marker genes. In search of the source of background noise, we find multiple lines of evidence that the majority of background molecules originates from ambient RNA. Finally, we use our genotype-based estimates to evaluate the performance of three methods (CellBender, DecontX, SoupX) that are designed to quantify and remove background noise. We find that CellBender provides the most precise estimates of background noise levels and also yields the highest improvement for marker gene detection. By contrast, clustering and classification of cells are fairly robust towards background noise and only small improvements can be achieved by background removal that may come at the cost of distortions in fine structure.

Conclusions: Our findings help to better understand the extent, sources and impact of background noise in single-cell experiments and provide guidance on how to deal with it.

Keywords: Single-cell RNA-sequencing, Background noise, Ambient RNA, Barcode swapping, Correction method comparison, (Gold) standard scRNA-seq data set

Background

Single cell and single nucleus RNA-seq (scRNA-seq, snRNA-seq) are in the process of revolutionizing medical and biological research. The typically sparse coverage per cell and gene is compensated by the capability of analyzing thousands of cells in one experiment. In droplet-based protocols such as 10x Chromium, this is achieved by encapsulating single cells in droplets together with beads that carry oligonucleotides. These usually consist of a oligo(dT) sequence which is used for priming reverse transcription, a bead-specific barcode that tags all transcripts encapsulated within the droplet as well as unique molecular identifiers (UMIs) that enable the removal of amplification noise [1–3]. As proof of principle that each droplet encapsulates only one cell, it is common to use mixtures of cells from human and mouse [3]. Thus doublets, i.e., droplets containing two cells, can be readily identified as they have an approximately even mixture of mouse and human transcripts. However, barcodes for which the clear majority of reads is either mouse or human, still contain a small fraction of reads from the other species [3–5]. Furthermore, presumably empty droplets also yield sequence reads [4].

One potential source of such contaminating reads or background noise is cell-free "ambient" RNA that leaked from broken cells into the suspension. The other potential source are chimeric cDNA molecules that can arise during library preparation due to so-called 'barcode swapping'. The pooling of barcode tagged cDNA after reverse transcription but before PCR amplification, is a decisive step to achieve high throughput. However, if amplification of tagged cDNA molecules occurs from unremoved oligonucleotides from other beads or from incompletely extended PCR products (originally called template jumping [6]), this generates a chimeric molecule with a "swapped" barcode and UMI [7, 8]. When sequencing this molecule, the cDNA is assigned to the wrong barcode and hence "contaminates" the expression profile of a cell. However, unless the swapping occurs between two different genes, the barcode and UMI will still be counted correctly. Another type of barcode swapping can occur during PCR amplification on a patterned Illumina flowcell before sequencing [9] with the same effects, although double indexing of Illumina libraries has reduced this problem substantially. This said, here we focus on barcode swapping that occurs during library preparation.

Irrespective of the source of background noise, its presence can interfere with analyses. For starters, background noise reduces the separability of cell type clusters as well as the power to pinpoint important (marker) genes via differential expression analysis. Moreover, reads from cell type-specific marker genes spill over to cells of other types, thus yielding novel marker combinations and hence implying the presence of novel cell types [8, 10]. Besides, background noise can also confound differential expression analysis between samples, e.g., when looking for expression changes within a cell type between two conditions. Varying amounts of background noise or differences in the cell type composition between conditions can result in dissimilar background profiles, which might generate false positives when identifying differentially expressed genes. To alleviate such problems during downstream analysis, algorithms to estimate and correct for the amounts of background noise have been developed.

SoupX estimates the contamination fraction per cell using marker genes and then deconvolutes the expression profiles using empty droplets as an estimate of the background noise profile [11]. In contrast, DecontX defaults to model the fraction of

background noise in a cell by fitting a mixture distribution based on the clusters of good cells [8], but also allows the user to provide a custom background profile, e.g., from empty droplets. CellBender requires the expression profiles measured in empty droplets to estimate the mean and variance of the background noise profile originating from ambient RNA. In addition, CellBender explicitly models the barcode swapping contribution using mixture profiles of the 'good' cells [4].

In order to evaluate method performance, one dataset of an even mix between one mouse and one human cell line [3] is commonly used to get an experimentally determined lower bound of background noise levels that is identified as counts covering genes from the other species [4, 8, 11, 12]. Since this dataset is lacking in cell type diversity, it is common to additionally evaluate performance based on other datasets that have a complex cell type mixture and where most cell types have well known profiles with exclusive marker genes. In such studies the performance test is whether the model removes the expression of the exclusive marker genes from the other cell types. In both cases, the feature space of the contamination does not overlap with the endogenous cell feature space. Mouse and human are too diverged, so that mouse reads only map to mouse genes and human reads only to human genes. Similarly, when using marker genes it is assumed that they are exclusively expressed in only one cell type, hence the features that are used for background inference are again not overlapping. However, in reality background noise will mostly induce shifts in expression levels that cannot be described in a binary on or off sense and it remains unclear how background correction will affect those profiles.

Here, we use a mouse kidney dataset representing a complex cell type mixture from three mouse strains of two subspecies, *Mus musculus domesticus* and *M. m. castaneus*. From both subspecies, inbred strains were used and thus we can distinguish exogenous and endogenous counts for the same features using known homozygous SNPs [13]. Hence, this dataset serves as a much more realistic experimental standard, providing a ground truth in a complex setting with multiple cell types which allows to analyze the variability, the source and the impact of background noise on single cell analysis. Moreover, this dataset enables us to better benchmark existing background removal methods.

Results

Mouse kidney single cell and single nucleus RNA-seq data

We obtained three replicates for single cell RNA-seq (rep1-3) data and two replicates for single nucleus RNA-seq (snRNA-seq, nuc2 and nuc3) data from the same samples that were used in scRNA-seq replicates 2 and 3, respectively. Each replicate consists of one channel of 10× [3] in which cells from dissociated kidneys of three mice each were pooled: one *M. m. castaneus* from the strain CAST/EiJ (CAST) and two *M. m. domesticus*, one from the strain C57BL/6J (BL6) and one from the strain 129S1/SvImJ (SvImJ) (Fig. 1A). Based on known homozygous SNPs that distinguish subspecies and strains, we assigned cells to mice (Fig. 1B). In total, we identified $> 40,000$ informative SNPs of which the majority (32,000) separates the subspecies and $\sim 10,000$ SNPs distinguish the two *M. m. domesticus* strains (Fig. 1C). On average, each cell had sufficient coverage for $\sim 1,000$ informative SNPs ($\sim$ 20% of total UMIs per cell) to provide us with unambiguous genotype calls for those sites. The coverage for the nuc2 data was much lower with only $\sim$ 100 SNPs (Fig. 1D).

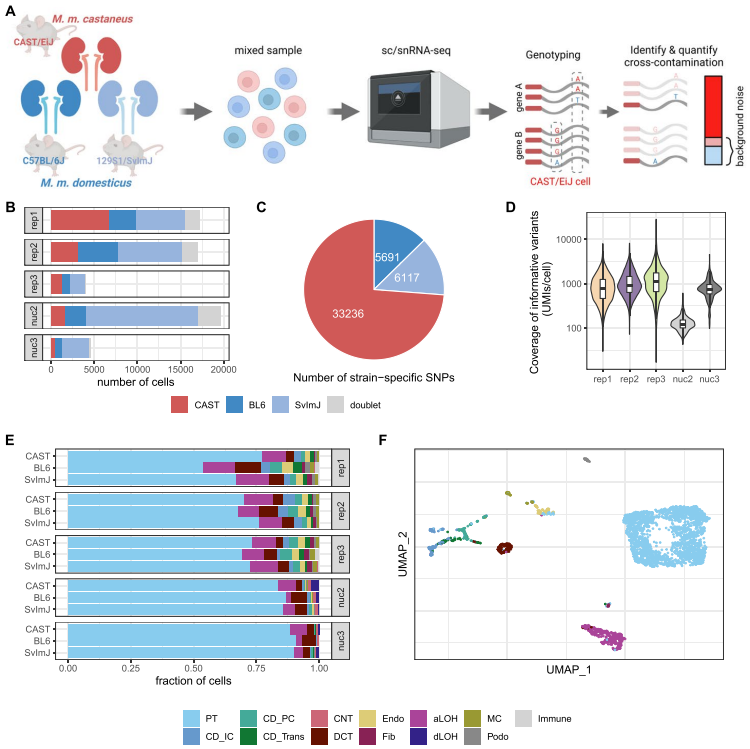


Fig. 1 Generation of mouse strain mixture datasets to quantify background noise. **A** Experimental design (created with BioRender.com). **B** Strain composition in 5 different replicates, subjected to scRNA-seq (rep1-3) or snRNA-seq (nuc2, nuc3). The replicates rep2 and nuc2 and rep3 and nuc3 were generated from the same samples each. CAST: CAST/EiJ strain; BL6: C57BL/6J strain; SvImJ: 129S1/SvImJ. **C** Number of homozygous SNPs with a coverage of more than 100 UMIs that distinguish one strain from the other two. **D** Per cell coverage in *M. m. castaneus* cells of informative variants that distinguish *M. m. castaneus* and *M. m. domesticus*. **E** Cell type composition per replicate and strain; labels were obtained by reference-based classification using mouse kidney data from Denisenko et al. [14] as reference. **F** UMAP visualization of *M. m. castaneus* cells in single-cell replicate 2, colored by assigned cell type. PT, proximal tubule; CD_IC, intercalated cells of collecting duct; CD_PC, principal cells of collecting duct; CD_Trans, transitional cells of collecting duct; CNT, connecting tubule; DCT, distal convoluted tubule; Endo, endothelial; Fib, fibroblasts; aLOH, ascending loop of Henle; dLOH, descending loop of Henle; MC, mesangial cells; Podo, podocytes

Overall, each experiment yielded 5000–20,000 good cells with 9–43% *M. m. castaneus* (Fig. 1B). Thus, the majority of background noise in any *M. m. castaneus* cell is expected to be from *M. m. domesticus* (Additional file 1: Fig. S1B) and therefore we expect that genotype-based estimates of cell-wise amounts of background noise for *M. m. castaneus* to be fairly accurate (Additional file 1: Fig. S2). Hence from here on out we focus on *M. m. castaneus* cells for the analysis of the origins of background noise and also as the ground truth for benchmarking background removal methods.

This dataset has two advantages over the commonly used mouse-human mix [3]. Firstly, the kidney data have a high cell type diversity. Using the data from Denisenko et al. [14] as reference dataset for kidney cell types, we could identify 13 cell types.

Encouragingly, the cell type composition is very similar across mouse strains as well as replicates with proximal tubule cells constituting 66–89% of the cells (Fig. 1E, F; Additional file 1: Fig. S3). Secondly, due to the higher similarity of the mouse subspecies, we can identify contaminating reads for the same features. $\sim 7,000$ genes carry at least one informative SNP about the subspecies. Because so many genes have informative SNPs, the fraction of UMIs that cover an informative SNP is a little higher for PTs, the most frequent cell type, but very comparable across all other cell types, allowing us to quantify contaminating reads (Additional file 1: Fig. S1A).

Background noise fractions differ between replicates and cells

Around 5–20% of the UMI counts are from molecules that contain a SNP that is informative about the subspecies of origin. We quantify in each *M. m. castaneus* cell how often an endogenous *M. m. castaneus* allele or a foreign *M. m. domesticus* allele was covered. Assuming that the count fractions covering the SNPs are representative of the whole cell, we detect a median of 2–27% counts from the foreign genotype over all cells per experiment (Additional file 1: Fig. S1C). This observed cross-genotype contamination fraction represents a lower bound of the overall amounts of background noise. As suggested in Heaton et al. [15], we then integrate over the foreign allele fractions of all informative SNPs to obtain a maximum likelihood estimate of the background noise fraction (ρ_{cell}) of each cell that extrapolates to also include contamination from the same genotype (see the "Methods" section, Additional file 1: Fig. S2). Based on these estimates, we find that background noise levels vary considerably between replicates and do not appear to depend on the overall success of the experiment measured as the cell yield per lane (Fig. 2). For example in scRNA-seq rep3 (3900 cells), we detected overall the fewest good cells, but most of those cells had less than 3% background noise, while the much more successful rep2 (15,000 cells) we estimated the median background noise level at around 11% (Fig. 2A). This said, the snRNA-seq data generated from frozen tissue have much higher background levels than the corresponding scRNA-seq replicates — 35% in nuc2 vs. 11% rep2 and 17% in nuc3 vs. 3% in rep3. How we define good cells based on the UMI counts has little impact on this variability. We still find by far the highest background levels in nuc2 and the lowest in rep3 (Additional file 1: Fig. S4). This high variability is not very surprising. This being a real life experiment and experimental conditions were improved for nuc3 based on the experience with nuc2 (see the "Methods" section). The

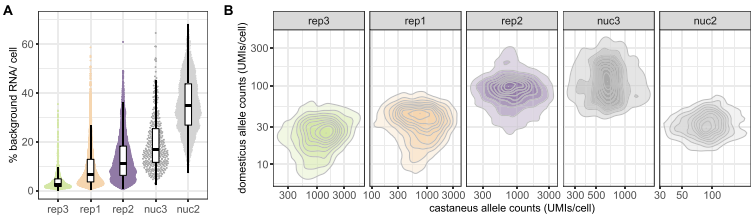


Fig. 2 The level of background noise is variable across replicates and single cells. **A** Estimated fraction of background noise per cell. The replicates on the *x*-axis are ordered by ascending median background noise fraction. **B** In *M. m. castaneus* cells both endogenous *M. m. castaneus* specific alleles (*x*-axis) and *M. m. domesticus* specific alleles (*y*-axis) have coverage in each cell. The detection of *M. m. domesticus* specific alleles can be seen as background noise originating from cells of a different mouse

number of contaminating RNA-molecules (UMIs) depends only weakly on the total UMI counts covering informative variants as a proxy for sequencing depth of the cell (Fig. 2B, Additional file 1: Table S1). Such a weak correlation could be explained by variation in the capture efficiency in each droplet. An alternative, but not mutually exclusive explanation of such a correlation could be that the source of some contaminating molecules is barcode swapping that can occur during library amplification.

However, by and large the absolute amount of background noise is approximately constant across cells and thus the contamination fraction mainly depends on the amount of endogenous RNA: the larger the cell, the smaller the fraction of background noise, pointing towards ambient RNA as the major source of the detected background (Fig. 2B).

Contamination profiles show a high similarity to ambient RNA profiles

In order to better understand the effects of background noise, it is helpful to understand its origins and composition. To this end, we constructed profiles representing endogenous, contaminating and ambient expression profiles by using *M. m. domesticus* allele counts in *M. m. domesticus* cells (endogenous), *M. m. domesticus* allele counts in *M. m. castaneus* cells (contamination) and *M. m. domesticus* allele counts in empty droplets (empty) (Fig. 3A , B; Additional file 1: Fig. S5A-E).

The number of contaminating UMI counts per cell is at a similar level as the UMI counts in empty droplets in all replicates (Fig. 3C, Additional file 1: Fig. S5F). Moreover, if the median UMI count in empty droplets is high for one replicate, we also observe

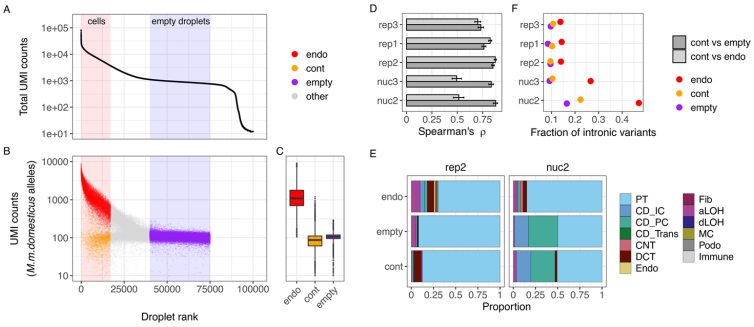


Fig. 3 Characterization of ambient RNA in cells and empty droplets. **A** Ordering droplet barcodes by their total UMI count to distinguish cell-containing droplets with high UMI counts from empty droplets that only contain cell-free ambient RNA and are identifiable as a plateau in the UMI curve, shown here for replicate 2. **B** UMI counts of reads covering *M. m. domesticus* specific alleles were used to construct three profiles depending on whether they were associated with *M. m. domesticus* cell barcodes (endogenous counts, endo), *M. m. castaneus* cell barcodes (contaminating counts, cont) or empty droplet barcodes (empty). Counts from droplets that are not clearly assignable as cell-containing or empty were excluded from further analysis (other). **C** UMI counts per cell for each of the three profiles. **D** Spearman rank correlation between pseudobulk profiles. Error bars indicate 95% confidence intervals obtained by bootstrapping over genes. **E** Deconvolution of cell type contributions to each pseudobulk profile, exemplified by replicates rep2 and nuc2. The stacked barplots depict the estimated fraction of each cell type in the profile as inferred by SCDC using the annotated single cell data of each replicate as reference. PT, proximal tubule; CD_IC, intercalated cells of collecting duct; CD_PC, principal cells of collecting duct; CD_Trans, transitional cells of collecting duct; CNT, connecting tubule; DCT, distal convoluted tubule; Endo, endothelial; Fib, fibroblasts; aLOH, ascending loop of Henle; dLOH, descending loop of Henle; MC, mesangial cells; Podo, podocytes. **F** Fraction of reads covering intronic variants in each of the three profiles

more contaminating UMIs, which is also consistent with ambient RNA as the main source for background noise.

In addition, when comparing pseudobulk aggregates of the three scRNA-seq replicates, we find that the contamination profiles correlate highly and similarly well with empty (Spearman's $\rho = 0.73 - 0.85$) and endogenous profiles (Spearman's $\rho = 0.70 - 0.87$), while for the nuc2 and nuc3 the contamination profiles are clearly more similar to the empty (Spearman's $\rho \sim 0.85$) than to the endogenous profiles (Spearman's $\rho \sim 0.50$) (Fig. 3B).

Using deconvolution analysis[16], we reconstructed the cell type composition from the pseudobulk profiles. In agreement with the correlation analysis, we find that in our scRNA-seq data the cell type compositions inferred for endogenous, contamination and empty counts are by and large similar with a slight increase in the PT-profile in empty droplets, suggesting that this cell type is more vulnerable to dissociation procedure than other cell types. In contrast, deconvolution of the empty droplet and contamination fraction of nuc2 and nuc3, that in contrast to the scRNA-seq data were prepared from frozen samples, shows a clear shift in cell type composition with a decreased PT fraction (Fig. 3C, Additional file 1: Fig. S6).

Moreover, we expect that cytosolic mRNA contributes more to the contaminating profile than to the endogenous profile. Indeed, in our snRNA-seq data we find that in good nuclei (endogenous molecules) more than 25% of the allele counts fall within introns, while out of the molecules from empty droplets less than 18% fall within introns (Fig. 3D). Similarly also in the scRNA-seq data, we find with $\sim$ 14% more intron variants than in empty droplets. The intron fraction of the contaminating molecules lies in-between the endogenous and the empty droplet fraction, but is in all cases much closer to the empty intron fraction, thus suggesting again that the majority of the background noise likely originates from ambient RNA.

Only little evidence for barcode swapping

In addition to ambient RNA, barcode swapping resulting from chimera formation during PCR amplification can also contribute to background noise. With the 12bp UMIs from 10x, the probability that we capture the same UMI-cell barcode combination twice independently is very low, hence how often we find the same combination of cell barcode and UMI associated with more than one gene is a good measure for barcode swapping [7]. The median fraction of such chimeric molecules varies between 0.2% for rep3 and 0.7% for nuc3 (Additional file 1: Fig. S7A). In line with our expectations outlined before, the absolute amount of swapping per cell correlates strongly with the total molecule count (Additional file 1: Table S1). In combination with the weak correlation between the number of contaminating with endogenous molecule counts, this supports the notion that the majority of background noise does not come from swapping. To be more quantitative, we combine the swapping and the total background fractions to estimate how much swapping could contribute to the total background and find that the median contribution of barcode swapping to background noise is lower than 10% for all replicates (Additional file 1: Fig. S7B).

Furthermore, molecules with a swapped barcode are expected to have a lower average number of reads per UMI. This is because chimera that are formed late during PCR

subsequently undergo less amplification [7]. Thus, if the majority of contaminating reads were to originate from barcode swapping, we would expect that the distribution of reads per UMI for cross-genotype contaminating molecules (cont) is similar to that of observed chimeras. This is not what we see (Additional file 1: Fig. S7C): The distribution of reads per UMI for contaminating reads is much more distinct from the distribution for chimeras (Kolmogorov-Smirnov distance, $\Delta_n = 0.381$ (rep3) to 0.595 (nuc3)) than for endogenous reads ($\Delta_n = 0.008$ (rep2) to 0.046 (rep3)). In summary, we find that barcode swapping during library preparation only contributes little to the overall background noise in this data.

The impact of contamination on marker gene analyses

The ability to distinguish hitherto unknown cell types and states is one of the greatest achievements made possible by single cell transcriptome analyses. To this end, marker genes are commonly used to annotate cell clusters for which available classifications appear insufficient. An ideal marker gene would be expressed in all cells of one type but in none of the other present cell types. Thus, when comparing expression levels of one cell type versus all others, we expect high log2-fold changes, the higher the change the more reliable the marker. However, such a reliance on marker genes also makes this type of analysis vulnerable to background noise. Our whole kidney data can illustrate this problem well, because with the very frequent proximal tubular (PT) cells we have a dominant cell type for which rather specific marker genes are known [17]. Slc34a1 encodes a phosphate transporter that is known to be expressed exclusively in PT cells [18, 19]. As expected, it is expressed highly in PT cells, but it is also present in a high fraction of other cells (Fig. 4A, E; Additional file 1: Fig. S8). Moreover, the log2-fold changes of Slc34a1 are smaller in replicates with larger background noise, indicating that

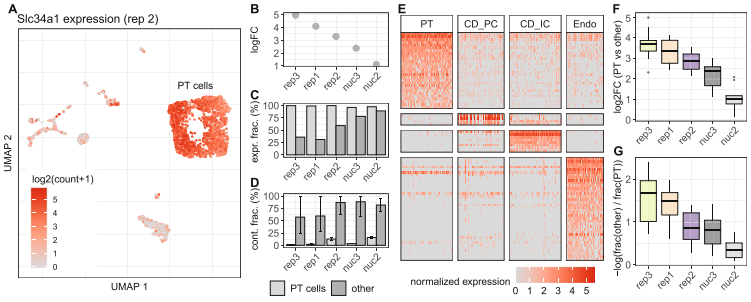


Fig. 4 Background noise affects differential expression and specificity of cell type specific marker genes. **A** UMAP representation of replicate 2 colored by the expression of Slc34a1, a marker gene for cells of the proximal tubule (PT). Besides high counts in a cluster of PT cells, Slc34a1 is also detected in other cell type clusters. Differential expression analysis between PT and all other cells shows a decrease of the detected log fold change of Slc34a1 (**B**) at higher background noise levels, as well as an increase of the fraction of non PT cells in which UMI counts of Slc34a1 were detected (**C**). **D** Estimation of the background noise fraction of Slc34a1 expression indicates that the majority of counts in non PT cells originates from background noise. Error bars indicate 90% profile likelihood confidence intervals. **E** Heatmap of marker gene expression for four cell types in replicate 2, downsampled to a maximum of 100 cells per cell type. **F** Comparison across replicates of log2 fold changes of 10 PT marker genes calculated based on the mean expression in PT cells against mean expression in all other cells. **G** For the same set of genes as in **E**, the log ratio of fraction of cells in which a gene was detected in others and PT cells shows how specific the gene is for PT cells

the detection of Slc34a1 in non-PT cells is likely due to contamination (Fig. 4B–D). We observe the same pattern for other marker genes as well: they are detected across all cell types (Fig. 4E, Additional file 1: Fig. S9) and an increase of background noise levels goes along with decreasing log2-fold changes and increasing detection rates in other cell types (Fig. 4F,G). Thus, the power to accurately detect marker genes decreases in the presence of background noise.

Benchmark of background noise estimation tools

Given that background noise will be present to varying degrees in almost all scRNA-seq and snRNA-seq replicates, the question is whether background removal methods can alleviate the problem without the information from genetic variants. SoupX [11], DecontX [16] and CellBender [4], all provide an estimate of the background noise level per cell. Here, we use our genotype-based background estimates as ground truth to compare it to the estimates of the three background removal methods (Fig. 5A, Additional file 1: Fig. S10). All methods have adjustable parameters, but also provide a set of defaults. For CellBender the user can adjust the nominal false positive rate to put a cap on losing information from true counts. For SoupX and DecontX the resolution of the clustering of cells that is later used to model the endogenous counts can be adjusted. In addition, SoupX can be provided with an expected background level and for DecontX the user can provide a custom background profile rather than using the

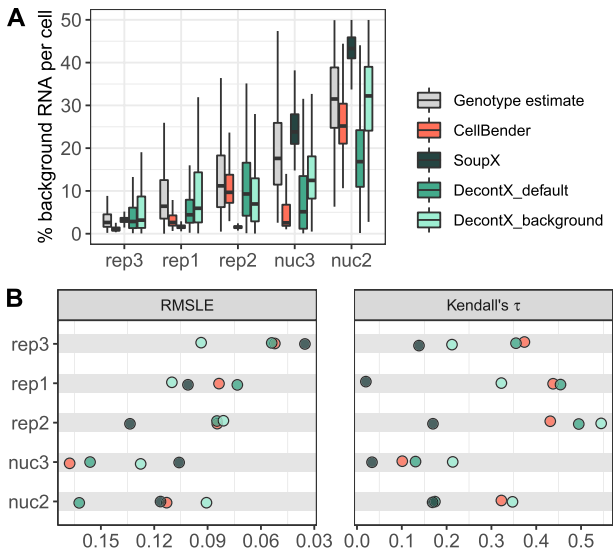


Fig. 5 Accuracy of computational background noise estimation. **A** Estimated background noise levels per cell based on genetic variants (gray) and different computational tools. **B** Taking the genotype-based estimates as ground truth, Root Mean Squared Logarithmic Error (RMSLE) and Kendall rank correlation serve as evaluation metrics for cell-wise background noise estimates of different methods. Low RMSLE values indicate high similarity between estimated values and the assumed ground truth. High values of Kendall's τ correspond to good representation of cell to cell variability in the estimated values

default estimation strategy for the background profile. At least with our reference dataset, CellBender does not seem to profit from changing the defaults, while SoupX's performance is boosted, if provided with realistic background levels (Additional file 1: Fig. S15). Because in a real case scenario, the true background level is unknown, we decided to report the SoupX performance metrics under default settings. DecontX defaults to estimating the putative background profile from averaging across intact cells. To ensure comparability, we report DecontX's performance with empty droplets as background profile (DecontX$_{background}$) in addition to DecontX with default settings (DecontX$_{default}$).

We find that CellBender and DecontX can estimate background noise levels similarly well for the scRNA-seq replicates, while SoupX tends to underestimate background levels and also cannot capture the cell to cell variation as measured by the correlation with the ground truth (Fig. 5B). For nuc2 and nuc3 , SoupX performs better at estimating global background levels, but as for the scRNA-seq still cannot capture cell to cell variation. In contrast, both CellBender and DecontX perform worse for nuc2 and nuc3. Moreover for nuc2 and nuc3, DecontX with default setting provides worse estimates than with empty droplets as background profile.

All in all, CellBender shows the most robust performance across replicates with default settings, while DecontX' and SoupX' performance seems to require parameter tuning. A drawback of CellBender is its runtime. While SoupX and DecontX take seconds and minutes to process one 10× channel, CellBender takes ∼ 45 CPU hours. However, parallelization is possible.

All methods struggled most with the nuc3 replicate that has the fewest *M. m. castaneus* cells and the lowest cell type diversity among our five data sets (Fig. 1B, E). This also presents a problem for other downstream analyses and thus we do not consider nuc3 further.

Effect of background noise removal on marker gene detection

Above we have shown that computational methods can estimate background noise levels per cell. Moreover, all three methods provide the user with a background corrected count matrix for downstream analysis. Here, we compare the outcomes of marker gene detection, clustering and classification when using corrected count matrices from SoupX, DecontX, and CellBender (Fig. 6A, Additional file 1: Fig. S11). To characterize the impact on marker gene detection, we first check in how many cells an unexpected marker gene was detected; for example, how often Slc34a1 was detected in cells other than PTs (Fig. 6B). Without correction we find Slc34a1 reads in ∼ 60% of non-PT cells of rep2, SoupX reduces this rate to 54%, CellBender to 7% and DecontX$_{background}$ to 9%. DecontX$_{default}$ manages to remove most contaminating reads reducing the Slc34a1 detection rate outside PTs to 2%. While we find a similar ranking when averaging across several marker genes from the PanglaoDB database [17] and scRNA-seq replicates (Fig. 6C), the ranking changes for nuc2: DecontX$_{default}$ fails: after correction, Slc34a1 is still found in 87% of non-PT cells while DecontX$_{background}$ is better with a rate of 20%. Here, CellBender and SoupX are clearly better with reducing the Slc34a1 detection rate to 4% and $< 1\%$, respectively (Additional file 1: Fig. S12).

Even though the changes in the marker gene detection rates outside the designated cell type seem dramatic (Additional file 1: Fig. S13A), the identification of marker genes

Janssen *et al. Genome Biology* (2023) 24:140 Page 11 of 22

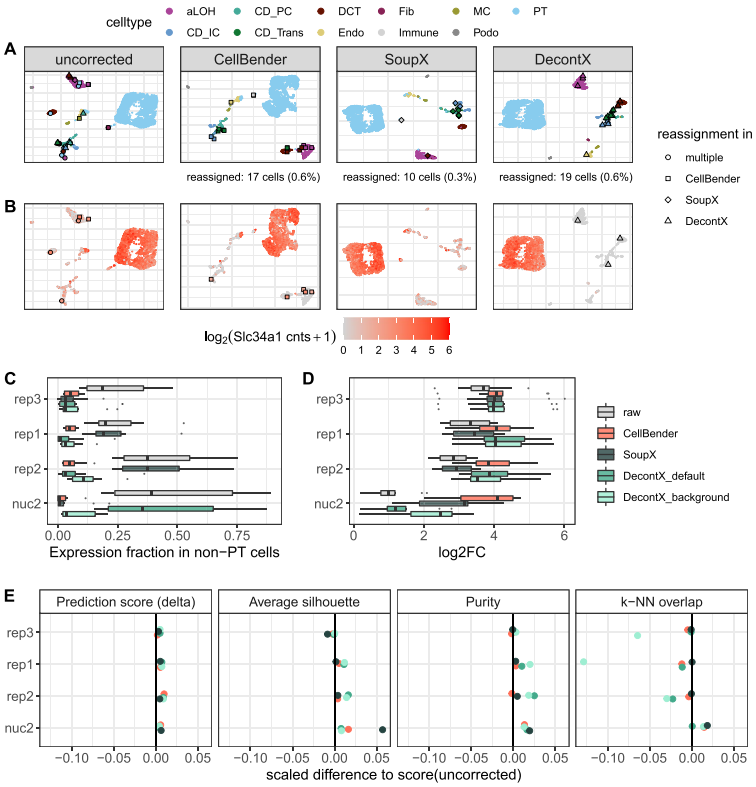


Fig. 6 Effect of background removal on downstream analysis. **A** UMAP representation of replicate 2 single-cell data before and after background noise correction, colored by cell type labels obtained from reference based classification. Individual cells that received a new label after correction are highlighted. PT, proximal tubule; CD_IC, intercalated cells of collecting duct; CD_PC, principal cells of collecting duct; CD_Trans, transitional cells of collecting duct; CNT, connecting tubule; DCT, distal convoluted tubule; Endo, endothelial; Fib, fibroblasts; aLOH, ascending loop of Henle; dLOH, descending loop of Henle; MC, mesangial cells; Podo, podocytes. **B** Expression of the PT cell marker Slc34a1 before and after background noise correction in replicate 2. Cells that were classified as PT cells in the uncorrected data, but got reassigned after correction, are highlighted. **C**, **D** Differential expression analysis of 10 PT markers, evaluating the expression fraction in non-PT cells (**C**) and the log2 fold change between PT and all other cells (**D**). **E** Evaluation metrics for the effect of background noise correction on classification and clustering. For each metric the change relative to the uncorrected data is depicted. The values were scaled by the possible range of each metric. Prediction score: cell-wise score "delta" of reference based classification with SingleR [20]. Average silhouette: Mean of silhouette widths per cell type. Purity: Cluster purity calculated on cell type labels as ground truth and Louvain clusters as test labels. *k*-NN overlap: overlap of the k=50 nearest neighbors per cell compared to genotype-cleaned reference *k*-NN graph

[21] is affected only a little. CellBender correction has the largest effect on marker gene detection, yet 8 from the top 10 genes without correction remain marker genes with CellBender correction (Spearman's correlation for top 100 $\rho = 0.84$). In contrast, in the nuc2 data with high background levels, the change in marker gene detection is dramatic. Here, only one of the top 10 marker genes remains after correction (Spearman's

correlation for top 100 $\rho = 0.04$). The largest improvement is achieved with CellBender: After correction, four out of the top 10 were known marker genes [17], while this overlap amounted to only one in the raw data (Additional file 1: Fig. S13B). Moreover, we find that background removal also increases the detected log-fold-changes of known marker genes across all replicates and methods, with CellBender providing the largest improvement (Fig. 6D, Additional file 1: Fig. S13C).

Effect of background noise removal on classification and clustering

One of the first and most important tasks in single cell analysis is the classification of cell types. As described above, we could identify 13 cell types in our uncorrected data using an external single cell reference dataset [14, 20]. Going through the same classification procedure after correction for background noise, changes the classification of only very few cells (Fig. 6A, Additional file 1: Fig. S11). For the scRNA-seq experiments $< 1\%$ and for the nuc2 up to 1.3% of cells change labels after background removal compared to the classification using raw data. Before correction, these cells are mostly located in clusters dominated by a different cell type (Fig. 6A). Moreover, these cells tend to have higher background levels as exemplified by the PT-marker gene Slc34a1 (Fig. 6B). Finally, background removal — irrespective of the method - improves the classification prediction scores (Fig. 6E, Additional file 1: Fig. S14). Together, this indicates that background removal improves cell type classification.

Similarly, background removal also results in more distinct clusters. Here, we reason that cells of the same cell type should cluster together and evaluate the impact of background removal (1) on the silhouette scores for cell types and (2) on the cell type purity of each cluster using unsupervised clustering (Fig. 6E). For the scRNA-seq data DecontX results in the purest and most distinct clusters, while for the nuc2 data SoupX wins in these categories.

All in all, it seems clear that all background removal methods sharpen the broad structure of the data a little, but how about fine structure? To answer this question, we turn again to the genotype cleaned data to obtain a ground truth for the k-nearest neighbors of a cell and calculate how much higher the overlap of the background corrected data is with this ground truth as compared to using the raw data (Fig. 6E). For the scRNA-seq data, DecontX has the largest improvement on the broad structure, but at same time in particular DecontX$_{background}$ lowers the overlap in k-NN with our assumed ground truth, suggesting that this change in structure is a distortion rather than an improvement. SoupX leaves the fine structure by and large unchanged in the scRNA-seq data, while both CellBender and DecontX make the fine structure slightly worse. In contrast, for the high background levels of the nuc2, all background removal methods achieve an improvement, with SoupX and CellBender performing best.

Discussion

Here we provide a dataset for the characterization of background noise in 10× Genomics data that is ideal to benchmark background removal methods. The mixture of cell types in our kidney data provides us with realistic cell type diversity and the mixture of mouse subspecies enables us to identify foreign alleles in a cell, thus resulting in a dataset that allows us to quantify background noise across diverse cell types and features. In

addition, the replicates exhibit varying degrees of contamination, enabling us to evaluate the effects of low, intermediate, and high background levels. Given that every sample poses new challenges for the preparation of a suspension of intact cells or nuclei that is needed for a 10× experiment, we expect that such variability in sample quality is not unusual. Consequently, marker gene identification is affected and markers appear less specific, as they are detected in cell types where they are not expressed. The degree of this issue directly depends on background noise levels (Fig. 4). This particular problem has been observed previously and has been used as a premise to develop background correction methods [4, 11, 22].

The novelty of this analysis is that — thanks to the mix of mouse subspecies — we are able to obtain expression profiles that describe the source of contamination in each sample and also have a ground truth for a more realistic dataset. We started to characterize background noise by comparing the contamination profile with the profile of empty droplets and that of endogenous counts of good cells. In agreement with the idea that ambient RNA is due to leakage of cytosol, we find that empty droplets show less evidence for unspliced mRNA molecules and that the unspliced fraction in the contamination profiles is similar to that of empty droplets. This is a first hint that a large proportion of the background noise is ambient RNA. In addition, we find only little direct evidence for barcode swapping as provided by chimeric UMIs, which only explains up to 10% of background noise (Additional file 1: Fig. S7B). Hence, also the observed correlation between cell size and the absolute amounts of background noise per cell in most of the replicates is likely due to variation in dropout rates [4] (Fig. 2B, Additional file 1: Table S1).

Another important insight from comparing contamination, empty and endogenous profiles is that we can deduce the origin of the contamination. While for rep1-3 all three profiles are highly correlated and are the result of very similar cell type mixtures, for nuc2 and nuc3 the empty and the contamination profiles are distinct from the expected endogenous mixture profile. Encouragingly the endogenous profiles of all replicates agree well with one another as well as with the cell type proportions from the literature [14, 23]. Moreover, the higher similarity of the contamination to the empty than to the endogenous profile supports the notion that the majority of background noise is ambient RNA and hence using the empty rather than the endogenous profile as a reference to model background noise is the better choice for our data. Indeed, the performance of DecontX for nuc2 is improved by providing the empty droplet profile as compared to the endogenous profile which is the default (Fig. 5A). We also observed that SoupX performs much better for the snRNA-seq data than the scRNA-seq data. We speculate that the marker gene identification that is the basis for estimating the experiment-wide average contamination is hampered by the fact that our dataset has one very dominant cell type that has the same prevalence in the empty droplets, thus masking all background. However, even if SoupX gets the overall background levels right, it by design grossly underestimates the variance among cells and cannot capture the cell to cell variation (Fig. 5B, C). Overall CellBender provides the most accurate estimates of the background noise levels and also captures the cell to cell variation rather well. We note that this finding is largely due to the robustness of CellBender to cell type composition and

the source of contamination, that determines the similarity between the contamination and the endogenous profiles.

In line with this, also marker gene detection is most improved by CellBender, which is the only method that removes marker gene molecules from other cell types and increases the log-fold-change consistently well. The effect of background removal on other downstream analyses is much more subtle. For starters, classification using an external reference is rather robust. Even with high levels of background noise, background removal improves classification only for a handful of cells and we cannot say that one method outperforms the others (Fig. 6E, Additional file 1: Fig. S14). Similarly, the broad structure of the data improves only minimally and this minimal improvement comes at the cost of disrupting fine structure (Fig. 6E). Here, again CellBender strikes the best balance between removing variation but preserving the fine structure, while DecontX tends to remove too much within-cluster variability, as the *k*-NN overlap with the genotype-based ground truth for DecontX is even lower than for the raw data. All in all, CellBender shows the best performance in removing background noise.

Conclusions

Levels of background noise can be highly variable within and between replicates and the contamination profiles do not always reflect the cell type proportions of the sample. Marker gene detection is affected most by this issue, in that known cell type specific marker genes can be detected in cell clusters where they do not belong. Existing methods for background removal are good at removing such stray marker gene molecule counts. In contrast, classification and clustering of cells is rather robust even at high levels of background noise. Consequently, background removal improves the classification of only few cells. Moreover, it seems that for low and moderate background levels the tightening of existing broad structures may go at the cost of fine structure. In summary, for marker gene analysis, we would always recommend background removal, but for classification, clustering and pseudotime analyses, we would only recommend background removal when background noise levels are high.

Methods

Mice

Three mouse strains were ordered from Jackson Laboratory at 6–8 weeks of age: C57BL/6J (000664), CAST/EiJ (000928), and 129S1/SvlmJ (002448). All animals were subjected to intracardiac perfusion of PBS to remove blood. Kidneys were dissected, divided into 1/4s, and subjected to the tissue dissociation protocol, stored in RNAlater, or snap-frozen in liquid nitrogen.

Tissue dissociation for single cell isolation

The single cell suspensions were prepared following an established protocol [24] with minor modifications. In detail, one of each kidney sagittal quarter from three perfused mice of different strains C57BL/6, CAST/EiJ and 129S1/SvImJ were harvested into cold RPMI (Thermo Fisher Scientific, 11875093) with 2% heat-inactivated Fetal Bovine Serum (Gibco, Thermo Fisher Scientific, 16140-071; FBS) and 1% penicillin/streptomycin (Gibco, Thermo Fisher Scientific, 15140122). Each piece of the tissue was then

minced for 2 min with a razor blade in 0.5 ml 1x liberase TH dissociation medium (10x concentrated solution from Millipore Sigma, 05401135001, reconstituted in DMEM/F12(Gibco, Thermo Fisher Scientific, 11320-033 in a petri dish on ice. The chopped tissue pieces were then pooled into one 1.5 ml Eppendorf tube and incubated in a thermomixer at 37°C for 1 hour at 600rpm with gentle pipetting for trituration every 10 min. The digestion mix was then transferred to a 15 ml conical tube and mixed with 10 ml 10% FBS RPMI. After centrifugation in a swinging bucket rotor at 500g for 5 min at 4°C and supernatant removal, the pellet was resuspended in 1ml red blood cell lysing buffer (Sigma Aldrich, R7757). The suspension was spun down at 500g for 5 min at 4°C followed by supernatant removal. The pellet cleared of the red blood cell ring was then resuspended in 250 μl Accumax (Stemcell Technologies, 7921) and incubated at 37°C for 3 mins. The reaction was stopped by mixing with 5 ml 10% FBS RPMI and spinning down at 500g for 5 min at 4°C followed by supernatant removal. The cell pellet was then resuspended in PBS with 0.4% BSA (Sigma, B8667) and passed through a 30 μm filter (Sysmex, 04-004-2326). The cell suspension was then assessed for viability and concentration using the K2 Cellometer (Nexcelom Bioscience) with the AOPIcell stain (Nexcelom Bioscience, CS2-0106-5ML).

Nuclei isolation from RNAlater preserved frozen tissue

The single nuclei suspensions were prepared following an established protocol [25] with minor modifications. In detail, the RNAlater reserved frozen tissue of 3 mice kidney quarters were thawed and transferred to one petri dish preloaded with 1 ml TST buffer containing 10 mM Tris, 146 mM NaCl, 1 mM CaCl2, 21 mM MgCl2, 0.03% Tween-20 (Roche, 11332465001), and 0.01% BSA (Sigma, B8667). It was minced with a razor blade for 10 min on ice. The homogenized tissue was then passed through a 40 μm cell strainer (VWR, 21008-949) into a 50 ml conical tube. One ml TST buffer was used to rinse the petri dish and collect the remaining tissue into the same tube. It was then mixed with 3 ml of ST buffer containing 10 mM Tris, 146 mM NaCl, 1 mM CaCl2, and 21 mM MgCl2 and spun down at 500g for 5 min at 4°C followed by supernatant removal. In the second experiment this washing step was repeated 2 more times. The pellet was resuspended in 100 μl ST buffer and passed through a 35 μm filter. The nuclei concentration was measured using the K2 Cellometer (Nexcelom Bioscience) with the AO nuclei stain (Nexcelom Bioscience, CS1-0108-5ML).

Single-cell and single-nucleus RNA-seq

The cells or nuclei were loaded onto a 10× Chromium Next GEM G chip (10x Genomics, 1000120) aiming for recovery of 10,000 cells or nuclei. The RNA-seq libraries were prepared using the Chromium Next GEM Single Cell 3' Reagent kit v3.1 (10× Genomics, 1000121) following vendor protocols. The libraries were pooled and sequenced on NovaSeq S1 100c flow cells (Illumina) with 28 bases for read1, 55 bases for read2 and 8 bases for index1 and aiming for 20,000 reads per cell.

Processing and annotation of scRNA-seq and snRNA-seq data

The scRNA-seq and snRNA-seq data were processed using Cell Ranger 3.0.2 using as reference genome and annotation mm10 version 2020A for the scRNA-seq data and and

a pre-mRNA version of mm10 2.1.0 as reference for snRNA-seq. In order to identify cell containing droplets we processed the raw UMI matrices with the DropletUtils package [5]. The function barcodeRanks was used to identify the inflection point on the total UMI curve and the union of barcodes with a total UMI count above the inflection point and Cell Ranger cell call were defined as cells.

For cell type assignment we used 3 scRNA-seq and 4 snRNA-seq experiments from Denisenko et al. [14] as a reference. Cells labeled as "Unknown" (n=46), "Neut" (n=17) and "Tub" (n=1) were removed. The reference was log-normalized and split into seven count matrices based on chemistry, preservation and dissociation protocol. Subsequently, a multi-reference classifier was trained using the function *trainSingleR* with default parameters of the R package SingleR version 1.8.1 [20]. After this processing, we could use the data to classify our log-normalized data using the *classifySingleR* function without fine-tuning (fine.tune = F). Hereby, each cell is compared to all seven references and the label from the highest-scoring reference is assigned. Some cell type labels were merged into broader categories after classification: cells annotated as "CD_IC," "CD_IC_A," or "CD_IC_B" were relabeled as "CD_IC," cells annotated as "T," "NK," "B," or "MPH" were relabeled as "Immune." Cells that were unassigned after pruning of assignments based on classification scores were removed for subsequent analyses.

Demultiplexing of mouse strains

A list of genetic variants between mouse strains was downloaded in VCF format from the Mouse Genomes Project [13], accessed on 21 October 2020. This reference VCF file was filtered for samples CAST_EiJ, C57BL_6NJ and 129S1_SvImJ and chromosomes 1–19. Genotyping of single barcodes was performed with cellsnp-lite [26], filtering for positions in the reference VCF with a coverage of at least 20 UMIs and a minor allele frequency of at least 0.1 in the data (–minCOUNT 20, –minMAF 0.1). Vireo [22] was used to demultiplex and label cells based on their genotypes. Only cells that could be unambiguously assigned to CAST_EiJ (CAST), C57BL_6NJ (BL6) or 129S1_SvImJ (SvImJ) were kept, cells labeled as doublet or unassigned were removed.

Genotype-based estimation of background noise

Based on the coverage filtered VCF-file (see above), we identified homozygous SNPs that distinguish the three strains and removed SNPs that had predominantly coverage in only one of the strains (1st percentile of allele frequency).

In most parts of the analysis, we focused on the comparison between the mouse subspecies, *M. m. domesticus* and *M. m. castaneus*. To this end, we subseted reads (UMI-counts) that overlap with SNPs that distinguish the two mouse subspecies.

To estimate background noise levels based on allele counts of genetic variants, an approach described in Heaton et al.[15] was adapted to estimate the total amount of background noise for each cells. First, the abundance of endogenous and foreign allele counts (i.e., cross-genotype background noise) was quantified per cell. Because of the filter for homozygous variants, there are two possible genotypes for each locus, denoted as 0 for the endogenous allele, i.e., the expected allele based on the strain assignment of

the cell, and 1 for the foreign allele. The probability for observable background noise at each locus l in cell c is given by

$$p = \rho_c * \frac{A_{l,1}}{A_{l,0} + A_{l,1}} \tag{1}$$

where ρ_c is the total background noise fraction in a cell and the experiment wide (over cells and empty droplets) foreign allele fraction is calculated from the foreign allele counts $A_{l,1}$ and the endogenous allele counts $A_{l,0}$. The foreign allele fraction is then used to account for intra-genotype background noise (contamination within endogenous allele counts).

The observed allele counts A_c per cell are modeled as draws from a binomial distribution with the likelihood function:

$$P(A_c|\rho_c) = \prod_{l \in L} \binom{A_{l,c,0} + A_{l,c,1}}{A_{l,c,1}} p^{A_{l,1}} (1-p)^{A_{l,0}} \tag{2}$$

A maximum likelihood estimate of ρ_c was obtained using one dimensional optimization in the interval [0,1].

The 95% confidence interval of each ρ_c estimate was calculated as the profile likelihood using the function *uniroot* of the R package stats [27].

Comparison of endogenous, contamination, and empty droplet profiles

Empty droplets were defined based on the UMI curve of the barcodes ranked by UMI counts, thus selecting barcodes from a plateau with $\sim 500 - 1000$ UMIs (Additional file 1: Fig. S5). For the following analysis, the presence of *M. m. domesticus* alleles in *M. m. domesticus* cells (i.e., endogenous), in *M. m. castaneus* cells (i.e., contamination) and empty droplets was compared. After this filtering, we summarized counts per gene and across barcodes of the same category to generate pseudobulk profiles.

In order to estimate cell type composition in the empty and contamination profiles, we used the deconvolution method implemented in SCDC[16], the endogenous single cell allele counts from the respective replicate were used as reference ($qcthreshold = 0.6$). In addition, cell type filtering (frequency>0.75%) was applied. Endogenous, contamination and empty pseudobulk profiles from each replicate were deconvoluted using their respective single cell/single nucleus reference.

To compare the correlation between the different profiles, pseudobulk counts were downsampled to the same total size.

Detection of barcode swapping events

Information about the number of reads per molecule and the combination of cell barcode (CB), UMI and gene were extracted from the molecule info file in the Cellranger output. We assume that a combination of CB and UMI corresponds to a single original molecule. Thus we define a PCR chimera as a non-unique CB-UMI combination in which multiple genes were associated with the same CB and UMI. Since we can only detect PCR chimera, if we detect at least 2 reads for a CB-UMI combination, we also

restrict the total molecule count to CB-UMI combinations with at least 2 reads for the calculation of the chimera fraction.

For the comparison of reads/UMI the identified chimera were intersected with identified cross-genotype contamination. To this end, the the analysis was restricted to *M. m. castaneus* cells and CB-UMI-gene combinations which can be associated with an informative SNP. The number of reads/UMI was summarized per CB-UMI-gene combination for chimera (as defined above), unique CB-UMI-gene combinations with coverage for an endogenous allele (endo) and unique CB-UMI-gene combinations with coverage for a foreign allele (cont).

Evaluation of marker gene expression

A list of marker genes for Proximal tubule cells (PT), Principal cells (CD_PC), Intercalated cells (CD_IC), and Endothelial cells (Endo) was downloaded from the public database PanglaoDB [17], accessed on 13 May 2022.

Log2 fold changes contrasting PT cells against all other cells were calculated with Seurat using the function *FindMarkers* after normalization with *NormalizeData*. The expression fraction e of PT markers was calculated as the fraction of cells for which at least 1 count of that gene was detected. To contrast expression fraction in PT cells against non-PT, the negative log-ratio was calculated as $-log((e_{PT}+1)/(e_{non-PT}+1))$.

Computational background noise estimation and correction methods

CellBender [4] makes use of a deep generative model to include various potential sources of background noise. Cell states are encoded in a lower-dimensional space and an integer matrix of noise counts is inferred, which is subsequently subtracted from the input count matrix to generate a corrected matrix.

The *remove-background* module of CellBender v0.2.0 was run on the raw feature barcode matrix as input, with a default *fpr* value of 0.01. For the comparison of different parameter settings, *fpr* values of 0.05 and 0.1 were also included in the analysis. For the parameter *expected-cells* the number of cells after cell calling and filtering in each replicate was provided. The parameter *total-droplets-included* was set to 25,000.

SoupX [11] estimates the experiment-wide amount of background noise based on the expression of strong marker genes that are expected to be expressed exclusively in one cell type. These genes can either be provided by the user or identified from the data. A profile of background noise is inferred from empty droplets. This profile is subsequently removed from each cell after aggregation into clusters to generate a corrected count matrix.

Cluster labels for SoupX were generated by Louvain clustering on 30 principal components and a resolution of 1 as implemented by *FindClusters* in Seurat after normalization and feature selection of 5000 genes. Providing the CellRanger output and cluster labels as input, data were imported into SoupX version 1.6.1 and the background noise profile was inferred with *load10X*. The contamination fraction was estimated using *autoEstCont* and background noise was removed using *adjustCounts* with default parameters.

For the comparison of parameter settings, different resolution values (0.5, 1, 2) for Louvain clustering were tested, alongside with manually specifying the contamination fraction (0.1, 0.2).

DecontX [8] is a Bayesian method that estimates and removes background noise by modeling the expression in each cell as a mixture of multinomial distributions, one native distribution cell's population and one contamination distribution from all other cell populations. The main inputs are a filtered count matrix only containing barcodes that were called as cells and a vector of cluster labels. The contamination distribution is inferred as a weighted combination of multiple cell populations. Alternatively, it is also possible to obtain an empirical estimation of the contamination distribution from empty droplets in cases where the background noise is expected to differ from the profile of filtered cells.

The function *decontX* from the R package celda version 1.12.0 was run on the filtered, unnormalized count matrix and clusters were inferred with the implemented default method based on UMAP dimensionality reduction and dbscan [28] clustering. For the "DecontX_default" results the parameter "background" was set to NULL, i.e., estimating background noise based on cell populations in the filtered data only. "DecontX_background" results were obtained by providing an unfiltered count matrix including all detected barcodes as "background" to empirically estimate the contamination distribution. Besides the default clustering method implemented in DecontX, cluster labels obtained from Louvain clustering (resolution 0.5, 1, and 2) were also provided to test different parameter settings.

Evaluation metrics

Estimation accuracy

The genotype-based estimates ρ_c for *M. m. castaneus* cells served as ground truth to evaluate the estimation accuracy of different methods. For each method cell-wise background noise fractions a_c were calculated from the corrected count matrix X and the uncorrected ("raw") count matrix R as

$$a_c = 1 - \frac{\sum_g x_{c,g}}{\sum_g r_{c,g}} \tag{3}$$

for cells c and genes g.

RMSLE The Root Mean Squared Logarithmic Error (RMSLE) is a lower bound metric that we use to quantify the difference between estimated background noise fractions per cell a_c from different computational background correction methods and the genotype-based estimates ρ_c, obtained from genotype based estimation. It is calculated as:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{c=1}^{n}(log(a_c + 1) - log(\rho_c + 1))^2} \tag{4}$$

Kendall's

τ To evaluate how well cell-to-cell variation of the background noise fraction is captured by the estimated values a_c, the Kendall rank correlation coefficient τ to the genotype-based estimates ρ_c was computed using the implementation in the R package stats [27] as $\tau = cor(a_c, \rho_c, method = "kendall")$.

Marker gene detection

The same set of 10 PT marker genes from PanglaoDB as in the "Evaluation of marker gene expression" section was used to evaluate the improvement on marker gene detection on corrected count matrices.

Log2 fold change for each gene between the average expression in PT cells and average expression in other cells were obtained using the *NormalizeData* and *FindMarkers* functions in Seurat version 4.1.1.

Expression fraction Entries in each corrected count matrix were first rounded to the nearest integer. The expression fraction of each gene in a cell population was calculated as the fraction of cells for which at least 1 count of that gene was detected. For evaluation of PT marker genes, unspecific detection is defined as the expression fraction in non-PT cells.

Cell type identification

Prediction score Each corrected count matrix was log-normalized and reference-based classification in SingleR [20] was performed with a pre-trained model (see "Processing and annotation of scRNA-seq and snRNA-seq data" section) on data from Denisenko et al. [14]. SingleR provides *delta* values as a measure for classification confidence, which depicts the difference of the assignment score for the assigned label and the median score across all labels. The *delta* values for each cell were retrieved using the function *getDeltaFromMedian* relative to the cells highest-scoring reference. A prediction score per cell type was calculated by averaging *delta* values across individual cells and a global prediction score per replicate was calculated by averaging across cell type prediction scores.

Average silhouette The silhouette width is an internal cluster evaluation metric to contrast similarity within a cluster with similarity to the nearest cluster. The cell type annotations from reference-based classification were used as cluster labels here. Count matrices were filtered to select for *M. m. castaneus* cells and cell types with more than 10 cells. Distance matrices were computed on the first 30 principal components using euclidean distance as distance measure. Using the cell type labels and distance matrix as input, the average silhouette width per cell type was computed with the R package cluster version 2.1.4. An *Average silhouette* per replicate was calculated as the mean of cell type silhouette widths.

Purity Purity is an external cluster evaluation metric to evaluate how well a clustering recovers known classes. Here, *Purity* was used to assess to what extent unsupervised cluster labels correspond to cell types. Count matrices were filtered to select for *M. m. castaneus* cells and cell types with more than 10 cells and Louvain clustering as implemented in *FindClusters* of Seurat version 4.1.1 on the first 30 principal components and with a resolution parameter of 1 was used to get a cluster label for each cell. Providing cell type annotations as true labels alongside the cluster labels, *Purity* was computed with the R package ClusterR version 1.2.6 [29].

k-NN overlap To evaluate the lower-dimensional structure in the data beyond clusters and cell-types *k*-NN overlap was used as described in Ahlmann-Eltze and Huber [30]. A ground truth reference *k*-NN graph was constructed on a 'genotype-cleaned' count matrix, only counting molecules that carry a subspecies-endogenous allele. Raw

and corrected count matrices were filtered to contain the same genes as in the reference and a query *k*-NN graph was computed on the first 30 principal components. The *k*-NN overlap summarizes the overlap of the 50 nearest neighbors of each cell in the query with the reference *k*-NN graph.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02978-x.

Additional file 1: Supplementary Material. This file contains Table S1 and Figures S1-15. **Table S1.** Spearman correlation analysis of background noise and barcode swapping. **Fig. S1.** Detection of cross-genotype contamination. **Fig. S2.** Estimation of background noise levels. **Fig. S3.** UMAP visualization showing the composition per replicate. **Fig. S4.** Definition of true cells and its effect on background noise estimates. **Fig. S5.** Definition of endogenous, empty droplet and contamination profiles across replicates. **Fig. S6.** Dissection of cell type contributions by deconvolution of pseudobulk profiles. **Fig. S7.** Identification of barcode swapping due to PCR chimeras. **Fig. S8.** Slc34a1 expression across replicates. **Fig. S9.** Expression of cell type marker genes. **Fig. S10.** Estimated background noise levels across cell types. **Fig. S11.** UMAP representations of all replicates before and after background noise correction. **Fig. S12.** Detected expression levels of Slc34a1 before and after background noise correction. **Fig. S13.** Effect of background noise correction on marker gene detection. **Fig. S14.** Evaluation metrics for cell type identification. **Fig. S15.** Evaluation of different parameter settings.

Additional file 2. Review history.

Acknowledgements
We thank Gabriela Stumberger for her help in benchmarking and Batuhan Akçabozan for his contribution to calculating genotype estimates. We thank the Broad Genomics Platform for sequencing.

Review history
The review history is available as Additional file 2.

Peer review information
Veronique van den Berghe was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
IH, WE, and PJ conceptualized this study. IH and PJ wrote the original draft. PJ, BV, and ZK conducted the formal analysis. SS did the data curation. XA, JM, and CM performed the experiments. JL supervised the experiments. WE, HH, and JL acquired funding. All authors reviewed and edited the manuscript (using https://credit.niso.org/).

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was supported and inspired by the CZI Standards and Technology Working Group and the Deutsche Forschungsgemeinschaft (DFG): BV HE7669/1-2 and PJ EN1093/5-1.

Availability of data and materials
The code used to analyse the data and benchmark the background methods is available on GitHub https://github.com/Hellmann-Lab/scRNA-seq_Contamination [31] under GPL-3.0 license and deposited in Zenodo under DOI 10.5281/zenodo.7941521 [32]. Larger files are available on a separate Zenodo repository [33]. All sequencing files were deposited in GEO under accession number GSE218853 [34].

Declarations

Ethics approval and consent to participate
All procedures performed are IACUC approved on Broad Institute animal protocol # 0061-07-15-1.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 14 November 2022 Accepted: 26 May 2023
Published online: 19 June 2023

References

1. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. Sci Rep. 2016;6:25533.

2. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017;65(4):631-643.e4.
3. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
4. Fleming SJ, Marioni JC, Babadi M. CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. bioRxiv. 2019;791699.
5. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, participants in the 1st Human Cell Atlas Jamboree, et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. 2019;20(1):63.
6. Pääbo S, Irwin DM, Wilson AC. DNA damage promotes jumping between templates during enzymatic amplification. J Biol Chem. 1990;265(8):4718–21.
7. Dixit A. Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments. bioRxiv. 2021;093237.
8. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. 2020;21(1):57.
9. Griffiths JA, Richard AC, Bach K, Lun ATL, Marioni JC. Detection and removal of barcode swapping in single-cell RNA-seq data. Nat Commun. 2018;9(1):2667.
10. Caglayan E, Liu Y, Konopka G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. Neuron. 2022;110:4043–4056.e5.
11. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience. 2020;9. https://doi.org/10.1093/gigascience/giaa151.
12. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol. 2020;38(6):737–46.
13. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011;477(7364):289–94.
14. Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. 2020;21(1):130.
15. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nat Methods. 2020;17(6):615–20.
16. Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief Bioinform. 2021;22(1):416–27.
17. Franzén O, Gan L-M, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database. 2019;2019. https://doi.org/10.1093/database/baz046.
18. Biber J, Hernando N, Forster I, Murer H. Regulation of phosphate transport in proximal tubules. Pflugers Arch. 2009;458(1):39–52.
19. Custer M, Lötscher M, Biber J, Murer H, Kaissling B. Expression of Na-P(i) cotransport in rat kidney: localization by RT-PCR and immunohistochemistry. Am J Physiol. 1994;266(5 Pt 2):F767-74.
20. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–72.
21. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-3587.e29.
22. Huang Y, McCarthy DJ, Stegle O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biol. 2019;20(1):273.
23. Clark JZ, Chen L, Chou CL, Jung HJ, Lee JW, Knepper MA. Representation and relative abundance of cell-type selective markers in whole-kidney RNA-Seq data. Kidney Int. 2019;95(4):787–96.
24. Subramanian A, Sidhom EH, Emani M, Vernon K, Sahakian N, Zhou Y, et al. Single cell census of human kidney organoids shows reproducibility and diminished off-target cells after transplantation. Nat Commun. 2019;10(1):5462.
25. Drokhlyansky E, Van N, Slyper M, Waldman J, Segerstolpe A, Rozenblatt-Rosen O, Regev A. HTAPP_TST- Nuclei isolation from frozen tissue v2. protocols.io. ZappyLab, Inc.; 2020. https://doi.org/10.17504/protocols.io.bhbcj2iw.
26. Huang X, Huang Y. Cellsnp-lite: an efficient tool for genotyping single cells. Bioinformatics. 2021;37:4569–71.
27. R Team Core. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. http://www.R-project.org/.
28. Hahsler M, Piekenbrock M, Doran D. dbscan: Fast density-based clustering with R. J Stat Softw. 2019;91:1–30.
29. Mouselimis L. Gaussian mixture models, K-means, mini-batch-kmeans, K-medoids and affinity propagation clustering [R package ClusterR version 1.2.7]. Comprehensive R Archive Network (CRAN). 2022. https://CRAN.R-project.org/package=ClusterR. Accessed 18 Aug 2022.
30. Ahlmann-Eltze C, Huber W. Comparison of transformations for single-cell RNA-seq data. Nat Methods. 2023;20:665–72.
31. Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, et al. The effect of background noise and its removal on the analysis of single-cell expression data. Github. 2022. https://github.com/Hellmann-Lab/scRNA-seq_Contamination. Accessed 14 May 2023.
32. Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, et al. The effect of background noise and its removal on the analysis of single-cell expression data. Zenodo Code. 2022. https://doi.org/10.5281/zenodo.7941521.
33. Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, et al. The effect of background noise and its removal on the analysis of single-cell expression data. Zenodo Data. 2022. https://doi.org/10.5281/zenodo.7328632.
34. Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, et al. The effect of background noise and its removal on the analysis of single-cell expression data. scRNA-seq and snRNA-seq datasets. Gene Expr Omnibus. 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE218853. Accessed 12 Dec 2022.

Supplementary Information

The effect of background noise and its removal on the analysis of single-cell expression data

Philipp Janssen[1] Zane Kliesmete[1], Beate Vieth[1], Xian Adiconis[2,3], Sean Simmons[2,3], Jamie Marshall[4], Cristin McCabe[2], Holger Heyn[5], Joshua Z. Levin[2,3], Wolfgang Enard[1], Ines Hellmann[1,*],

[1] Anthropology and Human Genomics, Department of Biology II, Ludwig-Maximilians Universitaet, Munich, Germany
[2] Klarman Cell Observatory, Broad Institute of Harvard and MIT, Cambridge, MA USA
[3] Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA USA
[4] Broad Institute of Harvard and MIT, Cambridge, MA USA
[5] CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain

[*] correspondence hellmann@bio.lmu.de

Supplementary Tables

Table S1 Spearman correlation analysis of background noise and barcode swapping. Endogenous and contaminating allele counts refer to *M.m. castaneus* and *M.m. domesticus* allele counts in *M.m. castaneus* cells, respectively. Chimera refer to barcode swapping events that are observable by the association of multiple genes with the same cell barcode (CB)-UMI combination.

	Endogenous vs contaminating allele counts per cell		Chimera vs unique BC-UMI-gene counts per cell	
replicate	rho	p-value	rho	p-value
rep3	0.27	<2.2e-16	0.81	<2.2e-16
rep1	0.07	9e-08	0.81	<2.2e-16
rep2	0.06	0.0021	0.64	<2.2e-16
nuc3	0.03	0.499	0.77	<2.2e-16
nuc2	0.15	2.56e-09	0.52	<2.2e-16

Supplementary Figures

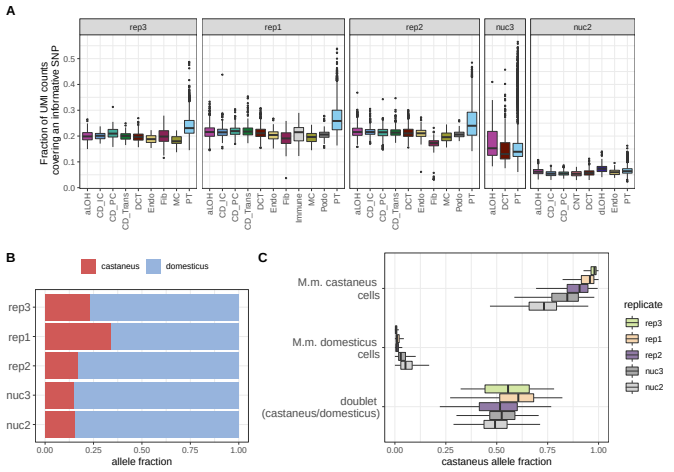


Fig. S1 Detection of cross-genotype contamination A) Variants that are variable between mouse subspecies were used to identify background noise. Boxplots show the fraction of detected molecules (UMI counts) that contain an informative SNP in each cell across cell types with more than 50 cells. CD_IC: intercalated cells of collecting duct; CD_PC: principal cells of collecting duct; CD_Trans: transitional cells of collecting duct; CNT: connecting tubule; DCT: distal convoluted tubule; Endo: endothelial; Fib: fibroblasts; aLOH: ascending loop of Henle; dLOH: descending loop of Henle; MC: mesangial cells; Podo: podocytes. B) For each informative SNP we can detect either the *M.m. castaneus* or the *M.m. domesticus* allele. The stacked bar plots indicate the allele fraction per replicate, integrating over all cells and SNPs. Since each replicate is a mixture experiment with a majority of *M.m. domesticus* cells, *M.m. domesticus* alleles are detected more often at covered informative SNPs. C) *M.m. castaneus* allele frequency per cell in cells from different subspecies and mixed-subspecies doublets. In all replicates varying amounts of *M.m. castaneus* alleles are detected in *M.m. domesticus* cells and vice versa, pointing towards background noise originating from cross-genotype contamination.

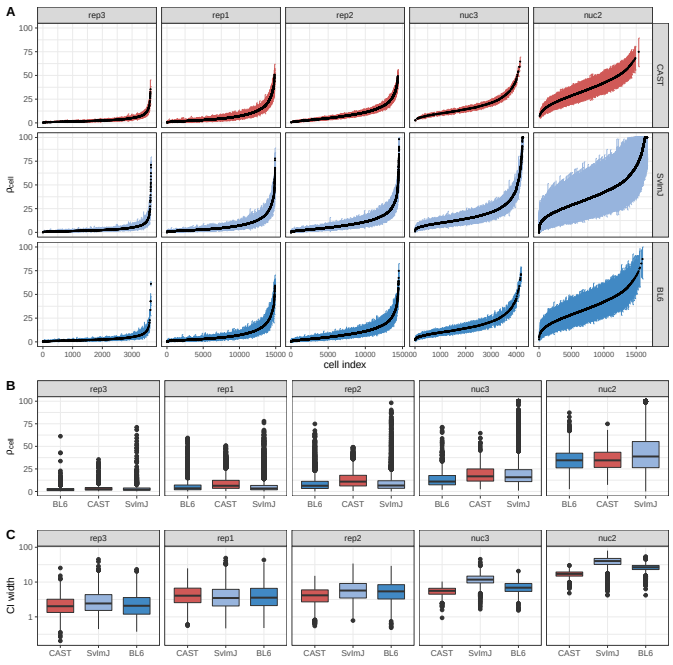


Fig. S2 Estimation of background noise levels. A) Estimates of background noise (ρ_{cell}) per cell. Cells were ordered by ascending ρ_{cell} in each replicate. Colored bars indicate 95% confidence intervals calculated by profile likelihood. B) Summary of ρ_{cell} estimates per strain. C) Width of confidence intervals for ρ_{cell}.

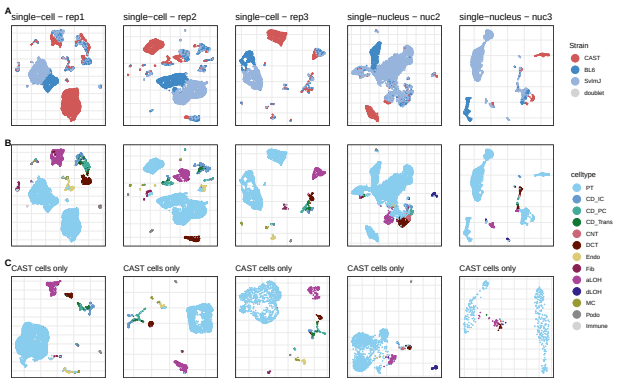


Fig. S3 UMAP visualization showing the composition per replicate of A) all cells, colored by strain assignment, B) all cells, colored by cell type assignment and C) *M. m. castaneus* cells only, colored by cell type assignment.

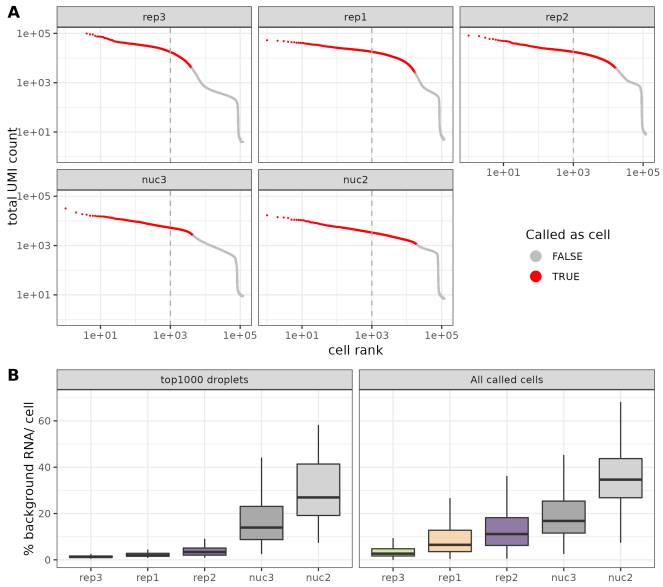


Fig. S4 Definition of true cells and its effect on background noise estimates A) UMI curves showing the total UMI counts for droplet barcodes arranged in descending order. True cells (red) were defined based on a combination of CellRanger cell calls and an inflection point in the UMI curve. The dashed line indicates a hard cutoff of 1000 cells that was used to check the robustness of background noise estimates. B) Estimated background noise levels per cell for the 1000 cells with the highest total UMI counts (left) and for all cells that were defined as true cells based on replicate specific cutoffs (right).

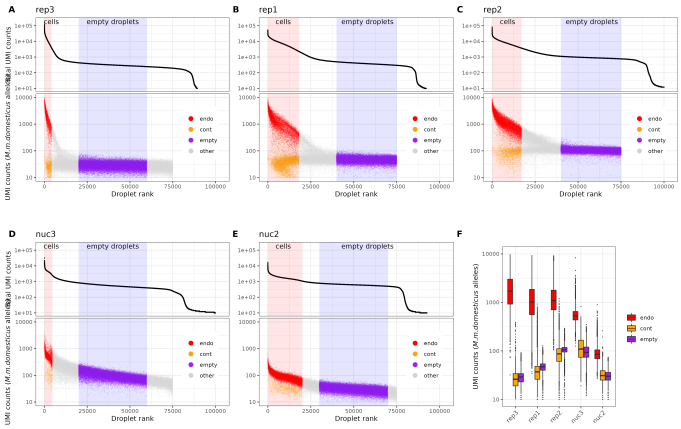


Fig. S5 Definition of endogenous, empty droplet and contamination profiles across replicates. A-E) Droplet barcodes were ordered by their total UMI counts and empty droplets were defined from this UMI curve as barcodes in the low UMI count plateau area (upper panel). UMI counts of reads covering *M. m. domesticus* specific alleles were used to construct three different profiles (lower panel left). *M. m. domesticus* allele counts in *M. m. domesticus* cells were defined as endogenous counts (endo), *M. m. domesticus* allele counts in *M. m. castaneus* cells as contaminating counts (cont) and *M. m. domesticus* allele counts associated with barcodes of the empty droplet plateau as empty droplet counts (empty). F) Boxplots indicate how many UMI counts could be obtained for each profile per droplet barcode.

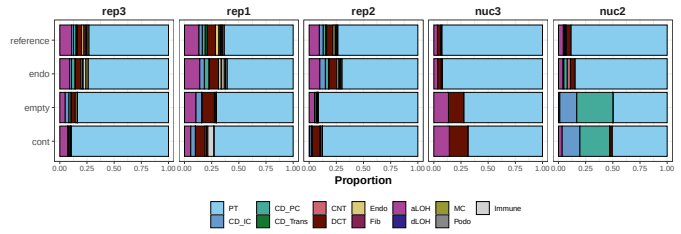


Fig. S6 Dissection of cell type contributions by deconvolution of pseudobulk profiles. The stacked bar plots of 'reference' depict the proportions of cell types in a single cell reference used for deconvolution with SCDC [16]. The 'endo', 'empty' and 'cont' bar plots show the estimated fraction of cell types after deconvolution of pseudobulk profiles that were aggregated for each category.

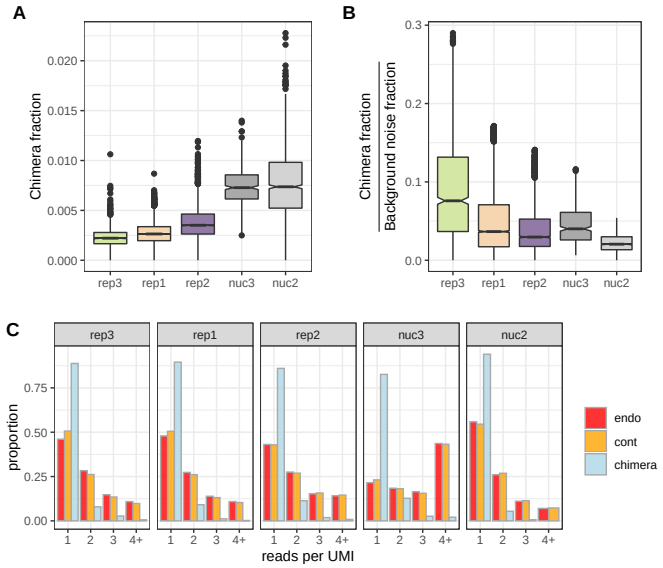


Fig. S7 Identification of barcode swapping due to PCR chimeras. A) Fraction of chimeras per cell. Chimeras were defined as non-unique combinations of cell barcode, UMI and gene. B) Relative fraction of chimeras relative to the estimated level of background noise per cell. C) Distribution of the number of reads per UMI for chimeric molecules, cross-genotype contamination (cont) and endogenous (endo) molecules.

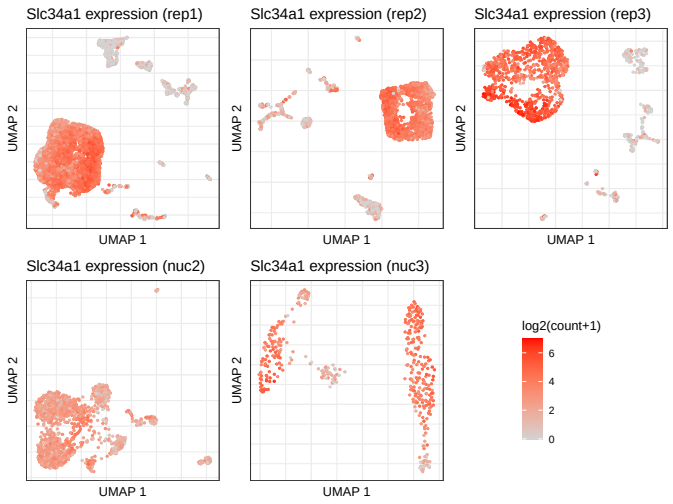


Fig. S8 Slc34a1 expression across replicates. UMAP representation *M. m. castaneus* cells coloured by Slc34a1 expression. Spurious detection of Slc34a1 in all cell clusters is observed in all replicates. In the replicates with the lowest background noise levels (rep1,rep3), Slc34a1 expression is most concentrated in PT cells.

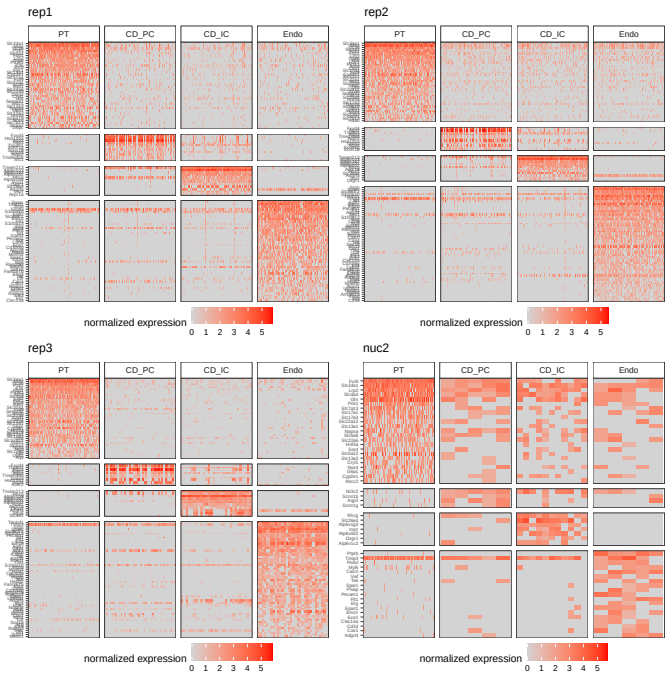


Fig. S9 Expression of cell type marker genes. Heatmaps show the normalized expression of known marker genes for four selected cell types across replicates. Marker genes were obtained from the PanlaoDB database [17] and filtered to select for genes that are detected in at least 50% of the cells of the cell type in which they are expected to be expressed. The replicate nuc3 was excluded from this figure due to an insufficient number of collecting duct and endothelial cells. PT: proximal tubule; CD_IC: intercalated cells of collecting duct; CD_PC: principal cells of collecting duct; Endo: endothelial

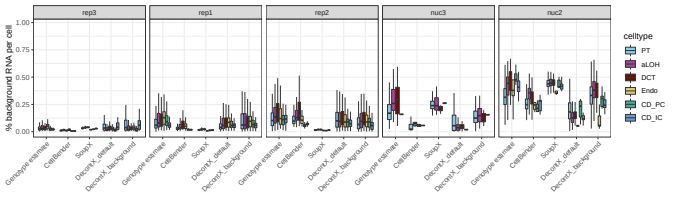


Fig. S10 Estimated background noise levels across cell types. Genotype estimates are inferred based on genetic variants. Cellbender, SoupX and DecontX estimates are calculated for each cell based on a corrected count matrix. PT: proximal tubule; aLOH: ascending loop of Henle; DCT: distal convoluted tubule; Endo: endothelial; CD_PC: principal cells of collecting duct; CD_IC: intercalated cells of collecting duct.

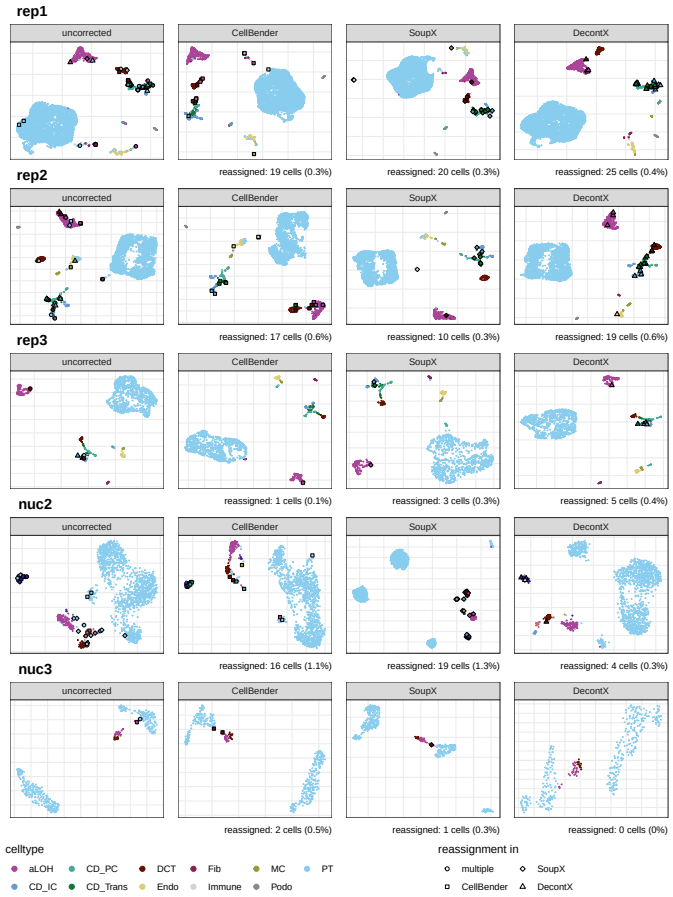


Fig. S11 UMAP representations of all replicates before and after background noise correction. Cells are colored by cell type labels obtained from reference based classification. Individual cells that received a new label after correction are highlighted. In case of the uncorrected data, all cells that received a new label after correction with any of the methods are highlighted. PT: proximal tubule; CD_IC: intercalated cells of collecting duct; CD_PC: principal cells of collecting duct; CD_Trans: transitional cells of collecting duct; CNT: connecting tubule; DCT: distal convoluted tubule; Endo: endothelial; Fib: fibroblasts; aLOH: ascending loop of Henle; dLOH: descending loop of Henle; MC: mesangial cells; Podo: podocytes

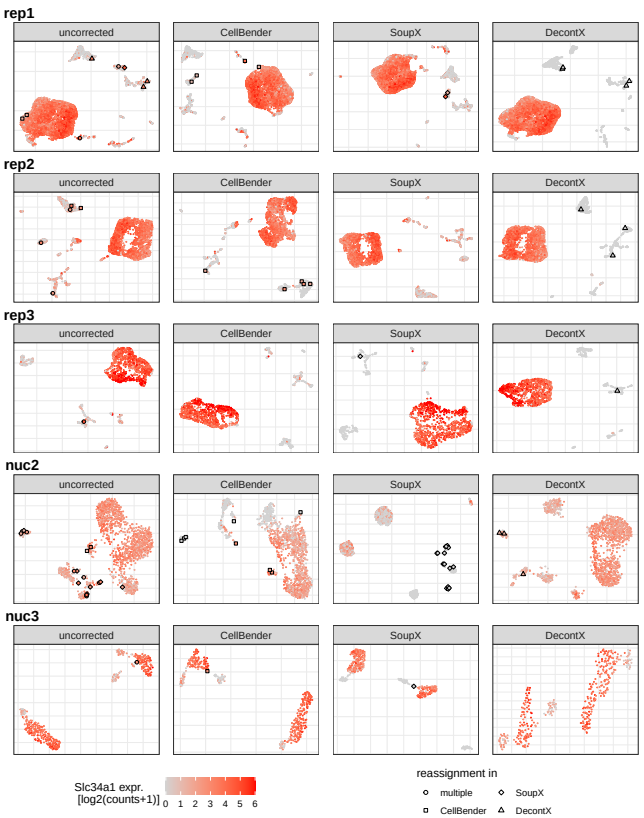


Fig. S12 Detected expression levels of Slc34a1 before and after background noise correction. Cells that were classified as PT cells in the uncorrected data, but got reassigned after correction, are highlighted.

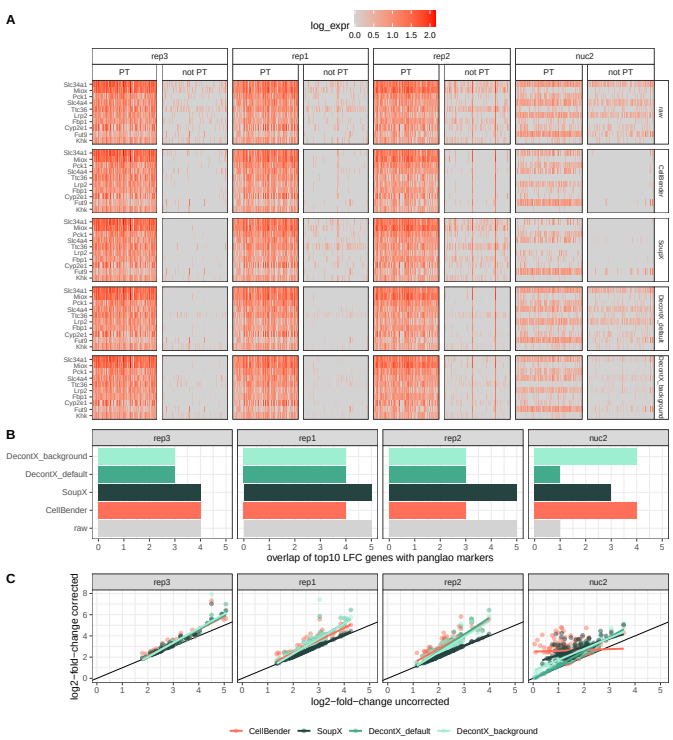


Fig. S13 Effect of background noise correction on marker gene detection. A) Heatmaps depicting the expression of 10 PT marker genes in 100 randomly sampled PT cells and 100 cells from other cell types. The first row of heatmaps is based on the uncorrected count matrix, rows 2-5 on the denoised count matrix output by different methods. B) Overlap of identified and known marker genes. Genes were ranked by log2 fold change between PT an other cells and the overlap of the top 10 genes in this ranking with known marker genes for Proximal Tubule cells from PanglaoDB [17] is shown. C) Log2 fold changes of PangloaDB PT cell marker genes after background noise correction compared to the uncorrected data.

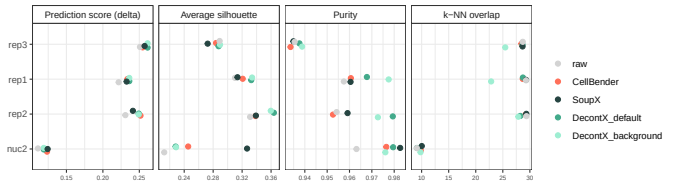


Fig. S14 Evaluation metrics for cell type identification. Prediction score: cell-wise score "delta" of reference based classification with SingleR [21]. Average silhouette: Mean of silhouette widths per cell type. Purity: Cluster purity calculated on cell type lables as ground truth and Louvain clusters as test labels. *k*-NN overlap: overlap of the k=50 nearest neighbors per cell compared to genotype-cleaned reference *k*-NN graph.

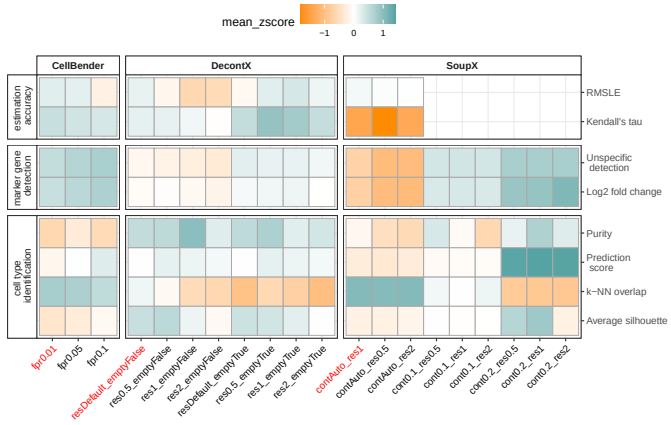


Fig. S15 Evaluation of different parameter settings. Combinations of the most impactful parameter/workflow choices of each method are evaluated. Default parameter settings are highlighted with red font color. For each metric, an average z-score across the replicates rep1, rep2, rep3 and nuc2 is shown, for which higher values indicate better performance. The following parameters were tuned: CellBender: *fpr* (0.01,0.05,0.1); DecontX: cluster lables *z* (resDefault: NULL, res0.5/1/2: vector of cluster labels from Louvain clustering with resolution 0.5/1/2), *background* (emptyFalse: NULL, emptyTrue: provide raw matrix containing empty droplets); SoupX: contamination fraction (contAuto: automatic estimation using *autoEstcont*, cont0.1/0.2: manually set using *setContaminationFraction* (0.1/0.2)), cluster labels (res0.5/1/2: vector of cluster labels from Louvain clustering with resolution 0.5/1/2)

2.2 A non-invasive method to generate induced pluripotent stem cells from primate urine

Johanna Geuder, Lucas E. Wange, Aleksandar Janjic, Jessica Radmer, **Philipp Janssen**, Johannes W. Bagnoli, Stefan Müller, Mari Ohnuki, Wolfgang Enard

"A non-invasive method to generate induced pluripotent stem cells from primate urine"

Scientific Reports 11, 3516 (2021). doi: 10.1038/s41598-021-82883-0

Supplementary Information is freely available at the publisher's website:

https://www.nature.com/articles/s41598-021-82883-0

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

A non-invasive method to generate induced pluripotent stem cells from primate urine

Johanna Geuder[1], Lucas E. Wange[1], Aleksandar Janjic[1], Jessica Radmer[1], Philipp Janssen[1], Johannes W. Bagnoli[1], Stefan Müller[2], Artur Kaul[3], Mari Ohnuki[1]✉ & Wolfgang Enard[1]✉

Comparing the molecular and cellular properties among primates is crucial to better understand human evolution and biology. However, it is difficult or ethically impossible to collect matched tissues from many primates, especially during development. An alternative is to model different cell types and their development using induced pluripotent stem cells (iPSCs). These can be generated from many tissue sources, but non-invasive sampling would decisively broaden the spectrum of non-human primates that can be investigated. Here, we report the generation of primate iPSCs from urine samples. We first validate and optimize the procedure using human urine samples and show that suspension- Sendai Virus transduction of reprogramming factors into urinary cells efficiently generates integration-free iPSCs, which maintain their pluripotency under feeder-free culture conditions. We demonstrate that this method is also applicable to gorilla and orangutan urinary cells isolated from a non-sterile zoo floor. We characterize the urinary cells, iPSCs and derived neural progenitor cells using karyotyping, immunohistochemistry, differentiation assays and RNA-sequencing. We show that the urine-derived human iPSCs are indistinguishable from well characterized PBMC-derived human iPSCs and that the gorilla and orangutan iPSCs are well comparable to the human iPSCs. In summary, this study introduces a novel and efficient approach to non-invasively generate iPSCs from primate urine. This will extend the zoo of species available for a comparative approach to molecular and cellular phenotypes.

Primates are our closest relatives and hence play an essential role in comparative and evolutionary studies in biology, ecology and medicine. We share the vast majority of our genetic information, and yet have considerable molecular and phenotypic differences[1]. Understanding this genotype–phenotype evolution is crucial to understand the molecular basis of human-specific traits. Additionally, it is biomedically highly relevant to interpret findings made in model organisms, such as the mouse, and to identify the conservation and functional relevance of molecular and cellular circuitries[2,3]. However, obtaining comparable samples from different primates, especially during development, is practically and—more importantly—ethically very difficult or even impossible.

Embryonic stem cells have the potential to partially overcome this limitation by their ability to differentiate into all cell types in vitro and divide indefinitely[4]. However, the necessary primary material collection from an embryo is in most cases impossible. Fortunately, a pluripotent state can also be induced in somatic cells by ectopically expressing four genes[5]. Since this discovery of induced pluripotency, great efforts have been made to identify suitable somatic cells[6] and optimize reprogramming methods[7]. Most of this research, however, has focused on human or mouse. While the methods are generally transferable and iPSCs from several different non-human primates[8–10] and other mammals[11,12] have been generated, these methods have not been optimized for non-model organisms.

One major challenge for establishing iPSCs of various non-human primates is the acquisition of the primary cells. So far iPSCs have been generated from fibroblasts, peripheral blood cells or vein endothelial cells derived during medical examinations or from post mortem tissue[8–10,13,14]. However, also these sources impose practical and ethical constraints and therefore limit the availability of the primary material.

To overcome these limitations, we adapted a method of isolating reprogrammable cells from human urine samples[15,16] and applied it to non-human primates (Fig. 1). We find that primary cells can be isolated from

[1]Anthropology and Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Großhaderner Straße 2, 82152 Martinsried, Germany. [2]Institute of Human Genetics, Munich University Hospital, Ludwig-Maximilians-University Munich, 80336 Munich, Germany. [3]Infection Biology Unit, German Primate Center, 37077 Göttingen, Germany. ✉email: ohnuki@biologie.uni-muenchen.de; enard@bio.lmu.de

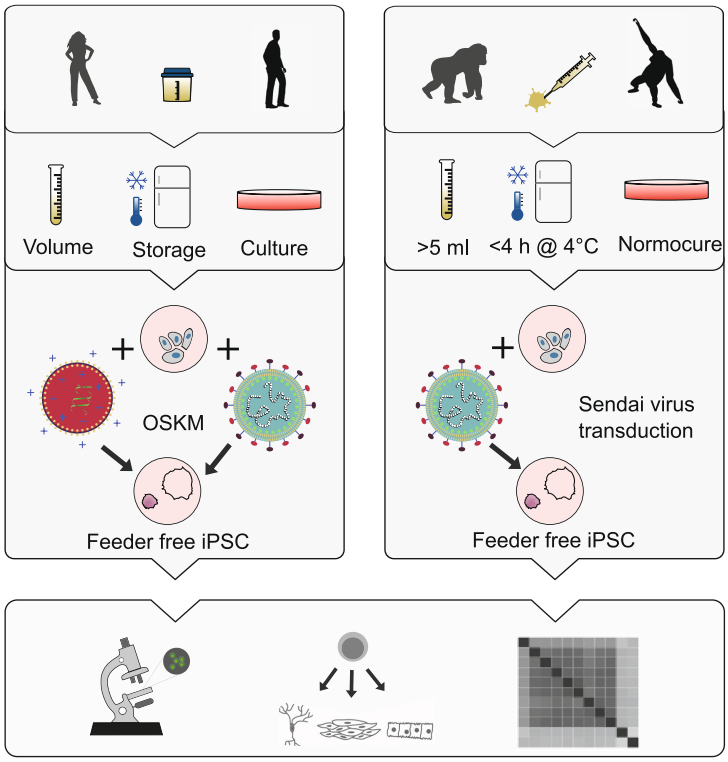


Figure 1. Workflow overview for establishing iPSCs from primate urine. We established the protocol for iPSC generation from human urine based on a previously described protocol[16]. We tested volume, storage and culture conditions for primary cells and compared reprogramming by overexpression of OCT3/4, SOX2, KLF4 and MYC (OSKM) via lipofection of episomal vectors and via transduction of a Sendai virus derived vector (SeV). We used the protocol established in humans and adapted it for unsterile floor-collected samples from non-human primates by adding Normocure to the first passages of primary cell culture and reprogrammed visually healthy and uncontaminated cultures using SeV. Pluripotency of established cultures was verified by marker expression, differentiation capacity and cell type classification using RNA sequencing.

unsterile urine sampled from the floor, can be efficiently reprogrammed using the integration-free Sendai Virus[17] and can be maintained under feeder-free conditions as shown by generating iPSCs from human, gorilla and orangutan.

Results

Isolating human urinary cells from small-volume and stored samples. To assess which method is most suitable for isolating and reprogramming primate cells, we first tested different procedures using urinary cells from human samples (Fig. 1). We collected urine from several humans in sterile beakers and processed them as described in Zhou et al.[15,16]. We found varying cell numbers in the urine samples (range 46–2250 cells per ml; Supplementary Table S1) with about 60% living cells. As previously reported[18,19], we initially observed two morphologically distinct colony types that became indistinguishable after the first passage and consisted of grain-shaped cells that proliferated extensively (Fig. 2a, Supplementary Figure S1b). In total we processed 19 samples of several individuals in 122 experiments using different volumes and storage times (Supplementary Table S2). Similar to previous reports[20], we isolated an average of 7.6 colonies per 100 ml of urine when processing samples immediately with a considerable amount of variation among samples (0–70 colonies per 100 ml, Supplementary Table S2) and among aliquots (0–160 per 100 ml; Supplementary Table S2; Fig. 2b), but no difference between sexes (Supplementary Table S2). Furthermore, storing samples for up to 4 h at room temperature or on ice did not influence the number of isolated colonies (9 samples, 7.4 colonies on average per 100 ml,

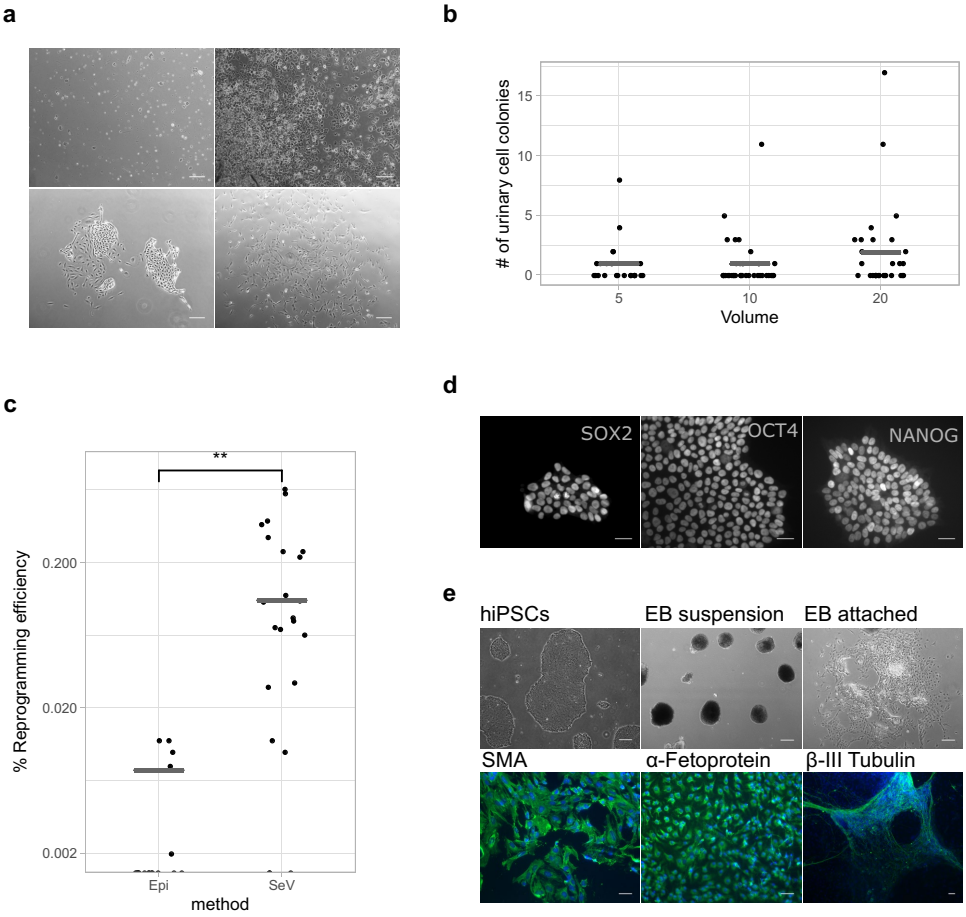


Figure 2. Establishing urinary cell isolation and reprogramming to iPSCs in human samples. (**a**) Human urine mainly consists of squamous cells and other differentiated cells that are not able to attach and proliferate (upper row). After ~ 5 days, the first colonies become visible and two types of colonies can be distinguished as described in Zhou (2012). Scale bars represent 500 µm. (**b**) Isolation efficiency of urine varies between samples. The efficiency between 5 ml, 10 ml and 20 ml of starting material is not different (Fisher's exact test $p > 0.5$). (**c**) SeV mediated reprogramming showed significantly higher efficiency than Episomal plasmids (Wilcoxon rank sum test: p = 1.1e−05). (**d**) Established human colonies transduced with SeV expressed Nanog, Oct4 and Sox2; Scale bars represent 50 µm and (**e**) differentiated to cell types of the three germ layers; scale bar represents 500 µm in the phase contrast pictures and 100 µm in the fluorescence pictures. See also Supplementary Figure S1.

range: 0–17). As sample volumes can be small for non-human primates, we also tested whether colonies can be isolated from 5, 10 or 20 ml of urine (Fig. 2b). We found no evidence that smaller volumes have lower success rates as we found that for 42% of the 5 ml samples, we could isolate at least one colony (Supplementary Table S2). Many more samples and conditions would be needed to better quantify the influence of different parameters on the isolation efficiency of colonies. However, in most practical situations such parameters would not be used to make a decision as one would anyway try to obtain colonies with the urine samples at hand, especially in our case where samples from primates are rare. Fortunately, low-volume human urine samples stored for a few hours at room temperature or on ice are a possible source to establish primary urinary cell lines. In summary, these experiments are a promising starting point for the use of small-volume urine samples from non-human primates to generate primary cell lines, which may then be reprogrammed into iPSCs.

Reprogramming human urinary cells is efficient when using suspension-Sendai Virus transduction. Next, we investigated which integration-free overexpression strategy would be the most suitable to induce pluripotency in the isolated urine cells. To this end we compared transduction by a vector derived from the RNA-based Sendai Virus[14,17] in suspension[10], to lipofection with episomal plasmids (Epi) derived from the Epstein Barr virus[21,22]. We chose to use the suspension transduction method as it yielded a significantly higher reprogramming efficiency than the method on attached cells (suspension reprogramming efficiency: 0.24%, N = 7; attached reprogramming efficiency: 0.09%, N = 7; Wilcoxon rank sum test: p = 0.003; Supplementary Table S3, Supplementary Figure S2d). Both systems have been previously reported to sufficiently induce reprogramming of somatic cells without the risk of genome integrations. In our experiments presented here, transduction of urinary cells with a Sendai Virus (SeV) vector containing Emerald GFP (EmGFP) showed substantially higher efficiencies than lipofection with episomal plasmids (~ 97% versus ~ 20% EmGFP+; Supplementary Figure S2a and S2b). We assessed the reprogramming efficiency of these two systems by counting colonies with a pluripotent-like cell morphology. Using SeV vectors, 0.19% of the cells gave rise to such colonies (Fig. 2c). In contrast, when using Episomal plasmids only 0.009% of the cells gave rise to colonies with pluripotent cell-like morphology (N = 23 and 18, respectively; Wilcoxon rank sum test: p = 0.00005), resulting in at least one colony in 87% and 28% of the cases. Furthermore, the first colonies with a pluripotent morphology appeared 5 days after SeV transduction and 14 days after Epi lipofection. To test whether the morphologically defined pluripotent colonies also express molecular markers of pluripotency, we isolated flat, clear-edged colonies from 5 independently transduced urinary cell cultures on day 10. All clones expressed POU5F1 (OCT3/4), SOX2, NANOG and differentiated into the three germ layers during embryoid body formation as shown by immunocytochemistry (Fig. 2d,e). Notably, while the transduced cells also expressed the pluripotency marker SSEA4, this was also true for the primary urinary cells (Supplementary Figure S2c). SSEA4 is known to be expressed in urine derived cells[18,23] and hence it is an uninformative marker to assess the reprogramming of urinary cells to iPSCs. Furthermore, SeV RNA was always absent after the first five passages (Supplementary Figure S3) and the pluripotent state could be maintained for over 100 passages (data not shown).

In summary, we find that the generation of iPSCs from human urine samples is possible from small volumes, and our results also reveal that reprogramming is most efficient when using suspension SeV transduction. Hence, we used this workflow for generating iPSCs from non-human primate cells.

Isolating cells from unsterile primate urine. For practical and ethical reasons, the collection procedure is a decisive difference when sampling urine from non-human primates (NHPs). Samples from chimpanzees, gorillas and orangutans were collected by zoo keepers directly from the floor, often with visible contamination. Initially, culturing these samples was not successful due to the growth of contaminating bacteria. The isolation and culture of urinary cells only became possible upon the addition of Normocure (Invivogen), a broad-spectrum antibacterial agent that actively eliminates Gram+ and Gram− bacteria from cell cultures. We confirmed that Normocure did not affect the number of colonies isolated from sterile human samples (Supplementary Table S2). Furthermore, many NHP samples also had volumes below 5 ml. We attempted to isolate cells from a total of 70 samples, but only 24 NHP samples showed collection parameters comparable to human urine samples as described above (≥ 5 ml of sample, < 4 h storage at RT or 4 °C and no visible contamination). From chimpanzees, gorillas and orangutans we collected a total of 87, 70 and 39 ml of urine in 11, 8 and 5 samples from several individuals and isolated 0, 5 and 2 colonies respectively (Supplementary Table S4). For gorilla and orangutan this rate (7.3 and 5.2 colonies per 100 ml urine) is not significantly different from the rate found for human samples (6.0 per 100 ml across all conditions in Supplementary Table S2, p = 0.8 and 0.6, respectively, assuming a Poisson distribution). However, obtaining zero colonies from 87 ml of chimpanzee urine is less than expected, given the rate found in human samples (p = 0.005). While isolating primary cells from urine samples seems comparable to humans in two great ape species, it seems to have at least a two- to threefold lower rate in our closest relatives, suggesting that the procedure might work in many but not in all NHPs. Fortunately, it is possible to culture many samples in parallel so that screening for urinary cells in a larger volume with more samples is relatively easy.

The first proliferating cells from orangutan and gorilla could be observed after six to ten days (Fig. 3a,b) in culture and could be propagated for several passages, which is comparable to human cells. While we observed different proliferation rates and morphologies among samples, these did not systematically differ among individuals or species (Fig. 3b). Infection with specific pathogens, including simian immunodeficiency virus (SIV), herpes B virus (BV, Macacine alphaherpesvirus 1), simian T cell leukemia virus (STLV) and simian type D retroviruses (SRV/D), was not detected in these cells (data not shown).

Expression patterns of urinary cells are most similar to mesenchymal stem cells, epithelial cells and smooth muscle cells. To characterize the isolated urinary cells, we generated expression profiles using prime-seq a 3′ tagged RNA-seq protocol[24–26], on early passage primary urinary cells (p1–3) from three humans, one gorilla and one orangutan. Note that some of these samples contained cells from 1–4 different colonies (Supplementary Table S2 and S4) and hence could be mixtures of different cell types. To classify these urinary cells we compared their expression profile to 713 microarray expression profiles grouped into 38 cell types[27] using the SingleR package[28]. SingleR uses the most informative genes from the reference dataset and iteratively correlates it with the expression profile to be classified. The most similar cell types were mesenchymal stem cells, epithelial cells and/or smooth muscle cells and at least two groups are evident among the six samples (Fig. 3c). To further investigate these cell types, we isolated 19 single colonies from six different individuals (Supplementary Table S1) and analyzed their expression profiles as described above. A principal component analysis revealed three clearly distinct clusters A, B and C with 10, 6 and 3 colonies, respectively (Fig. 3d). When we classified these 19 profiles using SingleR[27,28] as described above, we found the three colonies from cluster C

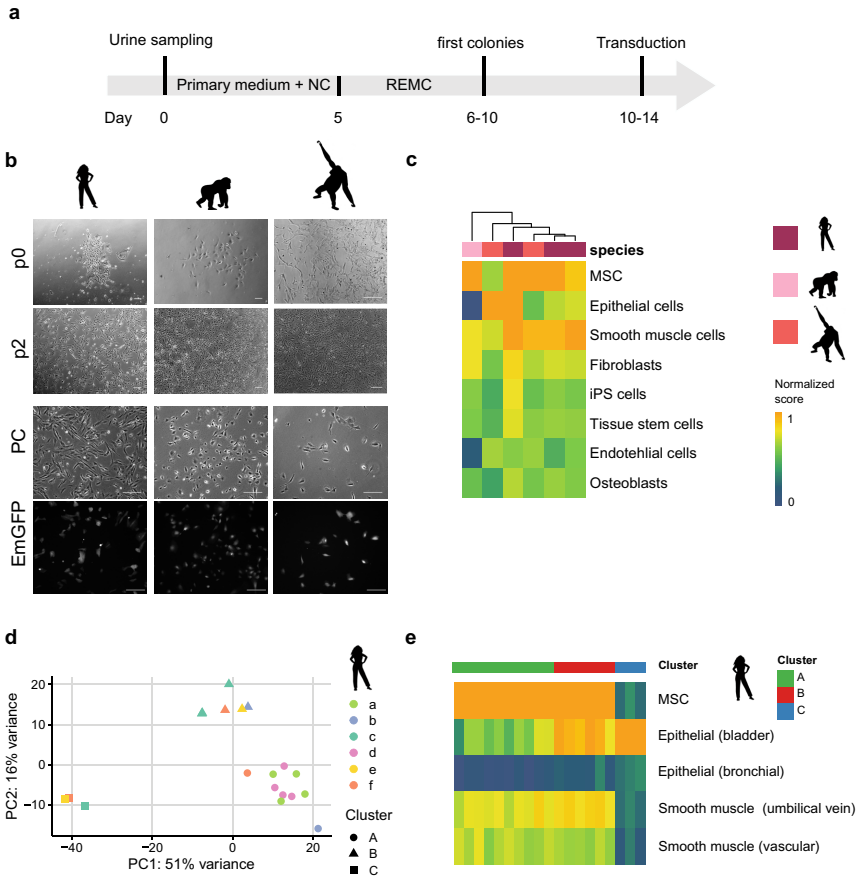


Figure 3. Isolation and characterization of primate urinary cells. (**a**) Workflow of cell isolation from primate urine samples. *NC* Normocure, *REMC* renal epithelial mesenchymal cell medium. (**b**) Primary cells obtained from human, gorilla and orangutan samples are morphologically indistinguishable and display similar EmGFP transduction levels. Scale bars represent 400 µm. (**c**) The package SingleR was used to correlate the expression profiles from six samples of primate urinary cells (passage 1–3) to a reference set of 38 human cell types. Normalized scores of the eight cell types with the highest correlations are shown (*MSC* mesenchymal stem cells, *SM* smooth muscle, *Epi* epithelial, *Endo* endothelial). Color bar indicates normalized correlation score. (**d**) Principal component analysis of primary cells from single colony lysates using the 500 most variable genes. (**e**) Heatmap of normalized SingleR scores show that cluster C is classified as epithelial cell originating from the bladder. The scores for MSCs in Cluster A and B are similarly high, although cluster B also shows higher scores for epithelial cells than cluster A. See also Supplementary Figure S5.

clearly classified as epithelial cells from the bladder (Fig. 3e). This cluster shows high KRT7 expression, as also described in Dörrrenhaus et al.[19] as well as high FOXA1 expression, both hinting towards an urothelial origin (Supplementary Figure S4). The colonies of the other two clusters are classified as MSCs, whereas cluster B also has a high similarity to epithelial profiles (Fig. 3e). They could resemble the two renal cell types described in Dörrrenhaus et al.[19] and are probably derived from the kidney as also evident by their PAX2 and MCAM expression (Supplementary Figure S4). We also used differential gene expression and Reactome pathway analysis[29] to further characterize the differences between these clusters (Supplementary Figure S4a, S4c). In sum, our findings indicate that at least three types of proliferating cells can be isolated from urine, one of urothelial and two of renal origin and that the same types can also be isolated from gorilla and orangutan.

Reprogramming efficiency of urinary cells is similar in humans and other primates. To generate iPSCs from the urinary cells isolated from gorilla and orangutan, we used Sendai Virus (SeV) transduction and the reprogramming timeline that we found to be efficient for human urinary cells (Fig. 4a). Human, gorilla and orangutan urinary cells showed similarly high transduction efficiencies with the EmGFP SeV vector (data not shown). Transduction with the reprogramming SeV vectors led to initial morphological changes after 2 days in all three species, when cells began to form colonies and became clearly distinguishable from the primary cells (Fig. 4b). When flat, clear-edged colonies appeared that contained cells with a large nucleus to cytoplasm ratio, these colonies were picked and plated onto a new dish. We found that the efficiency and speed of reprogramming was variable (Supplementary Figure S5b), probably depending on the cell type, the passage number and the acute state ("health") of the cells, in concordance with the variability and efficiency found in other studies utilizing urine cells as a source for iPSCs[15]. Also the mean reprogramming efficiency over all replicates was different (Kruskal–Wallis test, $p = 0.015$) for human (0.19%), gorilla (0.28%) and orangutan (0.061%). However, many more samples would be necessary to disentangle the effects of all these contributing factors. Of note, we observed that the orangutan iPSCs showed more variability in proliferation rates and morphology compared to human and gorilla iPSCs. Several subcloning steps were needed until a morphologically stable clone could be generated. However, the resulting iPSCs were stable and had the same properties as the other iPSCs (Fig. 4). To what extent this is indeed a property of the species is currently unclear. Importantly, from all primary samples that were transduced, colonies with an iPSC morphology could be obtained. So, while considerable variability in reprogramming efficiency exists, the overall success rate is sufficiently high and sufficiently similar in humans, gorillas and orangutans.

Urine derived primate iPSCs are comparable to human iPSCs. We could generate at least two lines per individual from each primary cell sample, all of which showed Oct3/4, TRA-1-60, SSEA4 and SOX2 immunofluorescence (Fig. 4c). Furthermore, karyotype analysis by G-banding in three humans, one gorilla and one orangutan iPS cell line revealed no recurrent numerical or structural aberrations in 33–60 metaphases analyzed per cell line. All five cell lines analyzed showed inconspicuous and stable karyotypes (Supplementary Figure S6). iPSCs from all species could be expanded for more than fifty passages, while maintaining their pluripotency, as shown by pluripotency marker expression (Fig. 4c) and differentiation capacity via embryoid body formation (Fig. 4d,e). Both the human and NHP iPSCs differentiated into ectoderm (beta-III Tubulin), mesoderm (α-SMA) and endoderm (AFP) lineages (Fig. 4e, Figure S7a). Dual-SMAD inhibition led to the formation of neurospheres in floating culture, as confirmed by neural stem cell marker expression (NESTIN+, PAX6+) using qRT-PCR (Supplementary Figure S7b).

To further assess and compare the urine-derived iPSCs, we generated RNA-seq profiles from nine human, three gorilla and four orangutan iPSC lines as well as the six corresponding primary urinary cells (see analysis above). As an external reference, we added a previously reported and well characterized blood-derived human iPS cell line that was generated using episomal vectors and adapted to the same feeder-free culture conditions as our cells (1383D2)[30]. All lines were grown and processed under the same conditions and in a randomized order in one experimental batch. We picked one colony per sample and used prime-seq, a 3′ tagged RNA-seq protocol[24–26] to generate expression profiles with 19,000 genes detected on average.

We classified the expression pattern of the iPSCs relative to the reference dataset of 38 cell types using SingleR as described for the urinary cells. ES cells or iPS cells are clearly the most similar cell type for all our iPS samples including the external PBMC-derived iPSC line (Fig. 5a). Principal component analysis of the 500 most variable genes (Fig. 5b), shows clear clustering of the samples according to cell type (54% of the variation in PC1) and species (23% of the variance in PC2). The external, human blood-derived iPSC line is interspersed among our human urine derived iPS cell lines. Using the pairwise Euclidean distances between samples to assess similarity, they also cluster first by cell type and then by species (Supplementary Figure S5d). When classifying the expression pattern of the iPSCs relative to a single cell RNA-seq dataset covering distinct human embryonic stem cell derived progenitor states (Chu et. al. 2016), again all our iPSC lines are most similar to embryonic stem cells and are indistinguishable from the external PBMC-derived iPSC line (Fig. 5c), also confirming the immunostainings. Finally, expression distances within iPS cells of the same species were similar, independent of the individual and donor cell type (Fig. 5d).

Taken together, these analyses do not only indicate that our urine derived iPS cells show a pluripotent expression profile and differentiate as expected for iPS cells but can also not be distinguished from an iPSC line derived in another laboratory from another cell type with another vector system. Hence, the expression differences among species are far larger than these technical sources of variation, indicating that these cells are well suited to assess species differences among primates in iPS cells as well as in cell types derived from these pluripotent cells by in vitro differentiation strategies.

Discussion

Here, we adapted a previously described protocol for human urine samples[16] to isolate proliferating cells from unsterile primate urine. We show that these urinary cells can be efficiently reprogrammed into integration-free and feeder-free iPSCs, which are closely comparable among each other and to other iPSCs. Our findings have implications for generating and validating iPSCs from primates and other species for comparative studies. Additionally, some aspects might also be of relevance when generating iPSCs from human urinary cells for medical studies.

Human urine mainly contains cells, such as squamous cells, which are terminally differentiated and cannot attach or proliferate in culture. The first proliferating cells from human urine were isolated in 1972[31] and since then a variety of different cells have been isolated and described that can proliferate, differentiate and be

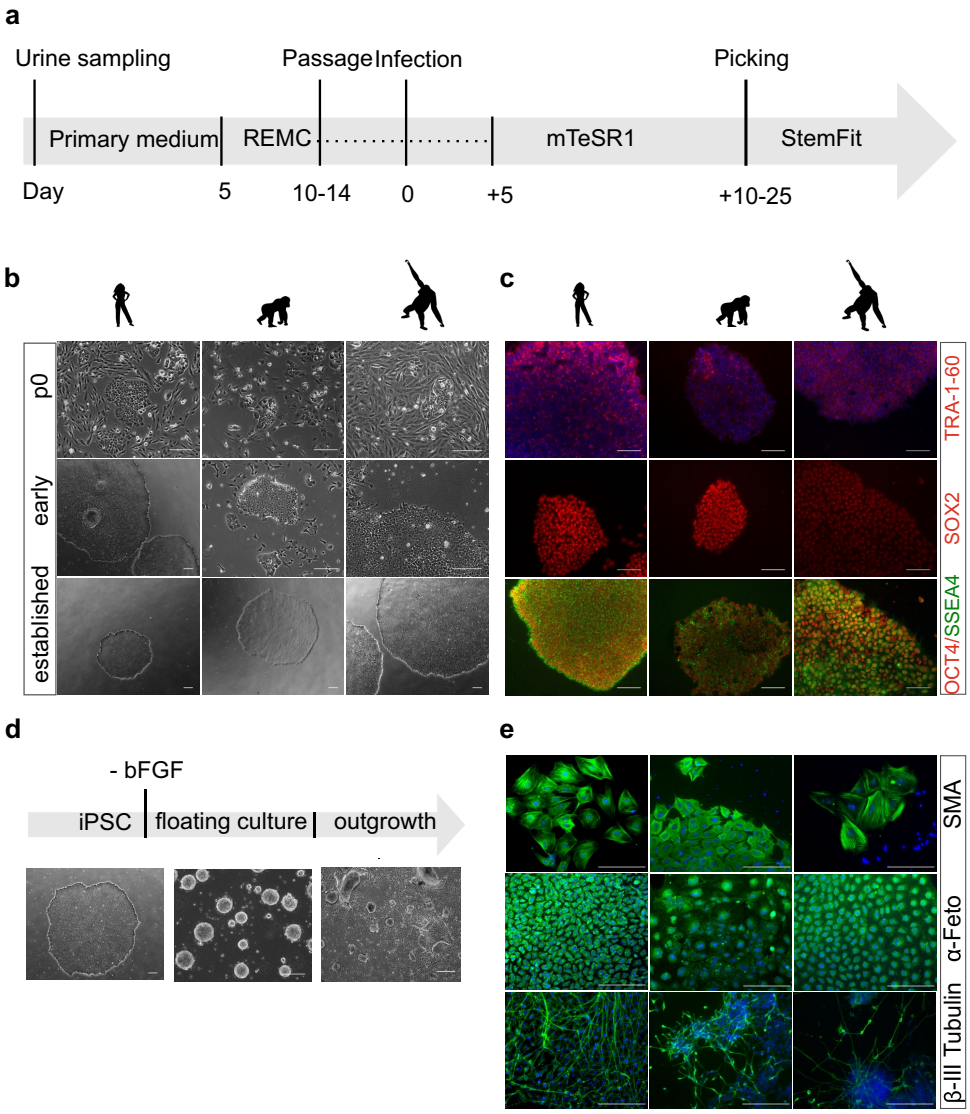


Figure 4. Generation and characterization of primate iPSCs. (**a**) Workflow for reprogramming of primate urinary cells. Urine collection and cell seeding is carried out in primary medium, then after 5 days changed to REMC medium, and only passaged for the first time after 10–14 days. When the cells reach confluency reprogramming is induced and after 5 days the medium is changed to mTeSR1. Once the reprogrammed cells are ready to be picked, the cells are seeded in StemFit medium. *REMC* renal epithelial mesenchymal cell medium. (**b**) Cell morphology of the three species is comparable before (p0), during (p1–3) and after reprogramming (~ p5). Scale bar represents 400 µm. (**c**) Immunofluorescence analysis of pluripotency associated proteins at passage 10–15: TRA-1-60, SSEA4, OCT4 and SOX2. Nuclei were counterstained with DAPI. Scale bars represent 200 µm. (**d**) Differentiation potency into the three germ layers. iPSC colony before differentiation, after 8 days of floating culture and after 8 days of attached culture. Scale bar represents 400 µm. (**e**) Immunofluorescence analyses of ectoderm (β-III Tubulin), mesoderm (α-SMA) and endoderm markers (α-Feto) after EB outgrowth. Nuclei were counterstained with DAPI. Scale bars represent 400 µm. See also Supplementary Figure S7a.

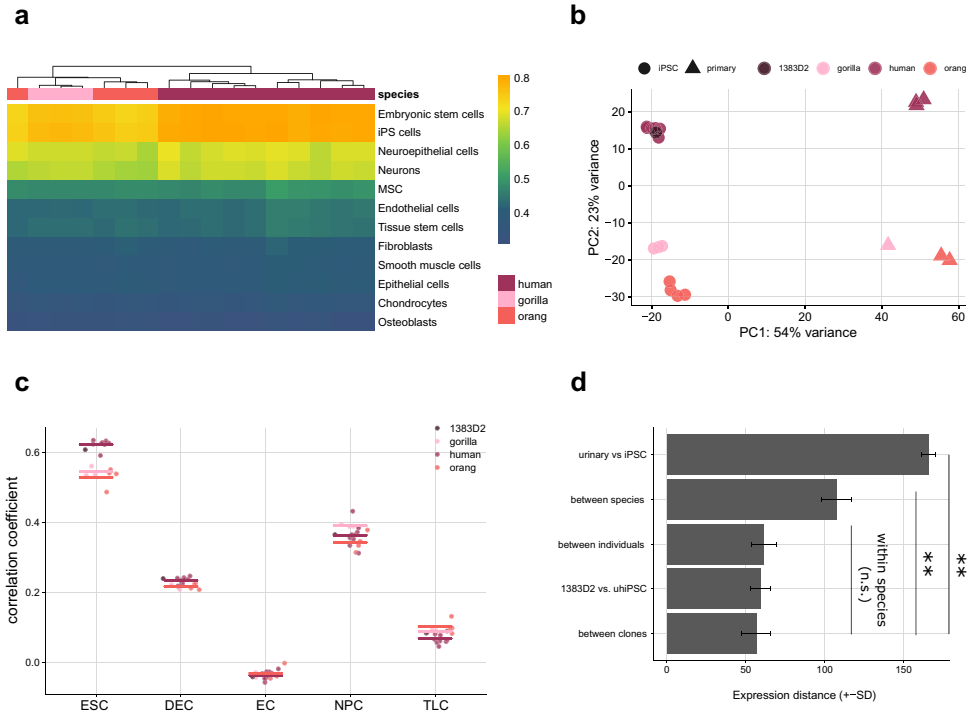


Figure 5. Characterization of primate iPSCs by expression profiling. (**a**) The package SingleR was used to correlate the expression profiles from seventeen samples of primate iPSCs (passage 1–3) to a reference set of 38 human cell types. The twelve cell types with the highest correlations are shown (*MSC* mesenchymal stem cells). All lines are similarly correlated to embryonic stem cells and iPS cells. Color bar indicates correlation coefficients. (**b**) Principal component analysis of primary cells and derived iPSC lines using the 500 most variable genes. PC1 separates the cell types and PC2 separates the species from each other. (**c**) Correlation coefficient of iPSCs compared to a single cell dataset covering distinct human embryonic stem cell derived progenitor states (Chu et al. 2016). (**d**) Expression distances of all detected genes are averaged from pairwise distances for six different groups of comparisons. Note that the distance between individuals and between species is calculated within iPSCs and distances between individuals within species. Pairwise t-tests are all below 0.01 (**) for comparisons to the cell-type and species distance and all above 0.05 (n.s.) for comparisons within the species. See also Supplementary Figure S5.

reprogrammed to iPSCs (see[32] for a recent overview). As these urine-derived stem cells (UDSCs) can be isolated non-invasively at low costs and reprogrammed efficiently[16], they are increasingly used to generate iPSCs from patients (e.g.[33–35]). Perhaps the only major drawback of using UDSCs for iPSC generation is that the number of UDSCs that can be grown per milliliter is quite variable among samples. While parameters such as body size, age and cell count correlate with the number of isolated colonies[20], isolation can fail despite large volumes and can be successful despite small volumes (Supplementary Table S1, Supplementary Table S2). As UDSC culturing is neither very cost- nor time-intensive, the best practical solution will in most cases be to try isolating UDSCs independent of those parameters.

While it is known for a long time that different types of UDSCs can be isolated, the quantitative relation between morphology, marker expression, potency and reprogramming efficiency among the different UDSCs is not clear. The RNA-seq profiles of single colonies presented here, allow for the first time to classify them based on genome-wide expression patterns. In agreement with previous findings using marker staining and morphological analysis[19], we find three different cell types, of which one is most similar to epithelial cells from the bladder and the other two are most similar to mesenchymal stem cells and probably originate from the kidney. Importantly, all three cell types seem to reprogram with sufficient efficiency and the expression of pluripotency markers like KLF4 and OCT3/4 in all three cell types (Supplementary Figure S4) might be one factor why the reprogramming efficiency of UDSCs is relatively high compared to other primary cells. Regarding the reprogramming method,

we find that transduction using the commercial Sendai Virus based vector in suspension[10] is substantially more efficient for UDSCs than lipofection of episomal plasmids, and also leads to a change in morphology within 2 days. While it is established that Sendai Virus reprogramming is an expensive but efficient method to generate iPSCs from fibroblasts[7,36], our findings indicate that the suspension method might be especially efficient for UDSCs. Finally, a relevant side note of our findings is that SSEA4, which is occasionally used as a marker for pluripotency[37,38], is not useful when starting from urinary cells as these express SSEA4 at already high levels (Supplementary Figure S2c). In summary, our findings contribute to a better understanding of human UDSCs and to a method to more efficiently reprogram them into iPSCs.

Maybe more important are the implications of our study for isolating urinary stem cells for the generation of iPSC from primates and other mammals. This could be useful in contexts where invasive sampling is difficult, as it is the case for non-model primates and many other mammals, and where iPSCs are needed for conservation[11] or comparative approaches as discussed below. So how likely is it that one can find UDSCs in other primates and mammals? In humans, UDSCs originate from the kidney and the urinary tract as also shown by our transcriptional profiles. We isolated UDSCs from orangutan and gorilla and found similar transcriptional profiles, morphologies and growth characteristics. Given the general similarity of the urinary tract in mammals and our successful isolation of UDSCs in two apes, it seems likely that most primates, and maybe even most mammals, shed UDSCs in their urine. However, our failure to isolate UDSCs from chimpanzees suggests that even very closely related species might have at least 2–3 times less of those cells in their urine. An alternative possibility is that the culture conditions, e.g. the FBS, do not work for isolating chimpanzee UDSCs. However, given that UDSCs from gorilla and orangutan can be isolated under these conditions and fetal calf serum works for tissue cultures of chimpanzee kidneys[39], we think that a lower concentration of UDSCs in some species is the more likely cause. Hence, from which species UDSCs can be isolated in practice might depend mainly on the concentration of UDSCs and the available amount of urine. Fortunately, this can be easily tested for any given species of interest, as culturing systems are very cost-efficient. Furthermore, our procedure to use unsterile samples from the ground to isolate such cells broadens the practical implementation of this approach considerably.

Given that it is possible to isolate UDSCs from a species, the efficiency of reprogramming and iPSC maintenance will determine whether one can generate stable iPSCs from them. Fortunately, the efficiency of reprogramming UDSCs is shown to be high, probably higher than for many other primary cell types[6]. This is especially true when using SeV transduction in suspension as is evident from the fact that we could generate iPSCs from all twelve UDSC reprogramming experiments (Supplementary Table S5). To what extent this reprogramming procedure works in other species is currently unclear, but as the Sendai virus is thought to infect all mammalian cells[40] it could be widely applicable. Additionally, iPSCs have been previously generated from many species, even avian species[11], when using human reprogramming factors and culture conditions, albeit with over tenfold lower reprogramming efficiencies[41,42]. So, while in principle it should be possible to isolate iPSCs from many or even all mammals, variation in reprogramming efficiency with human factors and culture conditions to keep cells pluripotent with and without feeder cells[42] will considerably vary among species and will make it practically difficult to obtain and maintain iPSCs from some species. Investigating the cause of this variation more systematically will be important to better understand pluripotent stem cells in general and to generate iPSCs from many species in practice. Recent examples of such fruitful investigations include the optimization of culture conditions for baboons[43], and the optimization of feeder-free culture conditions for rhesus macaques and baboons[42]. A related aspect of generating iPSCs from different species is testing whether iPSCs from a given species are actually *bona fide* iPSCs. While for humans a variety of tools exist, such as predictive gene expression assays, validated antibody stainings and SNP arrays for chromosomal integrity, these tools cannot be directly transferred to other species. Fortunately, due to the availability of genome sequences, RNA-sequencing in combination with human or mouse reference cell types to which generated iPSCs can be compared, but also rather traditional techniques such as karyotyping, the characterization of non-human iPSCs becomes feasible as also shown in this paper. In summary, while extending the zoo of comparable iPSCs is a daunting task and requires considerable more method development, we think our method to isolate UDSCs from unsterile urine could be a promising tool in this endeavor.

Assuming that our approach works in at least some non-human primates (NHPs), the effectiveness and non-invasiveness of the protocol allows sampling many more individuals and species than currently possible. Why is this important? So far, iPSCs have been generated from only a few individuals in a very limited set of NHP species. One main application is to model biomedical applications of iPSCs in primates such as rhesus macaques or marmosets[44]. As these species are used as model organisms, non-invasive sampling is less of an issue. Another main application are studies investigating the molecular basis of human-specific phenotypes e.g. by comparing gene expression levels in humans, chimpanzees and an outgroup[8,9,45,46] to infer human-specific changes more robustly[47]. A third type of application with considerable potential has been explored much less, namely using iPSCs in a comparative framework to identify molecular or cellular properties that are conserved, i.e. functional across species[2,3,48]. This is similar to the comparative approach on the genotype level in which DNA or protein sequences are compared in orthologous regions among several species to identify conserved, i.e. functional elements[49]. This information is crucial, for example, when inferring the pathogenicity of genetic variants[50]. Accordingly, it would be useful to know whether a particular phenotypic variant, e.g. a disease associated gene expression pattern, is conserved across species. This requires a comparison of the orthologous cell types and states among several species. Primates are well suited for such an approach, because they bridge the evolutionary gap between human and its most important model organism, the mouse, and because phenotypes and orthologous cell states can be more reliably compared in closely related species. However, for practical and ethical reasons, orthologous cell states are difficult to obtain from several different primates. Hence, just as human iPSCs allow one to study cell types and states that are for practical and ethical reasons not accessible, primate iPSCs extend the comparative approach to these cell types and states, leveraging unique evolutionary information that is not

only interesting per se, but could also be of biomedical relevance. As our method considerably extends the possibilities to derive iPSCs from primates, it could contribute towards leveraging the unique information generated during millions of years of primate evolution.

Methods

Experimental model and subject details. *Human urine samples.* Human urine samples from healthy volunteers were obtained with written informed consent and processed anonymously. This experimental procedure was ethically approved by the responsible committee on human experimentation (20-122, Ethikkommission LMU München). All experimental procedures were performed in accordance with relevant guidelines and regulations. Additional information on the samples is available in Supplementary Table S2.

Primate urine samples. Primate urine was collected at the Hellabrunn Zoo in Munich, Germany. Caretakers noted the time and most likely donor and took up available urine on the floor with a syringe, hence the collection procedure was fully non-invasive without any perturbation of the animals. Due to the collection procedure we do not know with certainty from which individual the samples were derived. Additional information on the samples can be found in Supplementary Table S4.

iPSC lines. iPSC lines were generated from human and non-human primate urinary cells. Reprogramming was done using two different techniques. Reprogramming using SeV (Thermo Fisher) was performed as suspension transduction as described before[10]. Episomal vectors were transfected using Lipofectamine 3000 (Thermo Fisher). iPSCs were cultured under feeder-free conditions on Geltrex (Thermo Fisher) -coated dishes in StemFit medium (Ajinomoto) supplemented with 100 ng/ml recombinant human basic FGF (Peprotech), 100 U/ml Penicillin and 100 µg/ml Streptomycin (Thermo Fisher) at 37 °C with 5% carbon dioxide. Cells were routinely subcultured using 0.5 mM EDTA. Whenever cells were dissociated into single cells using 0.5 × TrypLE Select (Thermo Fisher) or Accumax (Sigma Aldrich), the culture medium was supplemented with 10 µM Rho-associated kinase (ROCK) inhibitor Y27632 (BIOZOL) to prevent apoptosis.

Isolation of cells from urine samples. Urine from human volunteers was collected anonymously in sterile tubes. Usually a volume of 5–50 ml was obtained. Urine from NHPs was collected from the floor at Hellabrunn Zoo (Munich) by the zoo personnel, using a syringe without taking special precautions while collecting the samples. Samples were stored at 4 °C until processing for a maximum time span of 5 h. Isolation of primary cells was performed as previously described by Zhou et al. 2012. Briefly, the sample was centrifuged at 400×*g* for 10 min and washed with DPBS containing 100 U/ml Penicillin, 100 µg/ml Streptomycin (Thermo Fisher), 2.5 µg/ml Amphotericin (Sigma-Aldrich). Afterwards, the cells were resuspended in urinary primary medium consisting of 10% FBS (Life Technologies), 100 U/ml Penicillin, 100 µg/ml Streptomycin (Thermo Fisher), REGM supplement (ATCC) in DMEM/F12 (TH. Geyer) and seeded onto one gelatine coated well of a 12-well-plate. To avoid contamination stemming from the unsanitary sample collection, 100 µg/ml Normocure (Invivogen) was added to the cultures until the first passage. 1 ml of medium was added every day until day 5, where 4 ml of the medium was aspirated and 1 ml of renal epithelial and mesenchymal cell proliferation medium RE/MC proliferation medium was added. RE/MC consists of a 50/50 mixture of Renal Epithelial Cell Basal Medium (ATCC) plus the Renal Epithelial Cell Growth Kit (ATCC) and mesenchymal cell medium consisting of DMEM high glucose with 10% FBS (Life Technologies), 2 mM GlutaMAX-I (Thermo Fisher), 1 × NEAA (Thermo Fisher), 100 U/ml Penicillin, 100 µg/ml Streptomycin (Thermo Fisher), 5 ng/ml bFGF (PeproTech), 5 ng/ml PDGF-AB (PeproTech) and 5 ng/ml EGF (Miltenyi Biotec). Half of the medium was changed every day until the first colonies appeared. Subsequent medium changes were performed every second day. Passaging was conducted using 0.5 × TrypLE Select (Thermo Fisher). Typically 15×10^3 to 30×10^3 cells were seeded per well of a 12-well plate.

Single colony isolation from urine samples. For the UDSC single colony characterization experiment we seeded cells of 3 ml urine sample per well and chose the wells with only one colony for further characterization. The cells grew without further passage for two weeks (some colonies appeared only after one week) and were dissociated, counted and lysed in RLT Plus (Qiagen) as soon as they reached a sufficient size to be counted.

Generation of NHP iPSCs by Sendai virus vector infection. Infection of primary cells was performed with the CytoTune-iPS 2.0 Sendai Reprogramming Kit (Thermo Fisher) at a MOI of 5 using a modified protocol. Briefly, 7×10^5 urine derived cells were incubated in 100 µl of the CytoTune 2.0 SeV mixture containing three vector preparations: polycistronic Klf4–Oct3/4–Sox2, cMyc, and Klf4 for one hour at 37 °C. To control transduction efficiency 3.5×10^5 cells were infected with CytoTune-EmGFP SeV. Infected cells were seeded on Geltrex (Thermo Fisher) coated 12-well-plates, routinely 10×10^3 and 25×10^3 cells per well. Medium was replaced with fresh Renal epithelial and mesenchymal cell proliferation medium RE/MC (ATCC) every second day. On day 5, medium was changed to mTeSR1 (Stemcell Technologies), with subsequent medium changes every second day. After single colony picking, cells were cultured in StemFit (Ajinomoto) supplemented with 100 ng/ml recombinant human basic FGF (Peprotech), 100 U/ml Penicillin and 100 µg/ml Streptomycin (Thermo Fisher).

Immunostaining. Cells were fixed with 4% PFA, permeabilized with 0.3% Triton X-100, blocked with 5% FBS and incubated with the primary antibody diluted in 1% BSA and 0.3% Triton X-100 in PBS overnight at 4 °C. The following antibodies were used: Human alpha-Smooth Muscle Actin (R&D Systems, MAB1420), Human/Mouse alpha -Fetoprotein/AFP (R&D Systems, MAB1368), Nanog (R&D Systems, D73G4), Neuron-

specific beta-III Tubulin (R&D Systems, MAB1195), Oct-4 (NEB, D7O5Z), Sox2 (NEB, 4900S), SSEA4 (NEB, 4755), EpCAM (Fisher Scientific, 22 HCLC, TRA-1-60 (Miltenyi Biotec, REA157) and the isotype controls IgG2a (Thermo Fisher, eBM2a) and IgG1 (Thermo Fisher, P3.6.2.8.1). The next day, cells were washed and incubated with the secondary antibodies for one hour at room temperature. Alexa 488 rabbit (Thermo Fisher, A-11034) and Alexa 488 mouse (Thermo Fisher, A-21042) were used in a 1/500 dilution. Nuclei were counterstained using DAPI (Sigma Aldrich) at a concentration of 1 µg/ml.

Karyotyping. iPSCs at ~80% confluency were treated with 50 ng/ml colcemid (Thermo Fisher) for 2 h, harvested using TrypLE Select (Thermo Fisher) and treated with 75 mM KCL for 20 min at 37 °C. Subsequently, cells were fixed with methanol/acetic acid glacial (3:1) at −20 °C for 30 min. After two more washes of the fixed cell suspension in methanol/acetic (3:1) we followed standard protocols for the preparation of slides with differentially stained mitotic chromosome spreads using the G-banding technique. Between 33 and 60 metaphases were analyzed per cell line.

RT-PCR and PCR analyses. Total RNA was extracted from cells lysed with Trizol using the Direct-zol RNA Miniprep Plus Kit (Zymo Research, R2072). 1 µg of total RNA was reverse transcribed using Maxima H Minus Reverse Transcriptase (Thermo Fisher) and 5 µM random hexamer primers. Conditions were as follows: 10 min at 25 °C, 30 min at 50 °C and then 5 min at 85 °C. Quantitative polymerase chain reaction (qPCR) studies were conducted on 5 ng of reverse transcribed total RNA in duplicates using PowerUp SYBR Green master mix (Thermo Fisher) using primers specific for NANOG, OCT4, PAX6 and NESTIN. Each qPCR consisted of 2 min at 50 °C, 2 min at 95 °C followed by 40 cycles of 15 s at 95 °C, 15 s at 55 °C and 1 min at 72 °C. Cycle threshold was calculated by using default settings for the real-time sequence detection software (Thermo Fisher). For relative expression analysis the quantity of each sample was first determined using a standard curve and normalized to GAPDH and the average target gene expression (deltaCt/average target gene expression).

Genomic DNA for genotyping was extracted using DNeasy Blood and Tissue Kit (Qiagen). PCR analyses were performed using DreamTaq (Thermo Fisher). Primate primary cells were genotyped using primers that bind species-specific Alu insertions (adapted from[51]).

To confirm the transgene-free status of the iPSC lines, SeV specific primers were used described in CytoTune-iPS 2.0 Sendai Reprogramming Kit protocol (Thermo Fisher).

In vitro differentiation. For embryoid body formation iPSCs from one confluent 6-well were collected and subsequently cultured on a sterile bacterial dish in StemFit without bFGF. During the 8 days of suspension culture, medium was changed every second day. Subsequently, cells were seeded into six gelatin coated wells of a 6-well-plate. After 8 days of attached culture, immunocytochemistry was performed using α-fetoprotein (R&D Systems, MAB1368) as endoderm, α-smooth muscle actin (R&D Systems, MAB1420) as mesoderm and β-III tubulin (R&D Systems, MAB1195) as ectoderm marker.

For directed differentiation to neural stem cells (NSCs) cells were dissociated and 9×10^3 cells were plated into each well of a low attachment U-bottom 96-well-plate in 8GMK medium consisting of GMEM (Thermo Fisher), 8% KSR (Thermo Fisher), 5.5 ml 100 × NEAA (Thermo Fisher), 100 mM Sodium Pyruvate (Thermo Fisher), 50 mM 2-Mercaptoethanol (Thermo Fisher) supplemented with 500 nM A-83–01 (Sigma Aldrich), 100 nM LDN 193189 (Sigma Aldrich) and 30 µM Y27632 (biozol). Half medium change was performed at days 4, 8, 11. Neurospheres were lysed in TRI reagent (Sigma Aldrich) at day 7 and differentiation was verified using qRT PCR.

Bulk RNA-seq library preparation. In this study two bulk RNA-seq experiments were performed, one to validate the generated iPS cells and the corresponding primary cells and one to further characterize human UDSCs derived from single colonies. For the first experiment one colony per clone corresponding to ~2×10^4 cells and 2×10^3 primary cells of each individual was lysed in RLT Plus (Qiagen) and stored at − 80 °C until processing. While for the single colony urinary cell characterization experiment we used lysate from 500 to 1000 cells per colony. The prime-seq protocol, which is based on SCRB-seq[24–26], was used for library preparation[24–26]. The full protocol can be found on protocols.io (https://www.protocols.io/view/prime-seq-s9veh66). Even though prime-seq was used in both cases some minor differences between the two experiments exist. In particular in regards to the oligo dT primers that were used and the library preparation method as highlighted below. Briefly, proteins in the lysate were digested by Proteinase K (Ambion), RNA was cleaned up using SPRI beads (GE, 22%PEG). In order to remove isolated DNA, samples were treated with DNase I for 15 min at RT. cDNA was generated using oligo-dT primers containing well specific (sample specific) barcodes and unique molecular identifiers (UMIs). Unincorporated barcode primers were digested using Exonuclease I (New England Biolabs). cDNA was pre-amplified using KAPA HiFi HotStart polymerase (Roche) and pooled before library preparation. Sequencing libraries for the iPSC/primary cell experiment were constructed from 0.8 ng of preamplified cleaned up cDNA using the Nextera XT kit (Illumina). Sequencing libraries for the single colony experiment were constructed using NEBNext (New England Biolabs) according to the prime-seq protocol. In both cases 3′ ends were enriched with a custom P5 primer (P5NEXTPT5, IDT) and libraries were size-selected for fragments in the range of 300–800 bp.

Sequencing. Libraries were paired-end sequenced on an Illumina HiSeq 1500 instrument. Sixteen/twenty-eight bases were sequenced with the first read to obtain cellular and molecular barcodes and 50 bases were sequenced in the second read into the cDNA fragment.

Data processing and analysis. All raw fastq data were processed with zUMIs[52] using STAR 2.6.0a[53] to generate expression profiles for barcoded UMI data. All samples were mapped to the human genome (hg38).

Gene annotations were obtained from Ensembl (GRCh38.84). Samples were filtered based on number of genes and UMIs detected, and genes were filtered using HTS Filter. DESeq2[54] was used for normalization and variance stabilized transformed data was used for principal component analysis and hierarchical clustering.

Mitochondrial and rRNA reads were excluded and singleR (v1.4.0, https://bioconductor.org/packages/SingleR/) was used to classify the cells. SingleR was developed for unbiased cell type recognition of single cell RNA-seq data, however, here we applied the method to our bulk RNA seq dataset[28]. The 200 most variable genes were used in the 'de' option of SingleR to compare the obtained expression profiles to[55] as well as HPCA[27]. Based on the highest pairwise correlation between query and reference, cell types of the samples were assigned based on the most similar reference cell type.

We averaged and compared pairwise expression distances for different groups (Fig. 5d): the distances among iPSC clones within and between each species (N = 14 samples), the average of the distances between 1383D2 and the urinary derived human iPSCs (N = 9) and the average of the pairwise distance between and within individuals among iPSCs and species (within individuals: N = 6 (6 individuals with more than one clone), between individuals: N = 8).

Data availability

RNA-seq data generated here are available at GEO under accession number GSE155889.

Code availability

Code is available upon request.

Received: 21 August 2020; Accepted: 19 January 2021
Published online: 10 February 2021

References

1. Pecon-Slattery, J. Recent advances in primate phylogenomics. *Annu. Rev. Anim. Biosci.* **2**, 41–63 (2014).
2. Enard, W. Functional primate genomics-leveraging the medical potential. *J. Mol. Med.* **90**, 471–480 (2012).
3. Enard, W. The molecular basis of human brain evolution. *Curr. Biol.* **26**, R1109–R1117 (2016).
4. Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154–156 (1981).
5. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
6. Raab, S., Klingenstein, M., Liebau, S. & Linta, L. A comparative view on human somatic cell sources for iPSC generation. *Stem Cells Int.* **2014**, 768391 (2014).
7. Schlaeger, T. M. *et al.* A comparison of non-integrating reprogramming methods. *Nat. Biotechnol.* **33**, 58–63 (2015).
8. Wunderlich, S. *et al.* Primate iPS cells as tools for evolutionary analyses. *Stem Cell Res.* **12**, 622–629 (2014).
9. Gallego Romero, I. *et al.* A panel of induced pluripotent stem cells from chimpanzees: A resource for comparative functional genomics. *Elife* **4**, e07103 (2015).
10. Nakai, R. *et al.* Derivation of induced pluripotent stem cells in Japanese macaque (*Macaca fuscata*). *Sci. Rep.* **8**, 12187 (2018).
11. Stanton, M. M. *et al.* Prospects for the use of induced pluripotent stem cells (iPSC) in animal conservation and environmental protection. *Stem Cells Transl. Med.* https://doi.org/10.1002/sctm.18-0047 (2018).
12. Ezashi, T., Yuan, Y. & Roberts, R. M. Pluripotent stem cells from domesticated mammals. *Annu. Rev. Anim. Biosci.* **4**, 223–253 (2016).
13. Morizane, A. *et al.* MHC matching improves engraftment of iPSC-derived neurons in non-human primates. *Nat. Commun.* **8**, 385 (2017).
14. Fujie, Y. *et al.* New type of Sendai virus vector provides transgene-free iPS cells derived from chimpanzee blood. *PLoS ONE* **9**, e113052 (2014).
15. Zhou, T. *et al.* Generation of induced pluripotent stem cells from urine. *J. Am. Soc. Nephrol.* **22**, 1221–1228 (2011).
16. Zhou, T. *et al.* Generation of human induced pluripotent stem cells from urine samples. *Nat. Protoc.* **7**, 2080–2089 (2012).
17. Fusaki, N., Ban, H., Nishiyama, A., Saeki, K. & Hasegawa, M. Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* **85**, 348–362 (2009).
18. Bharadwaj, S. *et al.* Multipotential differentiation of human urine-derived stem cells: Potential for therapeutic applications in urology. *Stem Cells* **31**, 1840–1856 (2013).
19. Dörrenhaus, A. *et al.* Cultures of exfoliated epithelial cells from different locations of the human urinary tract and the renal tubular system. *Arch. Toxicol.* **74**, 618–626 (2000).
20. Lang, R. *et al.* Self-renewal and differentiation capacity of urine-derived stem cells after urine preservation for 24 hours. *PLoS ONE* **8**, e53980 (2013).
21. Okita, K. *et al.* A more efficient method to generate integration-free human iPS cells. *Nat. Methods* **8**, 409–412 (2011).
22. Okita, K. *et al.* An efficient nonviral method to generate integration-free human-induced pluripotent stem cells from cord blood and peripheral blood cells. *Stem Cells* **31**, 458–466 (2013).
23. Zhang, Y. *et al.* Urine derived cells are a potential source for urological tissue reconstruction. *J. Urol.* **180**, 2226–2233 (2008).
24. Bagnoli, J. W. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).
25. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*. https://doi.org/10.1101/003236 (2014).
26. Ziegenhain, C. *et al.* Comparative analysis of single-cell RNA sequencing methods: Molecular cell. *Mol. Cell* **65**, 631–643 (2017).
27. Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C. & Hume, D. A. An expression atlas of human primary cells: Inference of gene function from coexpression networks. *BMC Genomics* **14**, 632 (2013).
28. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
29. Yu, G. & He, Q.-Y. ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).
30. Nakagawa, M. *et al.* A novel efficient feeder-free culture system for the derivation of human induced pluripotent stem cells. *Sci. Rep.* **4**, 3594 (2014).
31. Sutherland, G. R. & Bain, A. D. Culture of cells from the urine of newborn children. *Nature* **239**, 231 (1972).
32. Bento, G. *et al.* Urine-derived stem cells: Applications in regenerative and predictive medicine. *Cells* **9**, 573 (2020).
33. Gaignerie, A. *et al.* Urine-derived cells provide a readily accessible cell type for feeder-free mRNA reprogramming. *Sci. Rep.* **8**, 14363 (2018).

34. Xue, Y. *et al.* Generating a non-integrating human induced pluripotent stem cell bank from urine-derived cells. *PLoS ONE* **8**, e70573 (2013).
35. Ernst, C. A roadmap for neurodevelopmental disease modeling for non-stem cell biologists. *Stem Cells Transl. Med.* **9**, 567–574 (2020).
36. Churko, J. M. *et al.* Transcriptomic and epigenomic differences in human induced pluripotent stem cells generated from six reprogramming methods. *Nat. Biomed. Eng.* **1**, 826–837 (2017).
37. Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147 (1998).
38. Pera, M. F., Reubinoff, B. & Trounson, A. Human embryonic stem cells. *J. Cell Sci.* **113**(Pt 1), 5–10 (2000).
39. Dick, E. C. Chimpanzee kidney tissue cultures for growth and isolation of viruses. *J. Bacteriol.* **86**, 573–576 (1963).
40. Nishimura, K. *et al.* Development of defective and persistent Sendai virus vector: A unique gene delivery/expression system ideal for cell reprogramming. *J. Biol. Chem.* **286**, 4760–4771 (2011).
41. Ben-Nun, I. F. *et al.* Induced pluripotent stem cells from highly endangered species. *Nat. Methods* **8**, 829–831 (2011).
42. Stauske, M. *et al.* Non-human primate iPSC generation, cultivation, and cardiac differentiation under chemically defined conditions. *Cells* **9**, 1349 (2020).
43. Navara, C. S., Chaudhari, S. & McCarrey, J. R. Optimization of culture conditions for the derivation and propagation of baboon (*Papioanubis*) induced pluripotent stem cells. *PLoS ONE* **13**, e0193195 (2018).
44. Hong, S. G. *et al.* Path to the clinic: Assessment of iPSC-based cell therapies in vivo in a nonhuman primate model. *Cell Rep.* **7**, 1298–1309 (2014).
45. Kanton, S. *et al.* Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).
46. Marchetto, M. C. N. *et al.* Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* **503**, 525–529 (2013).
47. Kelley, J. L. & Gilad, Y. Effective study design for comparative functional genomics. *Nat. Rev. Genet.* **21**, 385–386 (2020).
48. Housman, G. & Gilad, Y. Prime time for primate functional genomics. *Curr. Opin. Genet. Dev.* **62**, 1–7 (2020).
49. Alföldi, J. & Lindblad-Toh, K. Comparative genomics as a tool to understand evolution and disease. *Genome Res.* **23**, 1063–1068 (2013).
50. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
51. Herke, S. W. *et al.* A SINE-based dichotomous key for primate identification. *Gene* **390**, 39–51 (2007).
52. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs—A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, giy059 (2018).
53. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
54. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
55. Chu, L.-F. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* **17**, 173 (2016).

Acknowledgements
This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through LMUexcellent, SFB1243 (Subproject A14) and the Cyliax foundation. We thank Stefanie Färberböck for her expert technical assistance and enormous help in cell culture. We are grateful to Christine Gohl and the staff at the Zoo Hellabrunn for kindly collecting and providing the primate urine samples.

Author contributions
J.G., M.O. and W.E. conceived the study. J.G. and W.E. wrote the manuscript. J.G. established iPSC lines and conducted differentiation experiments. J.G. and J.R. performed EB differentiation and immunostaining experiments. J.G., L.E.W., A.J, J.W.B. and P.J. generated and analysed RNA-seq data. A.K. tested for virus absence in primate iPSCs. S.M. and J.G. performed karyotype analyses of iPSC lines.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-82883-0.

Correspondence and requests for materials should be addressed to M.O. or W.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2.3 Generation and characterization of three fibroblast-derived Rhesus Macaque induced pluripotent stem cells

Jessica Jocher, Fiona C. Edenhofer, **Philipp Janssen**, Stefan Müller, Dana C. Lopez-Parra, Johanna Geuder, Wolfgang Enard

"Generation and characterization of three fibroblast-derived Rhesus Macaque induced pluripotent stem cells"

Stem Cell Research 74, 103277 (2023).

doi: 10.1016/j.scr.2023.103277

Supplementary Information is freely available at the publisher's website:

https://www.sciencedirect.com/science/article/pii/S1873506123002635?via%3Dihub

Stem Cell Research 74 (2024) 103277

Contents lists available at ScienceDirect

Stem Cell Research

journal homepage: www.elsevier.com/locate/scr

Lab Resource: Animal Multiple Cell lines

Generation and characterization of three fibroblast-derived Rhesus Macaque induced pluripotent stem cell lines

Jessica Jocher [a], Fiona C. Edenhofer [a], Philipp Janssen [a], Stefan Müller [b], Dana C. Lopez-Parra [a], Johanna Geuder [a], Wolfgang Enard [a,*]

[a] Anthropology & Human Genomics, Faculty of Biology, Ludwig-Maximilians-Universität München, Großhaderner Straße 2, 82152 Martinsried, Germany
[b] Institute of Human Genomics, Munich University Hospital, Ludwig-Maximilians-Universität München, 80336 Munich, Germany

A B S T R A C T

Cross-species comparisons using pluripotent stem cells from primates are crucial to better understand human biology, disease, and evolution. An important primate model is the Rhesus macaque (*Macaca mulatta*), and we reprogrammed skin fibroblasts from a male individual to generate three induced pluripotent stem cell (iPSC) lines. These cells exhibit the typical ESC-like colony morphology, express common pluripotency markers, and can differentiate into cells of the three germ layers. All generated iPSC lines can be cultured under feeder-free conditions in commercially available medium and are therefore valuable resources for cross-species comparisons.

1. Resource Table

Unique stem cell lines identifier	MPC-MacMul-C00001 (83Ab1.1) MPC-MacMul-C00002 (83D1) MPC-MacMul-C00003 (87B1)
Alternative name(s) of stem cell lines	83Ab1.1 83D1 87B1
Institution	Faculty of Biology, Ludwig-Maximilians-Universität München
Contact information of distributor	Prof. Dr. Wolfgang Enard: enard@bio.lmu.de Jessica Jocher: jocher@bio.lmu.de
Type of cell lines	iPSCs
Origin	Rhesus Macaque (*Macaca Mulatta*)
Additional origin info	Sex: Male
Cell Source	iPSCs were derived from Rhesus macaque skin fibroblasts, kindly provided by the DPZ Göttingen
Clonality	Clonal
Method of reprogramming	Integration-free sendai virus based OSKM vectors (CytoTune-iPSC 2.0 Sendai Reprogramming Kit, Thermo Fisher Scientific) were used for reprogramming
Evidence of the reprogramming transgene loss (including genomic copy if applicable)	PCR analysis for transgene detection (negative)
Associated disease	N/A
Gene/locus	N/A
Date archived/stock date	November 2020
Cell line repository/bank	N/A
Ethical approval	The study was ethically approved by the Animal Welfare Committee at DPZ which is registered and authorized by the local and regional veterinary governmental authorities (Ref. no. 122910.3311900, PK 36674).

2. Resource utility

The three iPSC lines derived from one male Rhesus macaque skin sample can be used for cross-species comparisons investigating e.g., the molecular and cellular evolution of early primate development. Thereby, the three lines can help to assess clonal variation within one Rhesus macaque genetic background.

3. Resource details

Comparative analyses of human and non-human primates (NHP) can leverage unique information to understand evolutionary and developmental mechanisms and bridge the phylogenetic gap between humans and mice for translationally and biomedically relevant questions (Enard, 2012). Among NHPs, the Rhesus macaque (*Macaca mulatta*) is probably the most important model across biological disciplines and comprises 65 % of all NHP subjects used in the United States (Cooper et al., 2022). However, ethical, and practical limitations make it difficult to obtain

* Corresponding author.

https://doi.org/10.1016/j.scr.2023.103277
Received 6 September 2023; Received in revised form 30 October 2023; Accepted 6 December 2023
Available online 10 December 2023

comparable cells especially during development. Induced pluripotent stem cells (iPSCs) can be used to overcome these challenges (Juan et al., 2023).

Here, skin fibroblast obtained from Rhesus macaque were reprogrammed to iPSCs using Sendai viruses to introduce the Yamanaka factors OCT3/4, SOX2, KLF4 and C-MYC. Following reprogramming, colonies were picked, accustomed to feeder-free culture conditions and further characterized (Table 1). The established colonies exhibited a typical ESC-like morphology with defined borders, tight cellular packaging, and prominent nucleoli (Fig. 1A). A primate-specific SINE based PCR demonstrated that the iPSCs show the same Macaque-specific *ALU* element insertions as the parental skin fibroblasts, confirming that they were derived from the same species (Herke et al., 2007) (Supplementary Fig. S1A). In addition, single nucleotide polymorphisms (SNPs) were called from single-cell RNA-sequencing (scRNA-seq) data to profile the genotype of the cell lines. Around 4000 high quality SNPs with high coverage in all three clones were retrieved (Supplementary Fig. S1B). After some passages, karyotype analysis was performed and revealed no recurrent numerical or structural aberrations (Fig. 1B). In addition, one cell line was used for a detailed high resolution validation of numerical and structural chromosome integrity by FISH using human chromosome specific painting probes (Supplementary Fig. S1C). Immunofluorescence (IF) staining was performed, confirming the expression of the pluripotency associated proteins OCT3/4, SOX2 and NANOG, as well as the presence of the cell surface markers SSEA4 and EpCAM (Fig. 1D). Quantification of the IF staining for OCT3/4, SOX2 and NANOG showed that > 95 % of cells were positive for these pluripotency markers (Figure C). All three iPSCs were negative for mycoplasma DNA (Fig. 1E) and negative for the Sendai-based reprogramming vectors (Fig. 1F). Moreover, all iPSC lines had the ability to differentiate into cells of the three germ layers, confirmed by positive immunofluorescence staining of germ layer-specific markers. Endodermal cells were positively stained for alpha-fetoprotein (AFP), mesodermal cells expressed alpha-smooth muscle actin (SMA) and ectodermal cells displayed neuron-specific beta-III tubulin expression (Fig. 1G). Additionally, scRNA-seq of Embryoid bodies (EBs) confirmed their potential for trilineage differentiation and the expression of germ layer-specific marker genes (Supplementary Fig. S1D). In summary, these characteristics suggest the successful reprogramming of three lines to mycoplasma free, integration free and feeder-free iPSCs from Rhesus macaque (*Macaca mulatta*).

4. Materials and methods

4.1. Reprogramming of fibroblasts and iPSC maintenance

Fibroblasts were cultured on 0.2 % Gelatin-coated dishes in DMEM/F12 (Fisher Scientific) supplemented with 10 % FBS and 100 U/mL Penicillin and 100 µg/mL Streptomycin (Thermo Fisher Scientific) at 37 °C with 5 % CO_2. For reprogramming, a CytoTune™-iPS 2.0 Sendai Reprogramming Kit (Thermo Fisher Scientific) was used at a MOI of 5, using a modified protocol. Fibroblasts were incubated with the virus mix for 1 h at 37 °C in suspension, followed by seeding on feeder cell-coated wells. Cells were switched to mTesR1™ medium (STEMCELL Technologies) on day 5 after transduction. Emerging colonies were manually

Table 1
Characterization and validation.

Classification	Test	Result	Data
Morphology	Photography Bright field	Normal iPSC colony morphology	Fig. 1A Scale bar represents 500 µm
Phenotype	Qualitative analysis by immunocytochemistry	iPSCs were positively stained for OCT4, NANOG, SOX2, SSEA4 and EpCAM	Fig. 1D Scale bar represents 100 µm
	Quantitative analysis by immunocytochemistry counting	% total cells positive for pluripotency markers (mean ± SD): 83Ab1.1 OCT4: 97.6 % ± 1.1 % NANOG: 94.4 % ± 2.8 % SOX2: 98.3 % ± 0.7 % 83D1 OCT4: 97.8 % ± 2 % NANOG: 98.4 % ± 0.3 % SOX2: 98.7 % ± 0.5 % 87B1 OCT4: 98.7 % ± 0.2 % NANOG: 98.7 % ± 0.9 % SOX2: 98.9 % ± 0.6 %	Fig. 1C
Genotype	Karyotype (G-banding and FISH)	3x inconspicuous male karyotype, 42,XYNo recurrent numeric or structural aberrations, after G-banding analysis of 46 to 48 cells per cell line with up to approximately 400 bphs (bands per haploid set)	Fig. 1B and Supplementary Fig. S1C
Identity	SINE-based genotyping PCR SNP analysis	DNA profiling performed, matched between iPSCs and parental fibroblasts Variant calling performed resulting in 4000 high quality SNPs	Supplementary Fig. S1A Submitted in archive with journal Summary: Supplementary Fig. S1B
Mutation analysis (IF APPLICABLE)	N/A N/A		
Microbiology and virology	Mycoplasma Sendai virus	Mycoplasma testing by PCR: negative PCR analysis for Sendai virus presence: negative	Fig. 1E Fig. 1F
Differentiation potential	Embryoid body formation - IF staining	iPSCs are capable of differentiating into the three germ layers. Mesoderm: Smooth muscle actin (SMA) Endoderm: α-feto protein (AFP) Ectoderm: β-III Tubulin	Fig. 1G Scale bar represents 100 µm
	Embryoid body formation - scRNA-seq	Expression of multiple cell type and germ layer specific marker genes	Supplementary Fig. S1D
Donor screening (OPTIONAL)	N/A		
Genotype additional info (OPTIONAL)	N/A N/A		

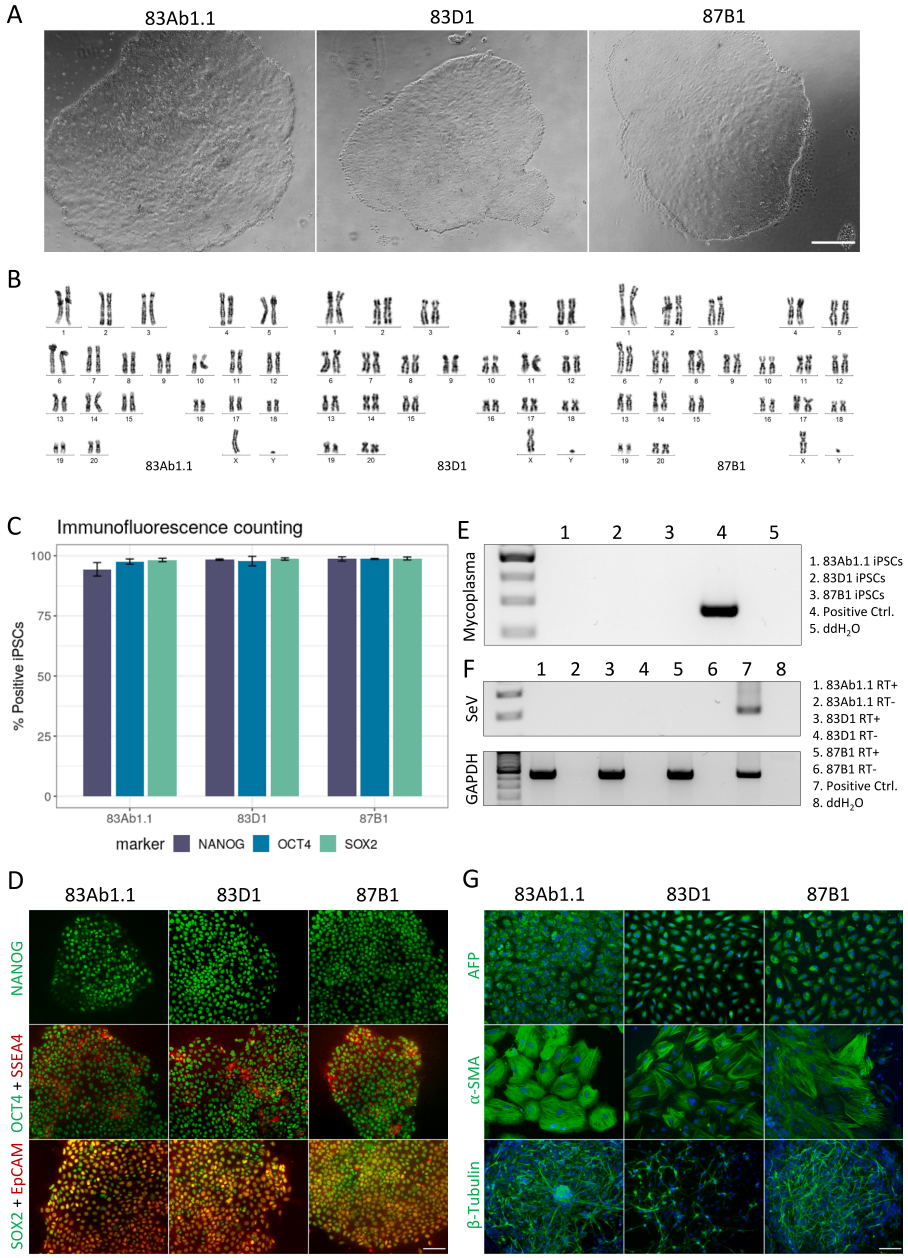


Fig. 1. Characterization of the three Rhesus macaque iPSC lines. (A) Phase contrast microscopy images of iPSC colonies. Scale bar represents 500 µm. (B) Karyotype analysis. (C) Immunofluorescence counting results for NANOG, OCT4 and SOX2. (D) Immunofluorescence staining for pluripotency markers. Scale bar represents 100 µm. (E) Mycoplasma test. (F) PCR for Sendai-based reprogramming vectors. (G) Immunofluorescence staining for germ layer-specific markers. Scale bar represents 100 µm.

picked on feeder cells and cultured in StemFit® Basic02 (Ajinomoto) supplemented with 100 ng/mL bFGF (Peprotech) and 100 U/mL Penicillin and 100 µg/mL Streptomycin. For generating feeder-free iPSCs, colonies were split using 0.5 mM EDTA on 1 % Geltrex™ (Thermo Fisher Scientific) -coated wells in feeder-conditioned StemFit. The ratio of feeder-conditioned to normal StemFit was reduced by 25 % after every second passage, until iPSCs could be cultured under feeder-free conditions. iPSCs were passaged using 0.5 mM EDTA at a ratio of 1:10–1:50 every 5 days, with medium changes every other day.

4.2. Immunocytochemistry

Attached cells (passage 15–20) were fixed for 15 min with 4 % PFA, permeabilized with 0.3 % Triton X-100 (Sigma Aldrich) and blocked with 5 % FBS for 30 min. Cells were incubated with primary antibodies (Table 2) diluted in staining buffer (PBS containing 1 % BSA and 0.3 % Triton X-100) overnight at 4 °C. The next day, cells were washed with PBS and incubated with secondary antibodies (Table 2) diluted in staining buffer for 1 h at room temperature. Nuclei were counterstained with 1 µg/mL DAPI. Positively-stained cells were quantified using the ImageJ software with the Cell Counter plugin.

4.3. Embryoid body formation

iPSCs at passage 15–20 were dissociated into clumps and cultured in sterile bacterial dishes containing StemFit Basic02 w/o bFGF at 37 °C with 5 % CO_2. A medium change was performed every other day during the first 8 days of floating culture. Afterwards, EBs were seeded into 6-wells coated with 0.2 % Gelatin for 8 days of attached culture. On day 16, differentiated cells were analyzed with specific antibodies for mesoderm, endoderm, and ectoderm (Table 2) using immunocytochemistry. In addition, cells were also sampled for scRNA-seq on day 16. Briefly, EBs were dissociated using Accumax, and sequencing libraries were generated using the 10x Genomics Chromium Next GEM Single Cell 3'Kit V3.1 workflow. Cluster analysis was performed in R using the package Seurat v5 and clusters were assigned to germ layers based on the expression of known marker genes.

4.4. Karyotyping

For Metaphase preparation, cells (passage 16–23) at 80 % confluency were incubated with 0.1 mg/mL Colcemid (Gibco) for 13 h and harvested using Accumax (Sigma Aldrich). Cells were treated with hypotonic Na-Citrate/NaCl for 35 min at 37 °C, followed by a subsequent fixation with methanol/acetic acid glacial (3:1) for 20 min at -20 °C. After pelleting, cells were washed twice with methanol/acetic acid as stated above. Differentially stained mitotic chromosome spreads were prepared using the G-banding technique and fluorescence in situ hybridization (FISH) was performed using human chromosome specific painting probes, following standard procedures.

4.5. Mycoplasma testing

The medium of a confluent 6-well with iPSCs at passage 15–20 was collected and pelleted, followed by resuspension in 100 µL PBS. After a 5 min incubation at 95 °C, 1 µL was used for a screening PCR with specific primers for the Mycoplasma 16S rRNA (Table 2).

4.6. Genotyping PCR

Total gDNA was isolated from cell pellets using the DirectPCR Lysis Reagent (VWR) supplemented with 20 mg/mL Proteinase K (Life Technologies), and a PCR (36 cycles) was performed with primers for the primate-specific *Alu* SINE (Table 2).

4.7. Variant calling

10x scRNA-seq data of day 16 EBs were used to call SNPs against the reference genome rheMac10 using GATK (Genome Analysis Tool Kit). High quality, biallelic SNPs were retained by joint genotyping of all three clones followed by quality filtering of the variants for high coverage (DP > 99) and quality by depth (QD > 2).

4.8. SeV detection

Total RNA was isolated from iPSCs at passage 10–15 using the Direct-zol RNA Microprep Kit (Zymo Research). After cDNA synthesis using the Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific), the cDNA was used to perform a PCR (36 cycles) with specific primers for SeV and GAPDH (Table 2).

Declaration of competing interest

The authors declare the following financial interests/personal

Table 2
Reagents details.

Antibodies used for immunocytochemistry				
	Antibody	Dilution	Company Cat #	RRID
Pluripotency Markers	Rabbit anti-OCT4	1:400	Cell Signaling Technology, Cat# 2750S	RRID: AB_823583
	Mouse anti-SOX2	1:400	Cell Signaling Technology, Cat# 4900S	RRID: AB_10560516
	Rabbit anti-Nanog	1:400	Cell Signaling Technology, Cat# 4903S	RRID: AB_10559205
	Mouse anti-SSEA4	1:500	NEB, Cat# 4755S	RRID: AB_1264259
	Rabbit anti-EpCAM	1:500	Thermo Fisher Scientific, Cat# 710524	RRID: AB_2532731
Differentiation Markers	Mouse anti- α-Smooth Muscle Actin	1:100	R&D Systems, Cat# MAB1420	RRID: AB_262054
	Mouse anti- Neuron-specific beta-III Tubulin	1:100	R&D Systems, Cat# MAB1195	RRID: AB_357520
	Mouse anti-alpha Fetoprotein	1:100	R&D Systems, Cat# MAB1368	RRID: AB_357658
Secondary Antibodies	Alexa Fluor 488 donkey anti-mouse IgG (H + L)	1:500	Thermo Fisher Scientific, Cat# A-21202	RRID: AB_141607
	Alexa Fluor 594 donkey anti-rabbit IgG (H + L)	1:500	Thermo Fisher Scientific, Cat# A-21207	RRID: AB_141637

Primers			
	Target	**Size of band**	**Forward/Reverse primer (5′-3′)**
Reprogramming factor clearance	Sendai Virus	180 bp	GGATCACTAGGTGATATCGAGC/ ACCAGACAAGAGTTTAAGAGATATGTATC
	GAPDH (housekeeping gene)	450 bp	ACCACAGTCCATGCCATCAC/TCCACCACCCTGTTGCTGTA
Mycoplasma testing	Mycoplasma 16S	270 bp	TGCACCATCTGTCACTCTGTTAACCTC/ GGGAGCAAACAGGATTAGATACCCT
Genotyping PCR	*Alu* (primate-specific SINE)	548 bp	CTCTCAGCTCCCTGTTTCTGTT/CATGGACATCAGACTAGCCACT

relationships which may be considered as potential competing interests: Wolfgang Enard reports financial support, article publishing charges, and travel were provided by German Research Foundation.

Acknowledgements

This work was supported by DFG EN 1093/5-1 (project number 458247426). We are grateful to Kerstin Mätz-Rensing and the staff at the DPZ for kindly providing the primary material. We thank Dr. Mari Ohnuki and Rudolf Hamburg, for picking up the material and for isolating the primary cells. Furthermore, we thank Vanessa Baltruschat for her substantial technical support in the lab.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scr.2023.103277.

References

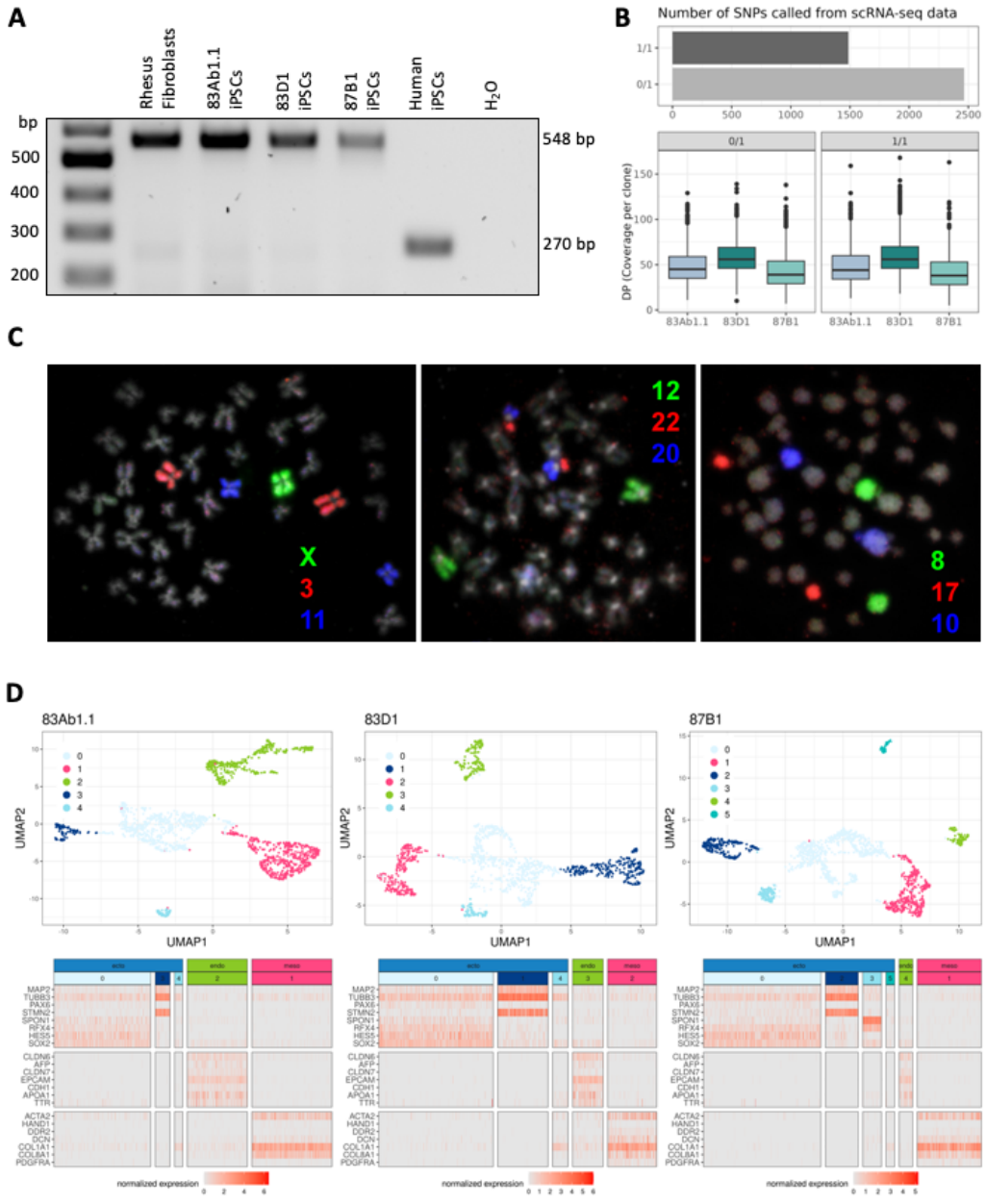
Cooper, E., Brent, L., Snyder-Mackler, N., Singh, M., Sengupta, A., Khatiwada, S., Malaivijitnond, S., Hai, Z., Higham, J., 2022. The Rhesus macaque as a success story of the anthropocene. eLive. https://doi.org/10.7554/elife.78169.

Enard, W., 2012. Functional primate genomics - leveraging the medical potential. J. Mol. Med. https://doi.org/10.1007/s00109-012-0901-4.

Herke, S., Xing, J., Ray, D., Zommerman, J., Cordaux, R., Batzer, M., 2007. A SINE-based dichotomous key for primate identification. Gene. https://doi.org/10.1016/j.gene.2006.08.015.

Juan, D., Santpere, G., Kelley, J., Cornejo, O.E., Marques-Bonet, T., 2023. Current advances in primate genomics: novel approaches for understanding evolution and disease. Nat. Rev. Genet. https://doi.org/10.1038/s41576-022-00554-w.

Supplementary Figure



2.4 Generation and characterization of two Vervet monkey induced pluripotent stem cell lines derived from fibroblasts

Jessica Jocher, Fiona C. Edenhofer, **Philipp Janssen**, Stefan Müller, Eva Briem, Johanna Geuder, Wolfgang Enard

"Generation and characterization of two Vervet monkey induced pluripotent stem cell lines derived from fibroblasts"

Stem Cell Research 75, 103315 (2024).

doi: 10.1016/j.scr.2024.103315

Supplementary Information is freely available at the publisher's website:

https://www.sciencedirect.com/science/article/pii/S1873506124000138?via%3Dihub

Stem Cell Research 75 (2024) 103315

Contents lists available at ScienceDirect

Stem Cell Research

journal homepage: www.elsevier.com/locate/scr

Generation and characterization of two Vervet monkey induced pluripotent stem cell lines derived from fibroblasts

Jessica Jocher [a], Fiona C. Edenhofer [a], Stefan Müller [b], Philipp Janssen [a], Eva Briem [a], Johanna Geuder [a], Wolfgang Enard [a,*]

[a] Anthropology & Human Genomics, Faculty of Biology, Ludwig-Maximilians-Universität München, Großhaderner Straße 2, 82152 Martinsried, Germany
[b] Institute of Human Genetics, Munich University Hospital, Ludwig-Maximilians-Universität München, 80336 Munich, Germany

A B S T R A C T

Cross-species comparisons using pluripotent stem cells from primates are crucial to better understand human biology, disease, and evolution. The Vervet monkey (*Chlorocebus aethiops sabaeus*) serves as an important primate model for such studies, and therefore we reprogrammed skin fibroblasts derived from a male and a female individual, resulting in two induced pluripotent stem cell lines (iPSCs). These iPSCs display the characteristic ESC-like colony morphology, express key pluripotency markers, and possess the ability to differentiate into cells representing all three germ layers. Importantly, both generated cell lines can be maintained in feeder-free culture conditions using commercially available medium.

1. Resource Table

Unique stem cell lines identifier	MPC-ChlSab-C00001 (76A3) MPC-ChlSab-C00002 (80B1)
Alternative name(s) of stem cell lines	76A3 80B1
Institution	Faculty of Biology, Ludwig-Maximilians-Universität München
Contact information of distributor	Prof. Dr. Wolfgang Enard: enard@bio.lmu.de Jessica Jocher: jocher@bio.lmu.de
Type of cell lines	iPSCs
Origin	Vervet monkey (*Chlorocebus aethiops sabaeus*)
Additional origin info	Sex: female (76A3) and male (80B1)
Cell Source	iPSCs were derived from fibroblasts established from skin biopsies collected at the Vervet Research Colony and kindly provided by the UCLA.
Clonality	Clonal
Method of reprogramming	Integration-free sendai virus based OSKM vectors (CytoTune-iPSC 2.0 Sendai Reprogramming Kit, Thermo Fisher Scientific) were used for reprogramming
Evidence of the reprogramming transgene loss	PCR analysis for transgene detection (negative)
Associated disease	N/A
Gene/locus	N/A
Date archived/stock date	November 2020
Cell line repository/bank	N/A
Unique stem cell lines identifier	MPC-ChlSab-C00001 (76A3) MPC-ChlSab-C00002 (80B1)
Ethical approval	Fibroblast sample collection was approved by the UCLA and VA Institutional Animal Care and Use Committees. Fibroblast import was approved by CITES (permit number: 18US12381D/9)

2. Resource utility

The two iPSC lines derived from fibroblasts of a male and a female Vervet monkey, provide a valuable resource for cross-species comparisons, e.g., enabling investigations of molecular and cellular processes during early primate development. Furthermore, the two cell lines can help to assess intra-species variation within the Vervet genetic background and allow to investigate sex-related genetic factors.

3. Resource details

To gain insights into evolutionary and developmental mechanisms, as well as to bridge the phylogenetic gap between humans and mice, comparative analyses of human and non-human primates (NHP) can provide valuable and unique information (Enard, 2012). Among NHPs, the Vervet monkey (also called African green monkey) is, next to Rhesus macaques, the most commonly investigated NHP in biomedical

* Corresponding author.
E-mail address: enard@bio.lmu.de (W. Enard).

https://doi.org/10.1016/j.scr.2024.103315
Received 6 September 2023; Received in revised form 21 December 2023; Accepted 16 January 2024
Available online 17 January 2024

2.4 Generation and characterization of two Vervet monkey induced pluripotent stem cell lines derived from fibroblasts 95

research. One reason for this is that they are a natural host to the Simian Immunodeficiency Virus (SIV) and can be used to study adaptations to lentiviral infections. In contrast to African Vervets, the Caribbean Vervets (*Chlorocebus aethiops sabaeus*) also used here, are SIV free which allows safe and controlled studies on SIV infections and has made them a valuable source for NHP genetics (Jasinska, 2013). Generating induced pluripotent stem cells (iPSCs) from Vervets links this important NHP model to stem cell biology and comparative primate genomics (Juan et al., 2023).

Here, Vervet monkey skin fibroblasts were reprogrammed to iPSCs using a commercially available Sendai virus kit to introduce OCT3/4, SOX2, KLF4 and C-MYC into the cells. Emerging colonies were picked, gradually transferred to feeder-free culture conditions, and further characterized (Table 1). The two derived clones exhibit the typical ESC-like morphology with tight cellular packaging, prominent nucleoli, and defined colony borders (Fig. 1A). Immunofluorescence (IF) staining confirmed the expression of the pluripotency associated proteins OCT3/4 and SOX2, as well as the presence of the cell surface markers TRA-1-60, SSEA4 and EpCAM (Fig. 1B). Quantification of the IF staining revealed that > 95 % of cells are expressing both pluripotency markers OCT3/4 and SOX2 (Fig. 1C). A primate-specific SINE based PCR confirmed the presence of Vervet-specific *ALU* element insertions in the iPSCs as well as the parental skin fibroblasts, confirming their derivation from the same primate species (Herke et al., 2007) (Supplementary Fig. S1A). In addition, single nucleotide polymorphisms (SNPs) were called from bulk RNA-sequencing (bulk RNA-seq) data to profile the genotype of the cell lines. Around 5100 and 4300 high quality SNPs with high coverage were retrieved for the cell lines 76A3 and 80B1, respectively (Supplementary Fig. S1B,C). All iPSCs were negative for mycoplasma contamination (Fig. 1E) and negative for Sendai-based reprogramming vectors (Fig. 1F). Furthermore, karyotype analysis was performed, revealing no recurrent numerical or structural aberrations in the two iPSC lines (Fig. 1D). In addition, a detailed high resolution validation of numerical and structural chromosome integrity was performed by FISH using human chromosome specific painting probes on the 76A3 line (Supplementary Fig. S1D). To assess their differentiation capacity, an in vitro differentiation to embryoid bodies (EBs) was conducted and stainings for alpha-fetoprotein (AFP) and SOX17 were used to verify endodermal differentiation, alpha-smooth muscle actin (SMA) and pro-collagen-1 alpha-1 (COL1A1) for mesodermal differentiation, and neuron-specific beta-III tubulin and PAX6 for ectodermal differentiation (Fig. 1G). In summary, these characteristics indicate the successful establishment of two feeder-free iPSC lines from Vervet monkey (*Chlorocebus aethiops sabaeus*).

4. Materials and methods

4.1. Reprogramming of fibroblasts and iPSC maintenance

Fibroblasts were cultured on 0.2 % Gelatin coated dishes in DMEM/F12 (Fisher Scientific) supplemented with 10 % FBS and 100 U/mL Penicillin and 100 µg/mL Streptomycin (Thermo Fisher Scientific) at 37 °C and 5 % CO_2. The CytoTune™-iPS 2.0 Sendai Reprogramming Kit (Thermo Fisher Scientific) was used for reprogramming following a modified protocol. Briefly, fibroblasts were incubated in suspension with the virus mix at a MOI of 5 for 1 h at 37 °C and then seeded onto a feeder layer. On day 5, medium was switched to mTesR1™ (STEMCELL Technologies). Appearing colonies were manually picked onto feeder cells and cultured in StemFit® Basic02 (Ajinomoto) supplemented with 100 ng/mL bFGF (Peprotech) and 100 U/mL Penicillin and 100 µg/mL Streptomycin. For a feeder free culture, cells were passaged using 0.5 mM EDTA on 1 % Geltrex™ (Thermo Fisher Scientific) coated wells in

Table 1
Characterization and validation.

Classification	Test	Result	Data
Morphology	Photography phase contrast	Normal colony morphology	Fig. 1A Scale bar represents 500 µm
Phenotype	Qualitative analysis by immunocytochemistry	iPSCs were positively stained for OCT4, SOX2, TRA-1–60, SSEA4 and EpCAM	Fig. 1B Scale bar represents 100 µm
	Quantitative analysis by immunocytochemistry counting	% total cells positive for pluripotency markers (mean ± SD): 76A3 OCT4: 97.2 % ± 1.5 % (2,436 cells counted) SOX2: 98.6 % ± 1.2 % (1,693 cells counted) 80B1 OCT4: 96 % ± 1.9 % (2,128 cells counted) SOX2: 97.5 % ± 2.4 % (1,508 cells counted)	Fig. 1C
Genotype	Karyotype (G-banding and FISH)	76A3: inconspicuous female karyotype, 60,XX 80B1: inconspicuous male karyotype, 60,XY	Fig. 1D and Supplementary Fig. S1D
Identity	SINE-based genotyping PCR	DNA profiling performed, matched between iPSCs and parental fibroblasts	Supplementary Fig. S1A
	SNP analysis	Variant calling performed resulting in 5100 (76A3) and 4300 (80B1) high quality SNPs	Submitted in archive with journal Summary: Supplementary Fig. S1B,C
Mutation analysis (IF APPLICABLE)	N/A N/A		
Microbiology and virology	Mycoplasma	Mycoplasma testing by PCR: negative	Fig. 1E
	Sendai virus	PCR analysis for Sendai virus presence: negative	Fig. 1F
Differentiation potential	Embryoid body formation	iPSCs are capable of differentiating into the three germ layers. Mesoderm: Smooth muscle actin (SMA) and COL1A1 Endoderm: α-feto protein (AFP) and SOX17 Ectoderm: β-III tubulin and PAX6	Fig. 1G Scale bar represents 100 µm
Donor screening (OPTIONAL)	N/A		
Genotype additional info (OPTIONAL)	N/A N/A		

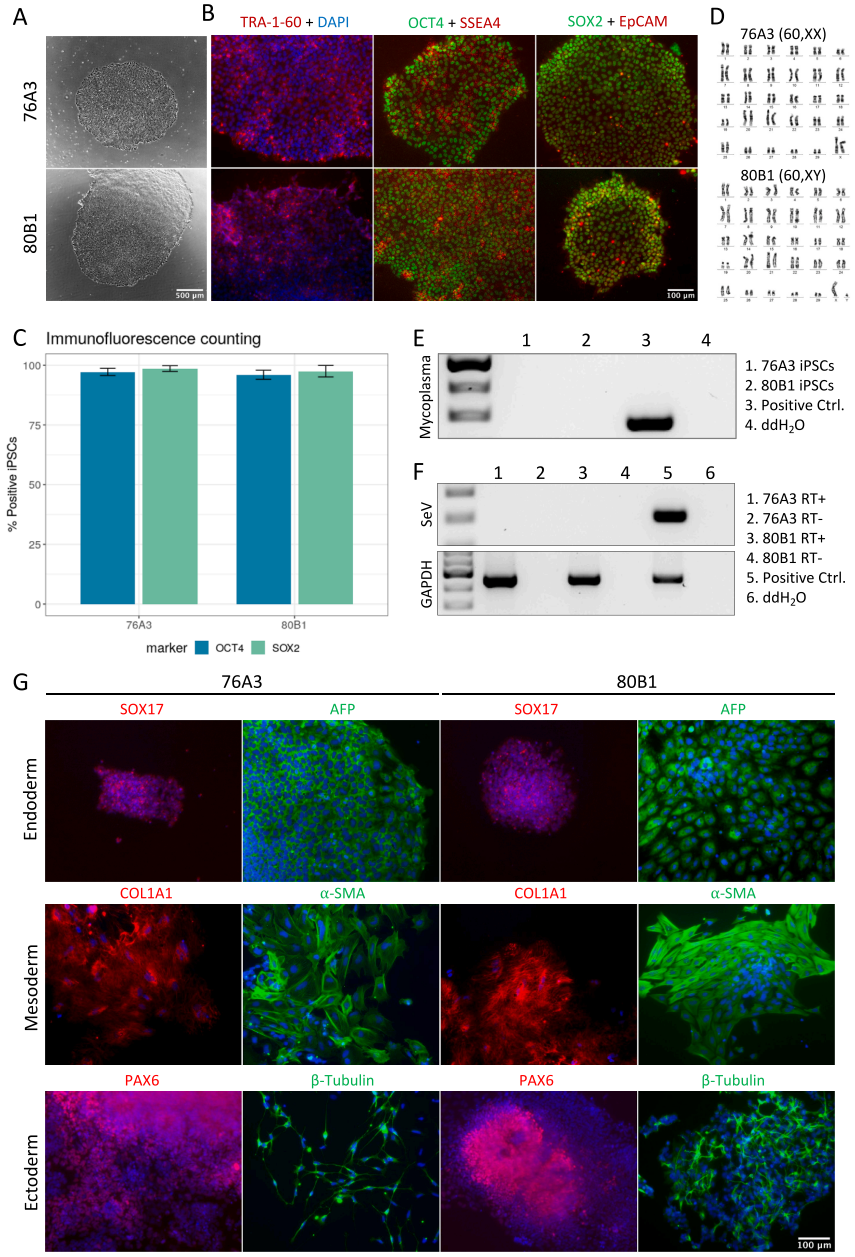


Fig. 1. Characterization of two Vervet monkey iPSC lines. (A) Phase contrast microscopy images of iPSC colonies. Scale bar represents 500 µm. (B) Immunofluorescence staining for pluripotency markers. Scale bar represents 100 µm. (C) Immunofluorescence couting results for OCT4 and SOX2. (D) Karyotype analysis. (E) Mycoplasma test. (F) PCR for Sendai-based reprogramming vectors. (G) Immunofluorescence staining for germ layer-specific markers. Scale bar represents 100 µm.

feeder-conditioned StemFit. The ratio of feeder-conditioned to normal StemFit was reduced by 25 % every other passage until iPSCs could be cultured under feeder-free conditions. iPSCs were passaged using 0.5 mM EDTA at a ratio of 1:10–1:40 every 5 days, changing the medium every other day.

4.2. Immunocytochemistry

Attached cells (passage 19–25) were fixed with 4 % PFA for 15 min at RT, permeabilized with 0.3 % Triton X-100 (Sigma Aldrich) and blocked for 30 min with 5 % FBS. Cells were incubated with primary antibodies (Table 2) diluted in staining buffer (PBS containing 1 % BSA and 0.3 % Triton X-100) overnight at 4 °C. Thereafter, cells were washed with PBS and incubated with secondary antibodies (Table 2) diluted in staining buffer for 1 h at RT. Nuclei were counterstained using DAPI at a concentration of 1 μg/mL. To obtain proportions of positively-stained cells, images of the fluorescence staining and the corresponding DAPI staining were counted using the Cell Counted plugin in ImageJ. Between 1,508 and 2,436 cells were counted for each marker and percentages were calculated based on the number of positively-stained cells divided by the number of DAPI stained nuclei. The standard deviation was calculated based on the difference between the counted images.

4.3. Embryoid body formation

One 6-well of iPSCs at passage 19–25 was dissociated into clumps, transferred to a sterile bacterial dish containing StemFit w/o bFGF and cultured at 37 °C with 5 % CO_2. During the first 8 days of floating culture, the medium was changed every other day. Then, EBs were seeded into 0.2 % Gelatin-coated 6-wells for 8 days of adherent culture. On day 16, the cells were stained with antibodies for mesoderm, endoderm, and ectoderm (Table 2) as stated above.

4.4. Karyotyping

Cells (passage 15–20) at 80 % confluency were incubated with 0.1 mg/mL Colcemid (Gibco) for 14 h and harvested using Accumax™ (Sigma Aldrich). Cells were treated with hypotonic Na-Citrate / NaCl for 35 min at 37 °C and then fixed with methanol / acetic acid glacial (3:1) for 20 min at -20 °C. After pelleting, cells were washed twice with methanol/acetic acid before conducting standard protocols for chromosome preparation, G-banding, and fluorescence in situ hybridization (FISH) using human chromosome specific painting probes.

4.5. Mycoplasma testing

The medium of a confluent 6-well with iPSCs at passage 15–25 was collected, pelleted, and resuspended in 100 μL PBS. After incubation at 95 °C for 5 min, 1 μL was used for a screening PCR with specific primers for the Mycoplasma 16S rRNA (Table 2).

4.6. Genotyping PCR

gDNA was isolated using the DirectPCR Lysis Reagent (VWR) supplemented with 20 mg/mL Proteinase K (Life Technologies), and a PCR (36 cycles) was carried out with primers for the primate-specific *Alu* SINE (Table 2).

4.7. SeV detection

Total RNA was isolated from iPSCs at passage 15–18 using the Direct-zol RNA Microprep Kit (Zymo Research) according to the manufacturer's instructions. After reverse transcription using the Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific), the cDNA was used to perform a PCR (36 cycles) with specific primers for SeV. The housekeeping gene GAPDH was used as positive control (Table 2).

4.8. Bulk RNA-sequencing and variant calling

iPSCs of both individuals were dissociated using Accumax, sampled in three biological replicates each and bulk RNA-seq libraries were generated using the Prime-seq workflow (https://www.protocols.io/view/prime-seq-81wgb1pw3vpk/v2). Bulk RNA-seq data of iPSCs were used to call SNPs against the reference genome chlSab2 using GATK (Genome Analysis Tool Kit). High quality, biallelic SNPs were retained by quality filtering of the variants for high coverage (DP > 49) and quality by depth (QD > 2).

Table 2
Reagents details.

	Antibodies used for immunocytochemistry/flow-cytometry			
	Antibody	Dilution	Company Cat #	RRID
Pluripotency Markers	Rabbit anti-OCT4	1:400	Cell Signaling Technology, Cat# 2750S	RRID: AB_823583
	Mouse anti-SOX2	1:400	Cell Signaling Technology, Cat# 4900S	RRID: AB_10560516
	Mouse anti-SSEA4	1:500	NEB, Cat# 4755S	RRID: AB_1264259
	Rabbit anti-EpCAM	1:500	Thermo Fisher Scientific, Cat# 710524	RRID: AB_2532731
	Mouse anti-TRA-1–60	1:100	Stem Cell Technologies, Cat# 60064	RRID: AB_2686905
Differentiation Markers	Mouse anti-α-Smooth Muscle Actin	1:100	R&D Systems, Cat# MAB1420	RRID: AB_262054
	Sheep anti-COL1A1	1:200	R&D Systems, Cat# AF6220	RRID: AB_10891543
	Mouse anti- Neuron-specific beta-III Tubulin	1:100	R&D Systems, Cat# MAB1195	RRID: AB_357520
	Rabbit anti-PAX6	1:100	Thermo Fisher Scientific, Cat# 42–6600	RRID: AB_2533534
	Mouse anti-alpha Fetoprotein	1:100	R&D Systems, Cat# MAB1368	RRID: AB_357658
	Rabbit anti-SOX17	1:500	Bio-Techne, Cat# NBP2-24568	RRID: AB_3075468
Secondary Antibodies	Alexa Fluor 488 donkey anti-mouse IgG (H + L)	1:500	Thermo Fisher Scientific, Cat# A-21202	RRID: AB_141607
	Alexa Fluor 594 donkey anti-rabbit IgG (H + L)	1:500	Thermo Fisher Scientific, Cat# A-21207	RRID: AB_141637
	Alexa Fluor 488 donkey anti-sheep (H + L)	1:500	Thermo Fisher Scientific, Cat# A-11015	RRID: AB_2534082
	Primers			
	Target	**Size of band**	**Forward/Reverse primer (5′-3′)**	
Reprogramming factor clearance	Sendai Virus	180 bp	GGATCACTAGGTGATATCGAGC / ACCAGACAAGAGTTTAAGAGATATGTATC	
	GAPDH (housekeeping gene)	450 bp	ACCACAGTCCATGCCATCAC / TCCACCACCCTGTTGCTGTA	
Mycoplasma testing	Mycoplasma 16S	270 bp	TGCACCATCTGTCACTCTGTTAACCTC / GGGAGCAAACAGGATTAGATACCCT	
Genotyping PCR	*Alu* (primate-specific SINE)	680 bp	CACAAAATACTAAAGGACTGTTAAAGG / CACAAAATACTAAAGGACTGTTAAAGG	

CRediT authorship contribution statement

Jessica Jocher: Conceptualizing, Investigation, Methodology, Visualization, Writing – original draft, Writing - review & editing. **Fiona C. Edenhofer:** Investigation, Visualization. **Stefan Müller:** Investigation, Methodology, Visualization. **Philipp Janssen:** Data curation, Formal analysis, Visualization, Writing – review & editing. **Eva Briem:** Investigation. **Johanna Geuder:** Methodology. **Wolfgang Enard:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Wolfgang Enard reports financial support, article publishing charges, and travel were provided by German Research Foundation.

Acknowledgements

This work was supported by DFG EN 1093/5-1 (project number 458247426). We thank Stephanie Färberböck and Vanessa Baltruschat for her technical assistance and substantial help in cell culture. We are grateful to Dr. Anna Jasinska from University of California, Los Angeles for providing primary fibroblasts and for supporting the CITES permit procedure. We are grateful to Dr. Nelson Freimer from University of California, Los Angeles, supported by his grants R01RR016300/0D010980 and to Dr. Matthew Jorgensen, Wake Forest University, supported by his grants NIH (P40RR019963/0D010965 (VRC) and P40-0D010965) for providing and enabling the collection of primary fibroblasts.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scr.2024.103315.

References

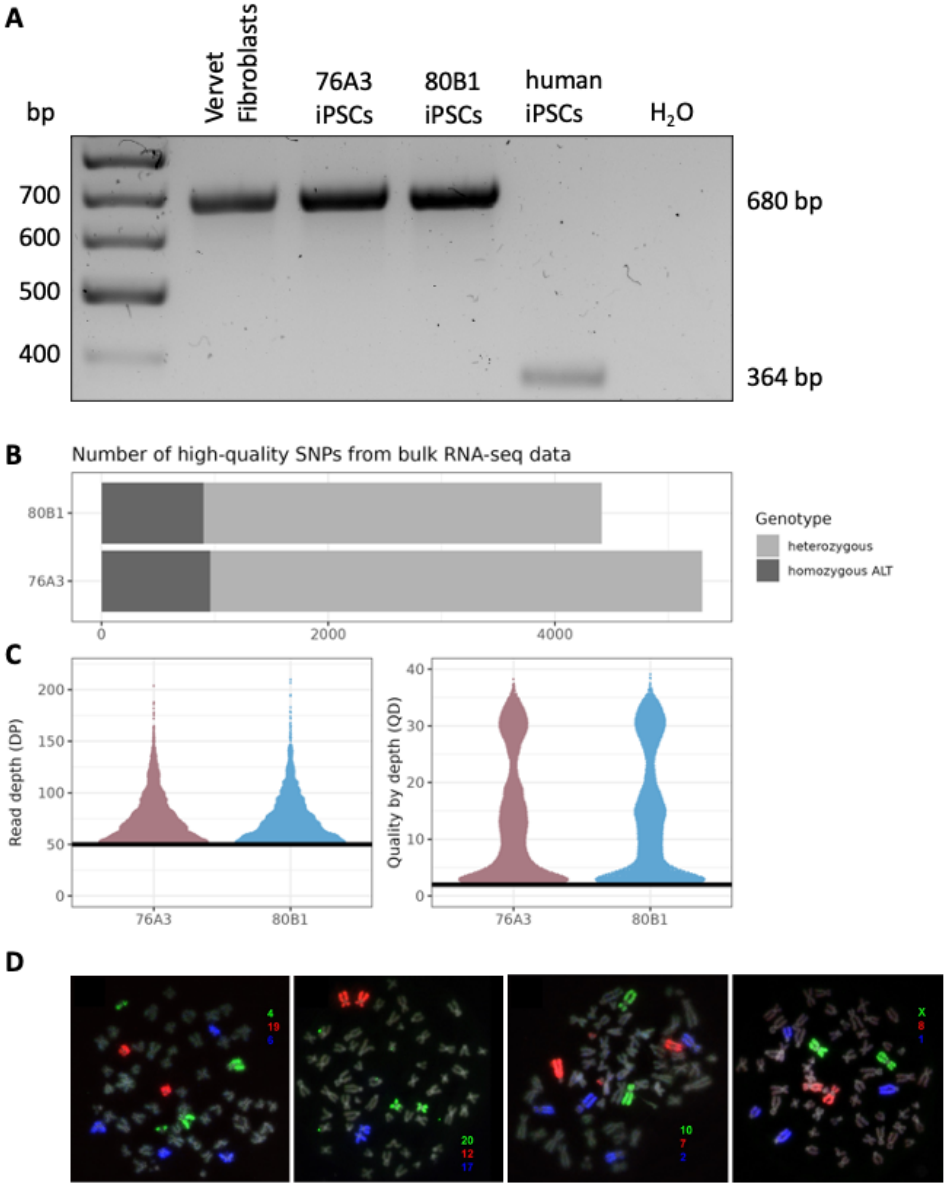
Enard, W., 2012. Functional primate genomics - leveraging the medical potential. J. Mol. Med. https://doi.org/10.1007/s00109-012-0901-4.

Herke et al., 2007; S. Herke, J. Xing, D. Ray, J. Zommerman, R. Cordaux, M. Batzer, A; SINE-based dichotomous key for primate identification; Gene (2007); https://doi.org/10.1016/j.gene.2006.08.015.

Jasinska et al., 2013; A. Jasinska, C. Schmitt, S. Service, R. Cantor, K. Dewar, J. Jentsch, J. Kaplan, T. Turner, W. Warren, G. Weinstock, R. Woods, N. Freimer; Systems Biology of the Vervet Monkey; ILAR Journal (2013); https://doi.org/10.1093/ilar/ilt0.

Juan et al., 2023; D. Juan, G. Santpere, J. Kelley, O. E. Cornejo, T. Marques-Bonet; Current advances in primate genomics: novel approaches for understanding evolution and disease; Nature reviews genetics (2023); https://doi.org/10.1038/s41576-022-00554-w.

Supplementary Figure



2.5 Generation and characterization of two fibroblast-derived Baboon induced pluripotent stem cell lines

Jessica Jocher, Fiona C. Edenhofer, **Philipp Janssen**, Stefan Müller, Eva Briem, Johanna Geuder, Wolfgang Enard

"Generation and characterization of two fibroblast-derived Baboon induced pluripotent stem cell lines"

Stem Cell Research 75, 103316 (2024).

doi: 10.1016/j.scr.2024.103316

Supplementary Information is freely available at the publisher's website:

https://www.sciencedirect.com/science/article/pii/S187350612400014X?via%3Dihub#s0080

Stem Cell Research 75 (2024) 103316

Contents lists available at ScienceDirect

Stem Cell Research

journal homepage: www.elsevier.com/locate/scr

Lab Resource: Animal Multiple Cell lines

Generation and characterization of two fibroblast-derived Baboon induced pluripotent stem cell lines

Jessica Jocher [a], Fiona C. Edenhofer [a], Stefan Müller [b], Philipp Janssen [a], Eva Briem [a], Johanna Geuder [a], Wolfgang Enard [a,*]

[a] Anthropology & Human Genomics, Faculty of Biology, Ludwig-Maximilians-Universität München, Großhaderner Straße 2, 82152 Martinsried, Germany
[b] Institute of Human Genetics, Munich University Hospital, Ludwig-Maximilians-Universität München, 80336 Munich, Germany

A B S T R A C T

Cross-species comparisons studying primate pluripotent stem cells and their derivatives are crucial to better understand the molecular and cellular mechanisms behind human disease and development. Within this context, Baboons (*Papio anubis*) have emerged as a prominent primate model for such investigations. Herein, we reprogrammed skin fibroblasts of one male individual and generated two induced pluripotent stem cell (iPSC) lines, which exhibit the characteristic ESC-like morphology, demonstrated robust expression of key pluripotency factors and displayed multilineage differentiation potential. Notably, both iPSC lines can be cultured under feeder-free conditions in commercially available medium, enhancing their value for cross-species comparisons.

1. Resource Table

Unique stem cell lines identifier	MPC-PapAnu-C00001 (100A1) MPC-PapAnu-C00002 (100B1.3)
Alternative name(s) of stem cell lines	100A1 100B1.3
Institution	Faculty of Biology, Ludwig-Maximilians-Universität München
Contact information of distributor	Prof. Dr. Wolfgang Enard: enard@bio.lmu.de Jessica Jocher: jocher@bio.lmu.de
Type of cell lines	iPSCs
Origin	Baboon (*Papio anubis*)
Additional origin info	Sex: male
Cell Source	iPSCs were derived from baboon skin fibroblasts
Clonality	Clonal
Method of reprogramming	Integration-free sendai virus based OSKM vectors (CytoTune-iPSC 2.0 Sendai Reprogramming Kit, Thermo Fisher Scientific) were used for reprogramming
Evidence of the reprogramming transgene loss	PCR analysis for transgene detection (negative)
Associated disease	N/A
Gene/locus	N/A
Date archived/stock date	April 2021
Cell line repository/bank	N/A
Ethical approval	Fibroblasts were isolated during an autopsy in an unrelated project that was approved by the Government of Upper Bavaria, Munich, Germany (reference number 55.2–1-54–2532-184–2014, September 2015).

(*continued on next column*)

(*continued*)

Unique stem cell lines identifier	MPC-PapAnu-C00001 (100A1) MPC-PapAnu-C00002 (100B1.3)

1.1. Resource utility

The utilization of two iPSC lines derived from one male Baboon skin sample enables cross-species comparisons, particularly for investigating the molecular and cellular evolution during early primate development. Additionally, these two lines offer the opportunity to evaluate clonal variation within the genetic background of one single Baboon.

2. Resource details

Comparative analyses of human and non-human primates (NHP) can provide valuable and unique information, allowing to gain insights into evolutionary and developmental mechanisms, as well as bridge the phylogenetic gap between humans and mice (Enard, 2012). The Baboon (*Papio anubis*) is a frequently used model in biomedical research, as well as in behavioral ecology (Fischer et al., 2019). However, availability of these animals is limited and obtaining comparable cells especially during development is practically and ethically challenging for many

* Corresponding author.
E-mail address: enard@bio.lmu.de (W. Enard).

https://doi.org/10.1016/j.scr.2024.103316
Received 6 September 2023; Received in revised form 21 December 2023; Accepted 16 January 2024
Available online 17 January 2024

Table 1
Characterization and validation.

Classification	Test	Result	Data
Morphology	Photography Bright field	Normal colony morphology	Fig. 1A
Phenotype	Qualitative analysis by immunocytochemistry	iPSCs were positively stained for OCT3/4, NANOG, SOX2, TRA-1–60 and SSEA4	Fig. 1B
	Quantitative analysis by immunocytochemistry counting	% of total cells positive for pluripotency markers (mean ± SD): 100A1 OCT3/4: 97.9 % ± 2.9 % (3,071 cell counted) NANOG: 98.2 % ± 0.3 % (3,070 cell counted) SOX2: 99.1 % ± 0.6 % (3,362 cell counted) 100B1.3 OCT3/4: 97.8 % ± 0.6 % (3,725 cell counted) NANOG: 98.8 % ± 1.1 % (2,845 cells counted) SOX2: 98.9 % ± 0.3 % (2,492 cells counted)	Fig. 1C
Genotype	Karyotype (G-banding) and resolution	2x inconspicuous male karyotype, 42,XY No recurrent numeric or structural aberrations, after G-banding analysis of more than 45 cells per cell line with up to approximately 400 bphs (bands per haploid set)	Fig. 1D and Supplementary Fig. S1D
Identity	SINE-based genotyping PCR	DNA profiling performed, matched between iPSCs and parental fibroblasts	Supplementary Fig. S1A
	SNP analysis	Variant calling performed resulting in 7000 high quality SNPs	Submitted in archive with journal Summary: Supplementary Fig. S1B,C
Mutation analysis (IF APPLICABLE)	N/A N/A		
Microbiology and virology	Mycoplasma	Mycoplasma testing by PCR: negative	Fig. 1E
	Sendai virus	PCR analysis for Sendai virus presence: negative	Fig. 1F
Differentiation potential	Embryoid body formation	iPSCs are capable of differentiating into the three germ layers. Mesoderm: Smooth muscle actin (SMA) and COL1A1, Endoderm: α-feto protein (AFP) and SOX17, Ectoderm: β-III Tubulin and PAX6	Fig. 1G
Donor screening (OPTIONAL)	N/A		
Genotype additional info (OPTIONAL)	N/A N/A		

studies. Therefore, generating induced pluripotent stem cells (iPSCs) from NHPs can aid in establishing renewable sample resources and help to overcome these challenges (Juan et al., 2023).

Here, Baboon skin fibroblasts were reprogrammed to iPSCs using a commercially available Sendai virus kit to introduce the Yamanaka factors OCT3/4, SOX2, KLF4 and C-MYC into the cells. Following transduction, emerging colonies were picked, gradually transitioned to feeder-free culture conditions, and further characterized (Table 1). The resulting iPSC clones exhibit characteristic ESC-like features, including compact cellular packaging, defined colony borders and a high nuclear / cytoplasm ratio (Fig. 1A). To confirm pluripotency of the iPSCs, immunofluorescence (IF) staining was performed, affirming the expression of the pluripotency-associated proteins NANOG, OCT3/4 and SOX2, in addition to the presence of cell surface markers TRA-1–60 and SSEA4 (Fig. 1B). Quantitative analysis of the IF staining demonstrated a substantial proportion of cells (>95 %) are expressing NANOG, OCT3/4 and SOX2 (Fig. 1C). A primate-specific SINE-based PCR was conducted, confirming the same Baboon-specific ALU element insertions as the parental skin fibroblasts, thereby validating their origin from the same species (Herke et al., 2007) (Supplementary Fig. S1A). In addition, single nucleotide polymorphisms (SNPs) were called from bulk RNA-sequencing (bulk RNA-seq) data to profile the genotype of the cell lines. Around 7000 high quality SNPs with high coverage in both clones and parental fibroblasts were retrieved (Supplementary Fig. S1B,C). All iPSCs were negative for mycoplasma contamination (Fig. 1E) and the absence of Sendai-based reprogramming vectors was proven by PCR (Fig. 1F). Karyotype analysis revealed no recurrent numerical or structural abnormalities (Fig. 1D). In addition, a detailed high resolution analysis of numerical and structural chromosome integrity was performed by FISH using selected human whole chromosome specific painting probes to further validate one of the cell lines. All chromosomes were stained as expected, with baboon chromosome 10 being stained with two probes based on an evolutionary fusion of human chromosome 20 and 22 homologs (Best et al., 1998) (Supplementary Figure S1D). To assess the differentiation potential of iPSCs, an in vitro differentiation to embryoid bodies (EBs) was performed, and subsequent staining for alpha-fetoprotein (AFP), SOX17, alpha-smooth-muscle actin (SMA), procollagen-1 alpha-1 (COL1A1), PAX6 and neuron-specific beta-III tubulin confirmed differentiation into the three germ layers (Fig. 1G). In

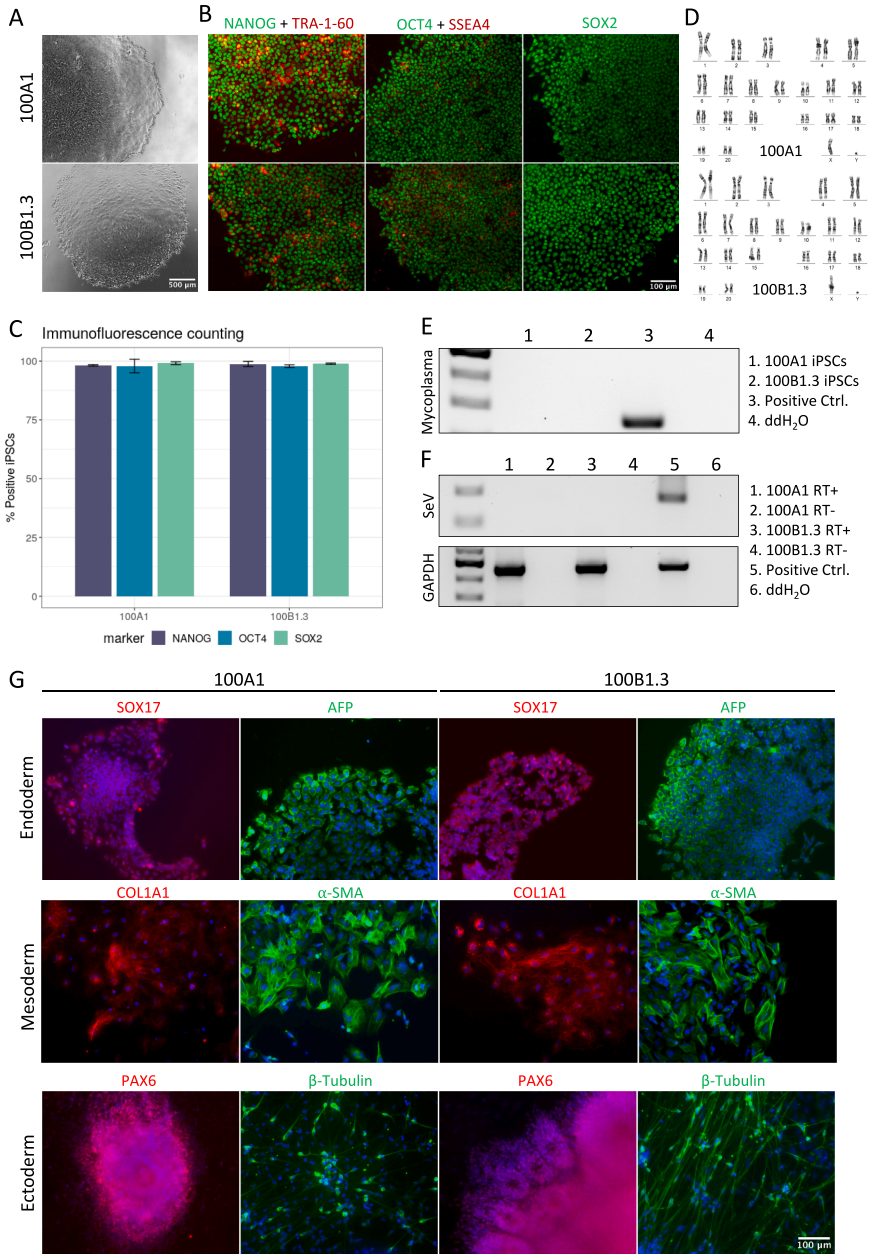


Fig. 1. Characterization of two Baboon iPSC lines. (A) Phase contrast microscopy images of iPSC colonies. Scale bar represents 500 µm. (B) Immunofluorescence staining for pluripotency markers. Scale bar represents 100 µm. (C) Immunofluorescence counting results for NANOG, OCT4 and SOX2. (D) Karyotype analysis. (E) Mycoplasma test. (F) PCR for Sendai-based reprogramming vectors. (G) Immunofluorescence staining for germ layer-specific markers. Scale bar represents 100 µm.

Table 2
Reagents details.

	Antibodies used for immunocytochemistry/flow-cytometry			
	Antibody	**Dilution**	**Company Cat #**	**RRID**
Pluripotency Markers	Rabbit anti-OCT4	1:400	Cell Signaling Technology, Cat# 2750S	RRID: AB_823583
	Mouse anti-SOX2	1:400	Cell Signaling Technology, Cat# 4900S	RRID: AB_10560516
	Rabbit anti-NANOG	1:400	Cell Signaling Technology, Cat# 4903S	RRID: AB_10559205
	Mouse anti-SSEA4	1:500	NEB, Cat# 4755S	RRID: AB_1264259
	Mouse anti-TRA-1-60	1:100	Stem Cell Technologies, Cat# 60064	RRID: AB_2686905
Differentiation Markers	Mouse anti-α-Smooth Muscle Actin	1:100	R&D Systems, Cat# MAB1420	RRID: AB_262054
	Sheep anti-COL1A1	1:200	R&D Systems, Cat# AF6220	RRID: AB_10891543
	Mouse anti-Neuron-specific beta-III Tubulin	1:100	R&D Systems, Cat# MAB1195	RRID: AB_357520
	Rabbit anti-PAX6	1:100	Thermo Fisher Scientific, Cat# 42-6600	RRID: AB_2533534
	Mouse anti-alpha Fetoprotein	1:100	R&D Systems, Cat# MAB1368	RRID: AB_357658
	Rabbit anti-SOX17	1:500	Bio-Techne, Cat# NBP2-24568	RRID: AB_3075468
Secondary Antibodies	Alexa Fluor 488 donkey anti-mouse IgG (H + L)	1:500	Thermo Fisher Scientific, Cat# A-21202	RRID: AB_141607
	Alexa Fluor 594 donkey anti-rabbit IgG (H + L)	1:500	Thermo Fisher Scientific, Cat# A-21207	RRID: AB_141637
	Alexa Fluor 488 donkey anti-sheep IgG (H + L)	1:500	Thermo Fisher Scientific, Cat# A-11015	RRID: AB_2534082
	Primers			
	Target	**Size of band**	**Forward/Reverse primer (5′-3′)**	
Reprogramming factor clearance	Sendai Virus	180 bp	GGATCACTAGGTGATATCGAGC / ACCAGACAAGAGTTTAAGAGATATGTATC	
	GAPDH (housekeeping gene)	450 bp	ACCACAGTCCATGCCATCAC / TCCACCACCCTGTTGCTGTA	
Mycoplasma testing	Mycoplasma 16S	270 bp	TGCACCATCTGTCACTCTGTTAACCTC / GGGAGCAAACAGGATTAGATACCCT	
Genotyping PCR	*Alu* (primate-specific SINE)	666 bp	TCTAAGGCAGCCATTGAGTG / CCAGGTTTGCCTCTGACTCC	

summary, these characteristics suggest the successful reprogramming of Baboon fibroblasts to two feeder-free iPSC lines.

3. Materials and methods

3.1. Reprogramming of fibroblasts and iPSC maintenance

Fibroblasts were cultured in DMEM/F12 (Fisher Scientific) supplemented with 10 % FBS, 100 U/ml Penicillin and 100 µg/ml Streptomycin (Thermo Fisher Scientific) on 0.2 % Gelatin-coated dishes at 37 °C with 5 % CO_2. For reprogramming, the CytoTune™-iPS 2.0 Sendai Reprogramming Kit (Thermo Fisher Scientific) was used at a MOI of 5 following a modified protocol. Briefly, a suspension infection with the virus mix was conducted for 1 h at 37 °C, followed by seeding on feeder cell (mitomycin-C treated mouse embryonic fibroblasts) -coated wells. The culture medium was switched to mTesR1™ (STEMCELL Technologies) on day 5. Emerging colonies were manually picked on feeder cells in StemFit® Basic02 (Ajinomoto) supplemented with 100 ng/mL bFGF (Peprotech) and PenStrep. To generate feeder-free iPSCs, cells were split using 0.5 mM EDTA on 1 % Geltrex™ (Thermo Fisher Scientific) -coated wells in feeder-conditioned StemFit. The ratio of normal to feeder-conditioned StemFit was increased by 25 % after every second passage, until iPSCs adapted to feeder-free culture conditions. Every 5–7 days, 0.5 mM EDTA was used for routine passaging at a ratio of 1:10–1:40, and the medium was exchanged every other day.

3.2. Immunocytochemistry

Cells at passage 20–25 were fixed with 4 % PFA for 15 min, permeabilized with 0.3 % Triton X-100 and blocked with 5 % FBS for 30 min. Cells were incubated with primary antibodies (Table 2) diluted in staining buffer (PBS containing 1 % BSA and 0.3 % Triton X-100) at 4 °C overnight. Following, cells were washed with PBS and incubated with secondary antibodies (Table 2) diluted in staining buffer for 1 h at RT. Nuclei were counterstained with 1 µg/mL DAPI. The percentage of positively-stained cells was quantified with the ImageJ software using the Cell Counter plugin. Between 2,492 and 3,725 cells were counted for each marker.

3.3. Embryoid body formation

One 6-well of iPSCs at passage 20–25 was dissociated to clumps and cultured in sterile bacterial dishes containing StemFit w/o bFGF at 37 °C with 5 % CO_2. Medium was changed every second day during the first 8 days of floating culture. On day 8, EBs were seeded into 6-wells coated with 0.2 % Gelatin allowing outgrowth of the EBs. On day 16, differentiated cells were stained using specific antibodies for mesoderm, endoderm, and ectoderm (Table 2).

3.4. Karyotyping

Cells at 80 % confluency (passage 15–20) were incubated with 0.1 mg/mL Colcemid (Gibco) for 15 h, dissociated using Accumax™ (Sigma Aldrich) and treated with hypotonic Na-Citrate / NaCl for 35 min at 37 °C. Subsequently, cells were fixed with methanol / acetic acid glacial (3:1) for 20 min at -20 °C and washed twice with methanol/acetic as stated above. A standard protocol for the preparation of differentially stained mitotic chromosome spreads using the G-banding technique was applied. Fluorescence in situ hybridization (FISH) was performed using human chromosome specific painting probes. In brief, mixtures of fluor conjugated paint probes were denatured at 75 °C for 5 min, added to the metaphase slide, covered with a cover slip and sealed with rubber cement. The slide was denatured at 75 °C for 2 min in a Hybrite (VYSIS, US) hybridization station and hybridized at 37 °C overnight, followed by a 2 min post-hybridization wash in 0.1xSSC buffer at 60 °C. Final slides were mounted in Vectashield embedding medium containing DAPI (Vector Laboratories, UK) and analyzed using an Axioplan 2 Imaging microscope (Zeiss, Germany).

3.5. Mycoplasma testing

The medium of one confluent 6-well with iPSCs at passage 19–22 was collected, pelleted, and resuspended in 100 µL PBS. After an incubation for 5 min at 95 °C, 1 µL was used for a screening PCR with specific primers for the Mycoplasma 16S rRNA (Table 2). A sample that had previously tested positive was used as an internal control.

3.6. Genotyping PCR

Total gDNA was isolated using the DirectPCR Lysis Reagent (VWR) supplemented with 20 mg/mL Proteinase K (Life Technologies), and a PCR (36 cycles) was conducted with primers for the primate-specific *Alu* SINE (Table 2).

3.7. SeV detection

Total RNA was isolated from one 6-well of iPSCs (passage 15–20) using the Direct-zol RNA Microprep Kit (Zymo Research) and cDNA was synthesized using the Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific). 50 ng cDNA were used to perform a PCR (36 cycles) with specific primers for SeV and GAPDH as housekeeping gene (Table 2).

3.8. Bulk RNA-sequencing and variant calling

iPSCs and parental fibroblasts were dissociated using Accumax, sampled in three biological replicates each and bulk RNA-seq libraries were generated using the Prime-seq workflow (https://www.protocols.io/view/prime-seq-81wgb1pw3vpk/v2). Bulk RNA-seq data of iPSCs and parental fibroblasts were used to call SNPs against the reference genome papAnu4 using GATK (Genome Analysis Tool Kit). High quality, biallelic SNPs were retained by joint genotyping of data from both iPSC lines and fibroblasts followed by quality filtering of the variants for high coverage (DP > 99) and quality by depth (QD > 2).

CRediT authorship contribution statement

Jessica Jocher: Conceptualization, Investigation, Methodology, Visualization, Writing - original draft, Writing - review & editing. **Fiona C. Edenhofer:** Investigation, Visualization. **Stefan Müller:** Investigation, Methodology, Visualization. **Philipp Janssen:** Data curation, Formal analysis, Visualization, Writing - review & editing. **Eva Briem:** Investigation. **Johanna Geuder:** Methodology. **Wolfgang Enard:** Conceptualization, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Wolfgang Enard reports financial support, article publishing charges, and travel were provided by German Research Foundation.

Acknowledgements

This work was supported by DFG EN 1093/5-1 (project number 458247426). We are grateful to Dr. Jan-Michael Abicht and the team of the SFB Xenotransplantation at LMU clinic for kindly providing the primary material. We thank Vanessa Baltruschat for her substantial technical support in the lab.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scr.2024.103316.

References

Enard, W., 2012. Functional primate genomics - leveraging the medical potential. J. Mol. Med. https://doi.org/10.1007/s00109-012-0901-4.

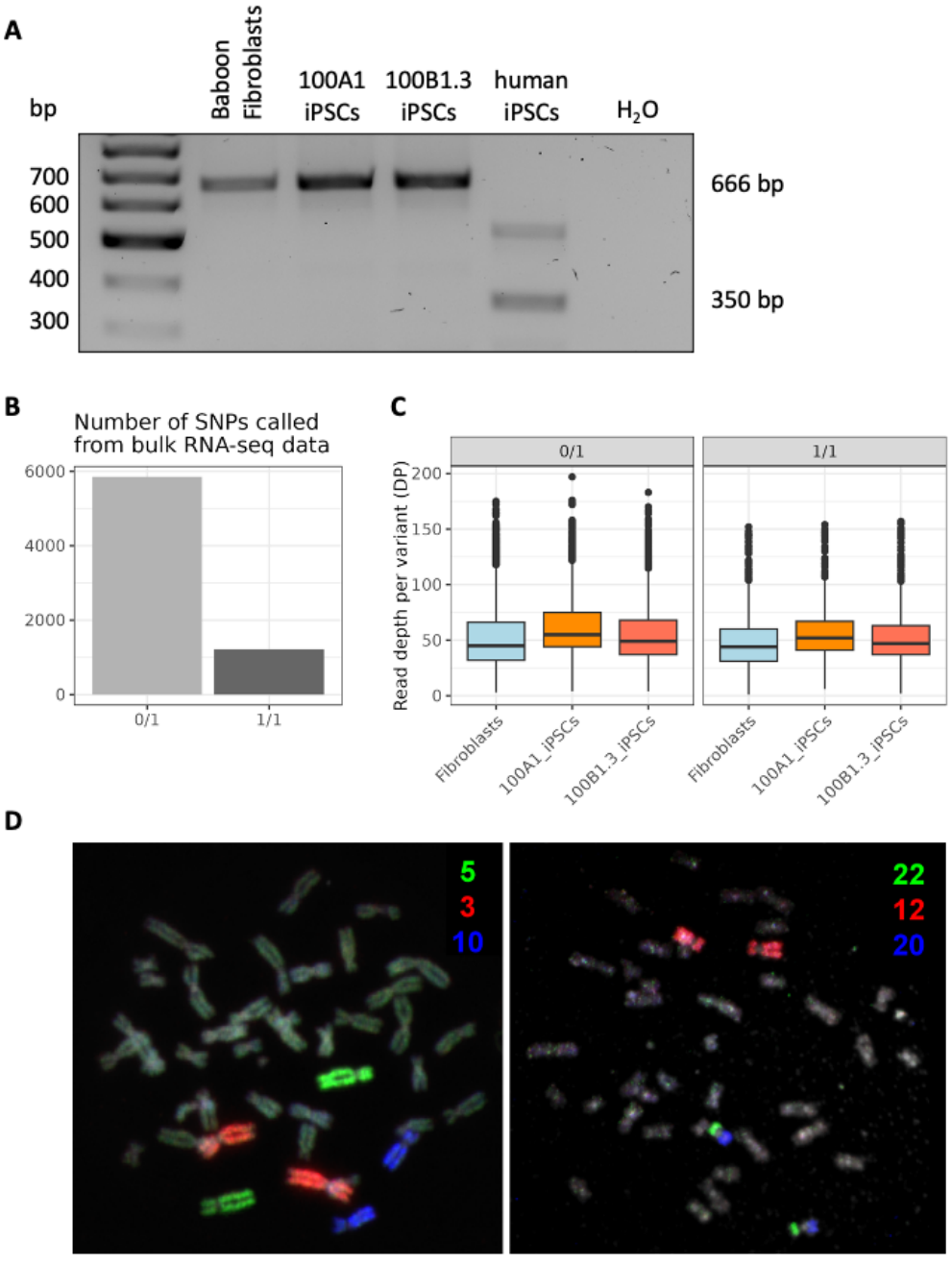
Best, R.G., Diamond, E., Crawford, F.S., Grass, C., Janish, T.L., Lear, D., Soenksen, A.A., Szalay, C.M Moore, 1998. Baboon/human homologies examined by spectral karyotyping (SKY): a visual comparison. Cytogenet. Cell Genet. https://doi.org/10.1159/000015070.

Fischer, J., Higham, J., Alberts, S.C., Barrett, L., Beehner, J.C., Bergman, T.J., Carter, A. J., Collins, A., Elton, A., Fagot, J., Ferreira da Silva, M.J., Hammerschmidt, K., Henzi, P., Jolly, C.J., Knauf, S., Kopp, G.H., Rogers, J., Roos, C., Ross, C., Seyfarth, R. M., Silk, J., Snyder-Mackler, N., Staedele, V., Swedell, L., Wilson, M.L., Zinner, D., 2019. The Natural History of Model Organisms: Insights into the evolution of social systems and species from baboon studies. eLIFE. https://doi.org/10.7554/eLife.50989.001.

Herke, S., Xing, J., Ray, D., Zommerman, J., Cordaux, R., Batzer, M., 2007. A SINE-based dichotomous key for primate identification. Gene. https://doi.org/10.1016/j.gene.2006.08.015.

Juan, D., Santpere, G., Kelley, J., Cornejo, O.E., Marques-Bonet, T., 2023. Current advances in primate genomics: novel approaches for understanding evolution and disease. Nat. Rev. Genet. https://doi.org/10.1038/s41576-022-00554-w.

Supplementary Figure



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

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

*contributed equally

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

Reviewed Preprint at *eLife* 14:RP105398 (2025).

doi: 10.7554/eLife.105398.1

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).

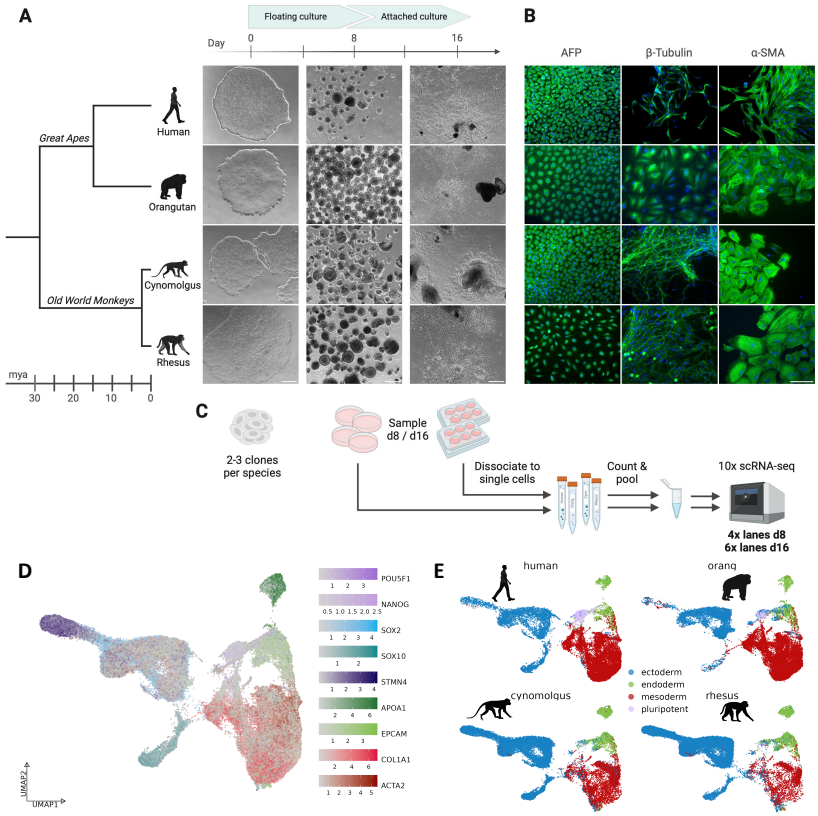


Figure 1. Generation of primate embryoid bodies. A) Overview about the EB differentiation workflow of the four primate species human (*Homo sapiens*), orangutan (*Pongo abelii*), cynomolgus (*Macaca fascicularis*) and rhesus (*Macaca mulatta*), including their phylogenetic relationship. Scale bar represents 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.

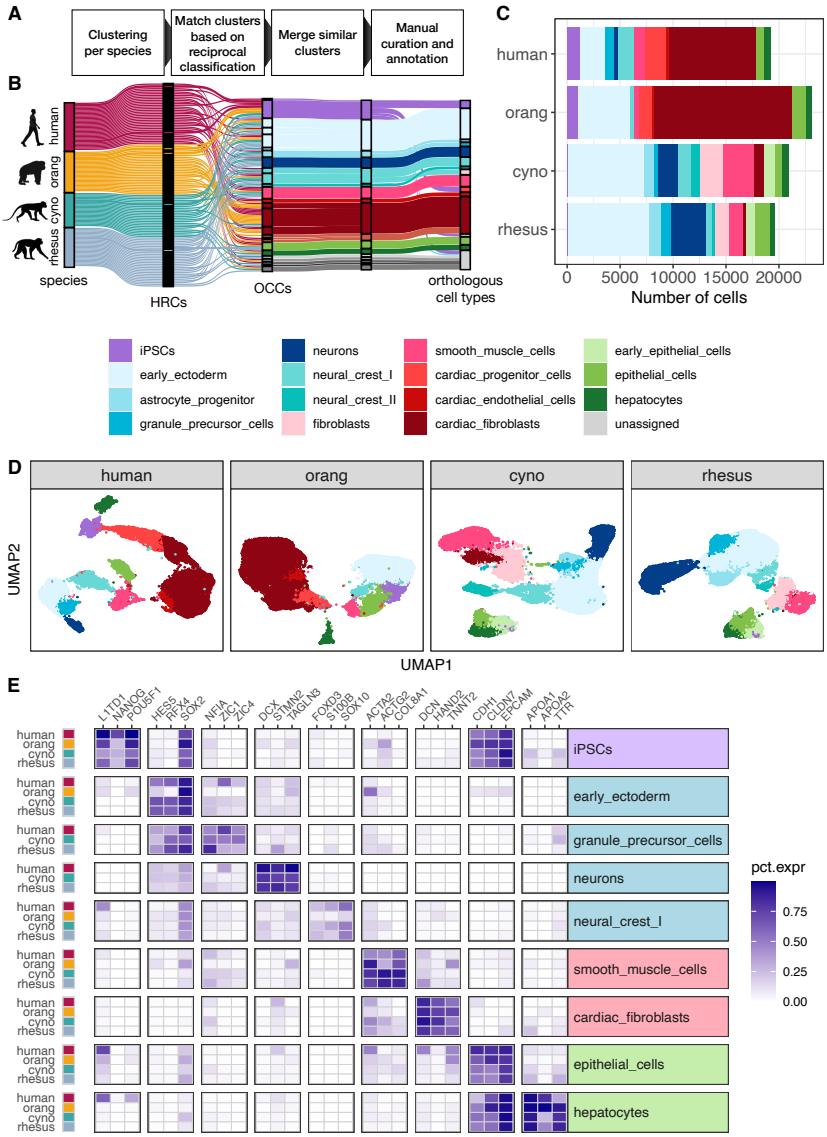


Figure 2. Assignment of orthologous cell types across species. A) Schematic overview of the pipeline to match clusters between species and assign orthologous cell types. B) Sankey plot visualizing the intermediate steps of the cell type assignment pipeline. Each line represents a cell which are colored 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.

Cell type-specific genes have less conserved expression levels

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 separate cell lines in 3 independent replicates. 9 of these were detected in at least 3 species, and 7 cell types were highly reproducibly detected in all four species (Figure 2C, D; Supplementary Figure S6). 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). Based on the premise that it is not necessarily the expression level, but rather the expression breadth that determines expression conservation [31], 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* [32] 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 S7A).

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 S7B).

Unsurprisingly, broadly expressed genes also showed higher average expression levels [33] (Supplementary Figure S7D). 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 S7E,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 S7C).

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[35]. 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 S8).

Given the special role of transcriptional regulators for the definition of a cell type [36] and the differences in conservation between protein-coding and non-coding RNAs [37], we analyzed the comparability of marker genes of different types. To this end, we assessed the

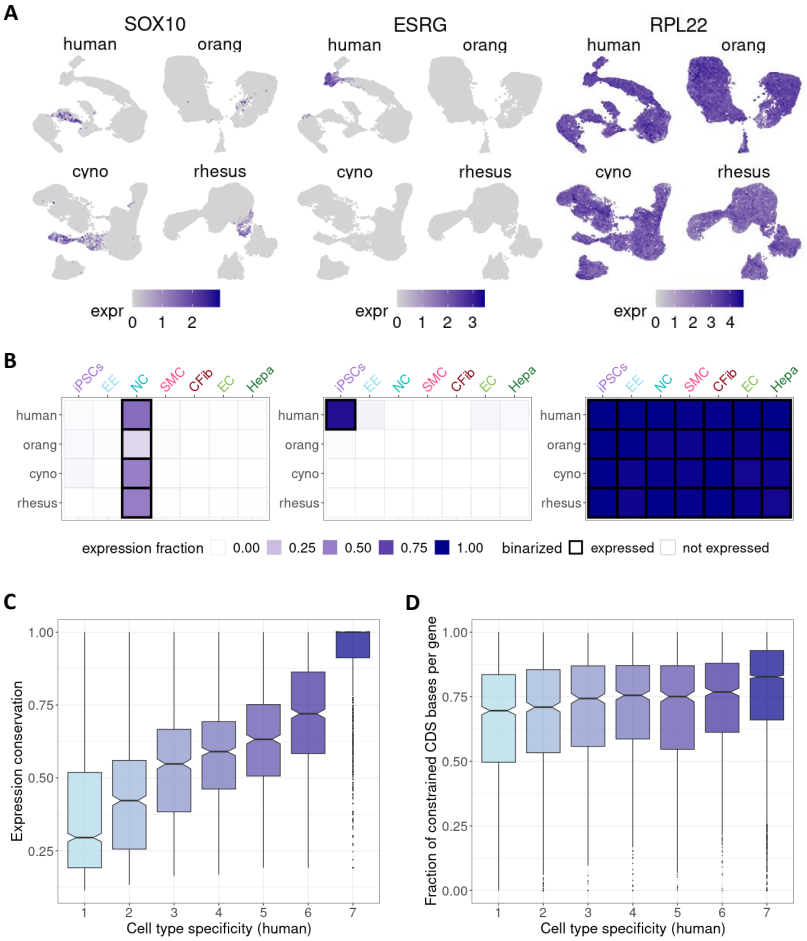


Figure 3. Effect of cell type specificity on expression conservation. A) UMAP visualizations depicting expression patterns of selected example genes: *SOX10* (conserved cell type-specific expression in neural crest cells), *ESRG* (species-specific and cell type-specific expression in human iPSCs), and *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 with different levels of cell type specificity in human. D) Boxplot of the fraction of coding sequence sites that were found to evolve under constraint based on a 43 primate phylogeny [34], stratified by human cell type specificity.

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 [38]. 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 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 S9). 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 S10). 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.

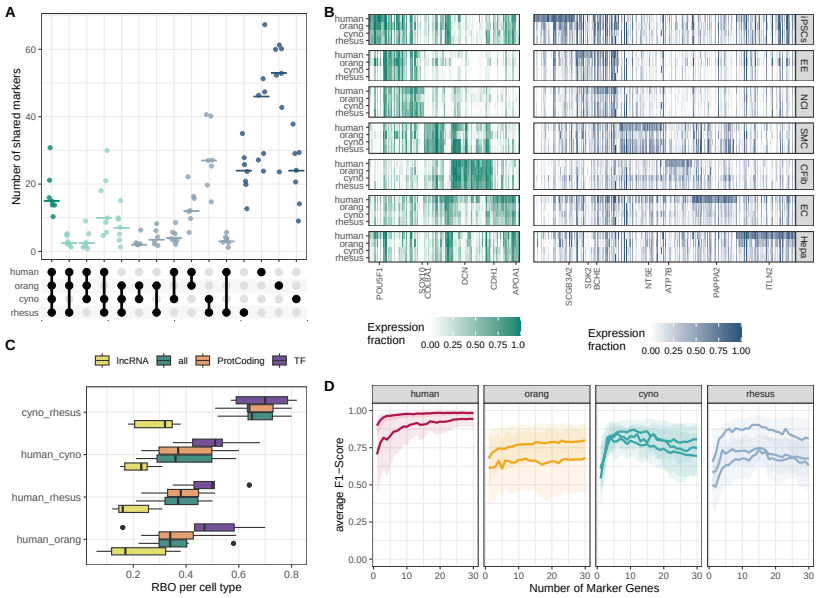


Figure 4. Evaluation of marker gene conservation. A) UpSet plot illustrating the overlap between species for the top 100 marker genes per cell type. B) Heatmap showing the expression fractions of marker genes: on the left, markers shared among all species, and on the right, markers unique to the human ranking. For each cell type, one representative gene is labeled and further detailed in Supplementary Figure S8. 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, 39]. 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

[40] 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).

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 [41, 31]. 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, 42]. 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 [37]. On the other hand, TFs, which have been proposed as central elements of a Core Regulatory Complex (CoRC) that defines cell type identity [36], 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 S10). 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 [43, 44]. 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 [45]. 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 [46], followed by removal of transcripts with partial mapping (<50%), low sequence identity (<50%) or excessive length (>100 bp difference and >2 length ratio).

Species and individual demultiplexing

Since we pool cells from multiple species on each 10x lane, we use cellsnp-lite [47] version 1.2.0 and vireo [48] 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 [49] 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 [50], 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 [51] 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.

Besides the separate processing per species, we also create an integrated dataset of all 4 species together using Harmony version 0.1.1 [28]. We identify clusters on the first 20 Harmony-integrated PCs with Seurat at a resolution of 0.1 (Figure 1D,E).

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.

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 [52]. 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. [53]. 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. [34]. 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'. We identify marker genes using the p-values ($p_{adj} < 0.1$) determined by ZIQ-Rank [35] 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) [38] 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 [54], JASPAR 2022 vertebrate unvalidated [54] and IMAGE [55]. 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 (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** (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] 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

[32] 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

[33] 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., 279001–124 (2024). doi:10.1101/gr.279001.124

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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

[45] 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

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

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

[48] 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

[49] 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

[50] 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

[51] 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

[52] 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

[53] 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

[54] 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

[55] 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

[56] 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

[57] 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

[58] 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

[59] 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

[60] 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

[61] 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

[62] 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

[63] 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

[64] 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

[65] 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

[66] 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

[67] 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

[68] 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

[69] 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

[70] 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

[71] 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

[72] 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

[73] 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

[74] 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

[75] 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

[76] 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

[77] 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

[78] 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

[79] 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

[80] 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

[81] 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

[82] 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

[83] 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

[84] 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

[85] 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

[86] 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

[87] 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

[88] 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

[89] 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

[90] 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

[91] 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

[92] 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

[93] 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

[94] 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

[95] 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

[96] 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

[97] 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

[98] 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

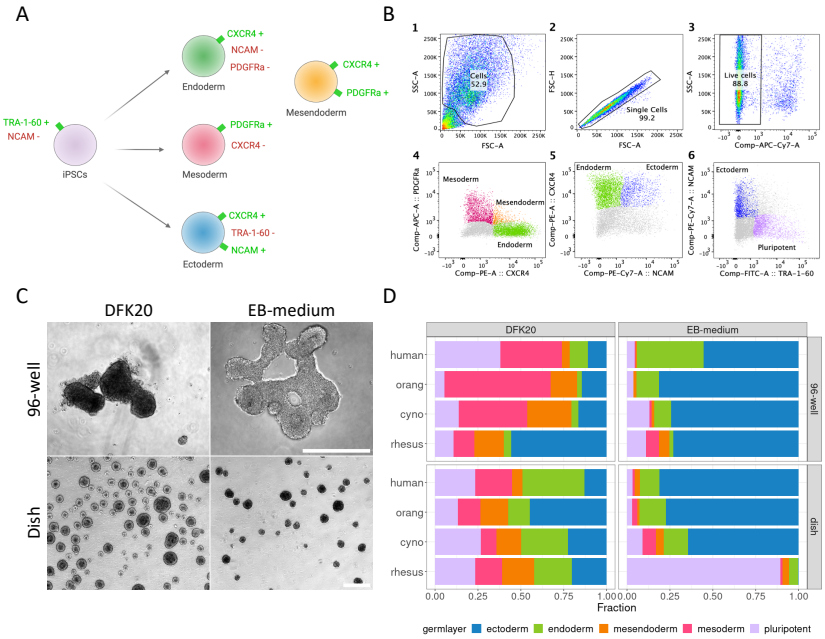
Jessica Jocher[1,*], Philipp Janssen[1,*], Beate Vieth[1], Fiona C. Edenhofer[1], Tamina Dietl[2], Anita Térmeg[1], Johanna Geuder[1], Wolfgang Enard[1,**], Ines Hellmann[1,**]

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

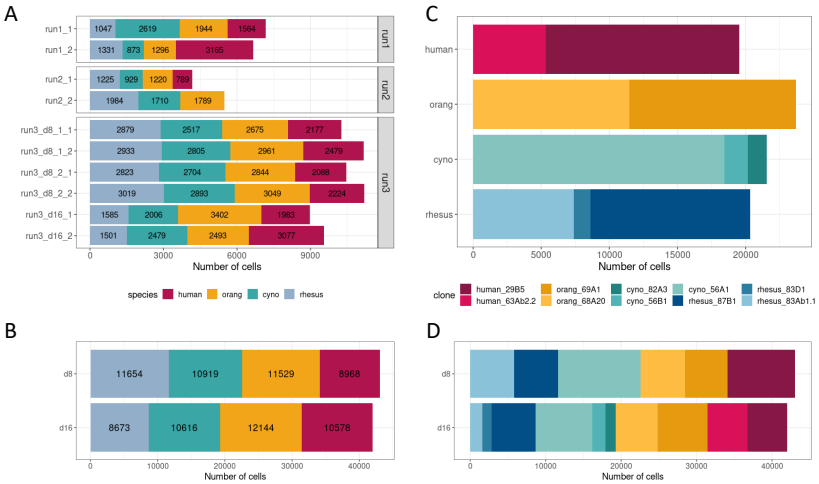
[2]Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt: Munich, Germany

[*] contributed equally

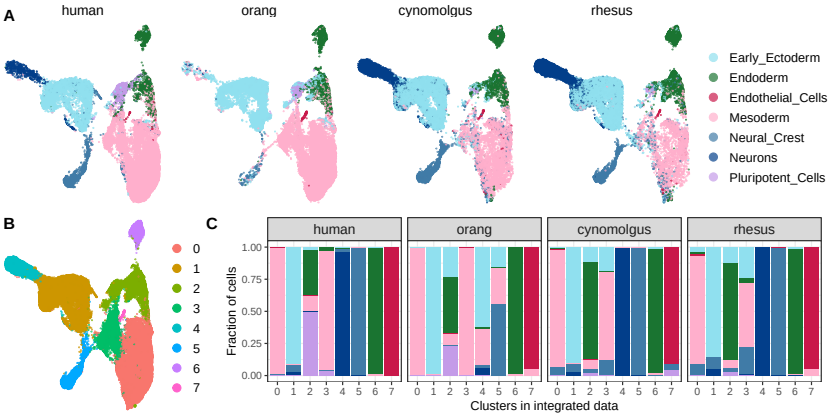
[**] correspondence hellmann@bio.lmu.de



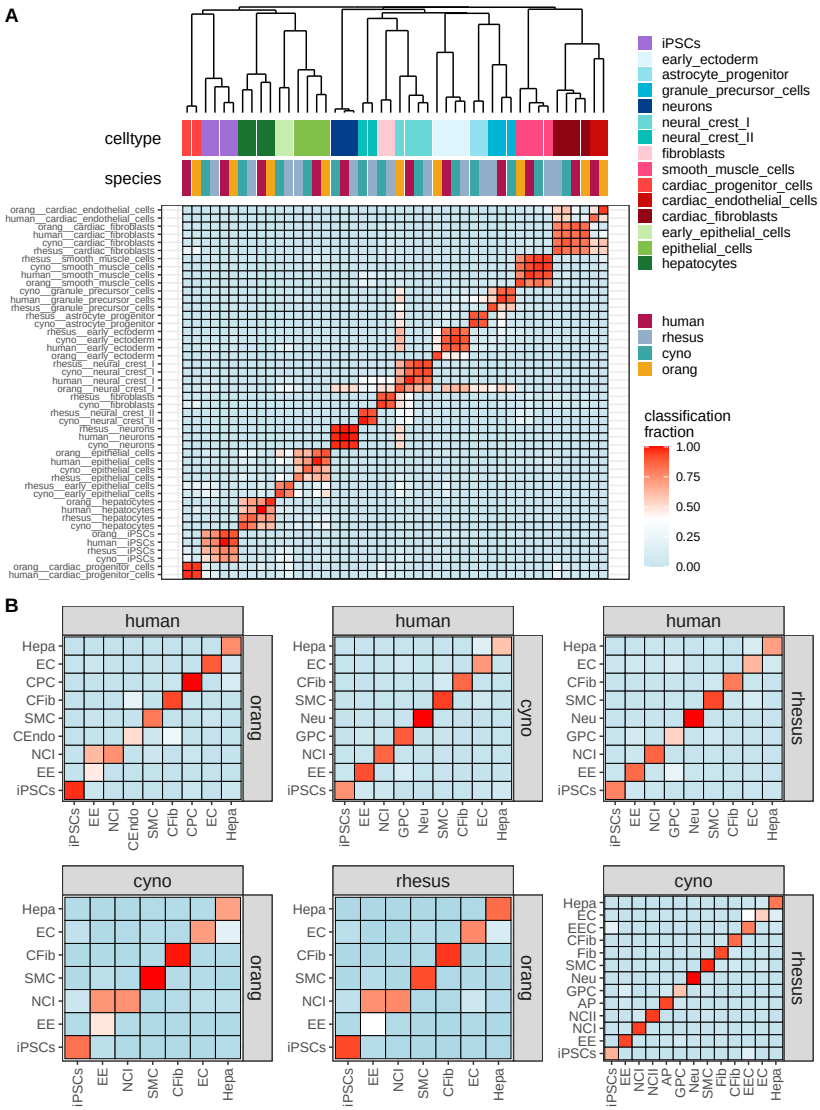
Supplementary Figure S1. Comparison of EB differentiation protocols using flow cytometry. A) Antibody combination to analyze iPSCs and cells of the three primary germ layers in a single sample. B) Flow cytometry gating overview using human EBs at day 7 of differentiation. 1. Gating of cell population. 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.



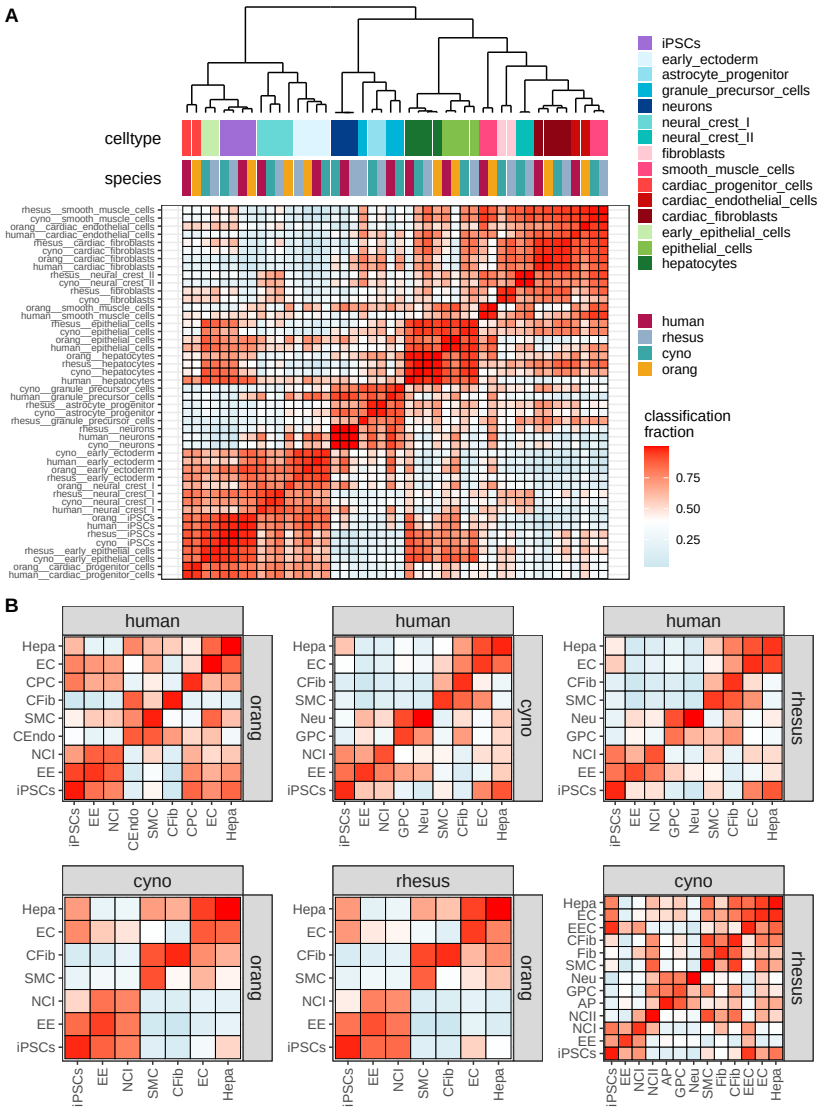
Supplementary Figure S2. Total number of recovered cells. A) Barplot of cell numbers per species and experimental batch and 10x lane. B) Barplot of cell numbers per species and day of differentiation. C) Barplot of cell numbers per clone. D) Barplot of cell numbers per clone and day of differentiation.



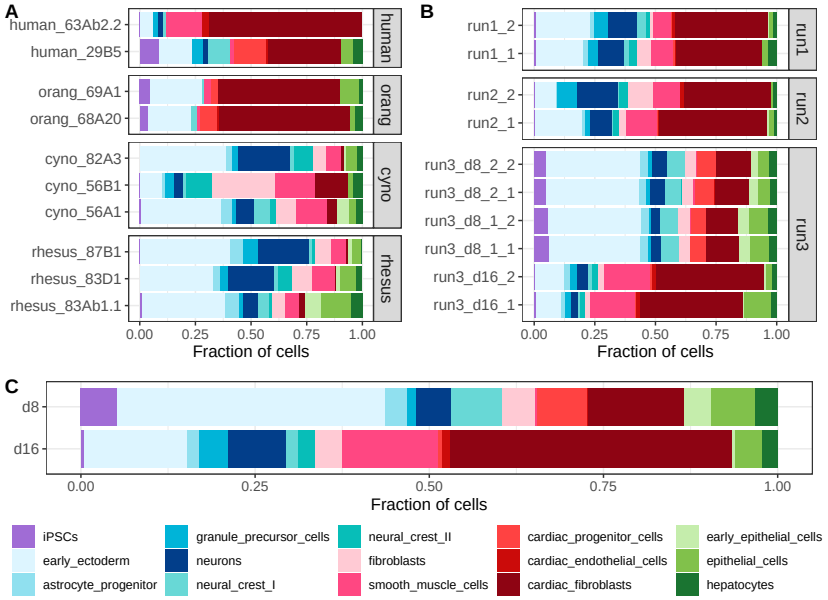
Supplementary Figure S3. Reference based cell type classification. A) UMAP representations colored by labels from a classification with a reference dataset of day 21 human embryoid bodies [18]. B) Single cell clusters in integrated data from all 4 species. C) Stacked bar plot of the proportions of predicted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



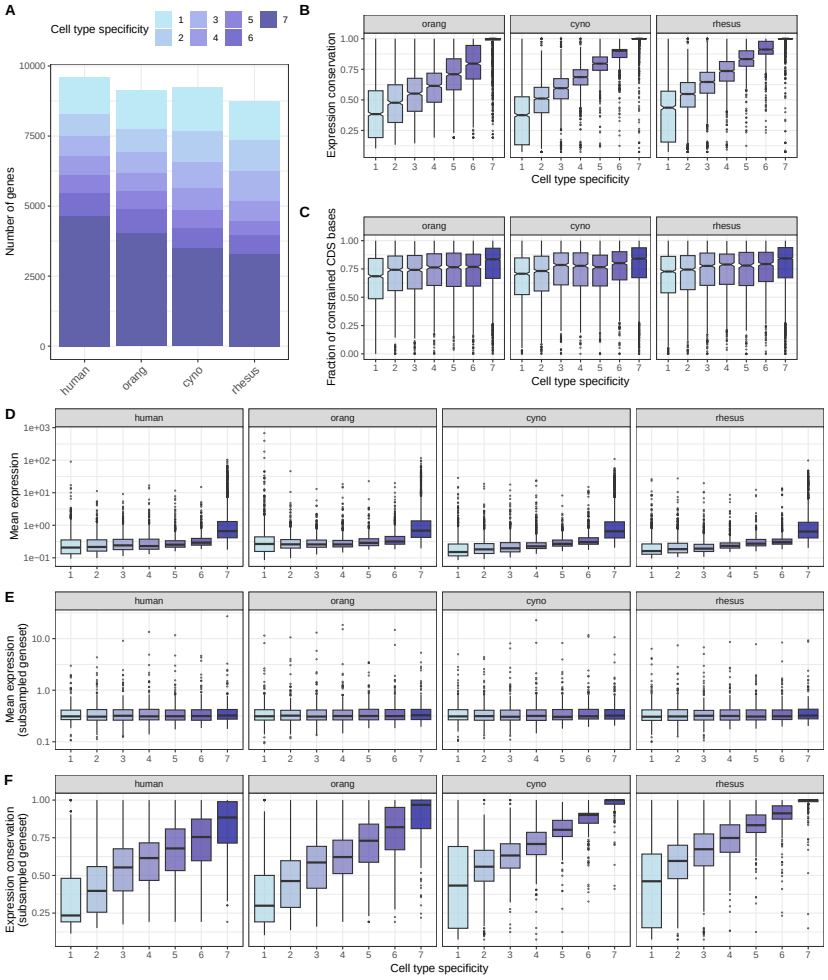
Supplementary Figure S4. Replicability of cell types across species measured by reciprocal classification. A) Heatmap illustrating 'all vs all' similarities of cell types from all four species. For each cell type pair the similarity represents the average classification fraction obtained through reciprocal 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.



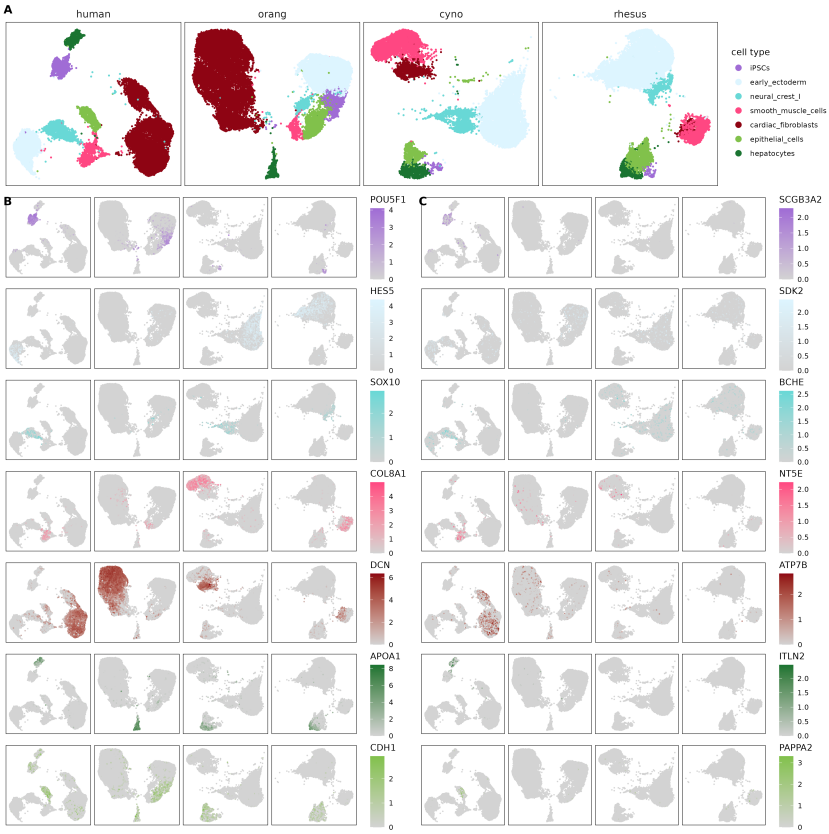
Supplementary Figure S5. Replicability of cell types across species measured with MetaNeighbor. A) Heatmap illustrating 'all vs all' similarities of cell types from all four species. For each cell type pair the similarity represents area under the receiver operator characteristic curve (AUROC) scores 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.



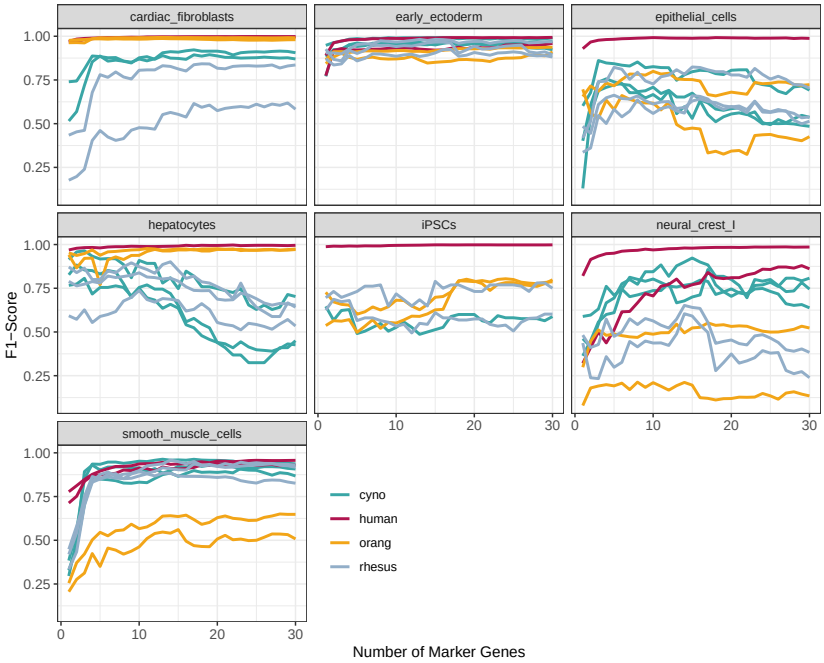
Supplementary Figure S6. Cell type annotation. A) Barplot of cell type fractions per species and clone. B) Barplot of cell type fractions per experimental batch and 10x lane. C) Barplot of cell type fractions per day of differentiation.



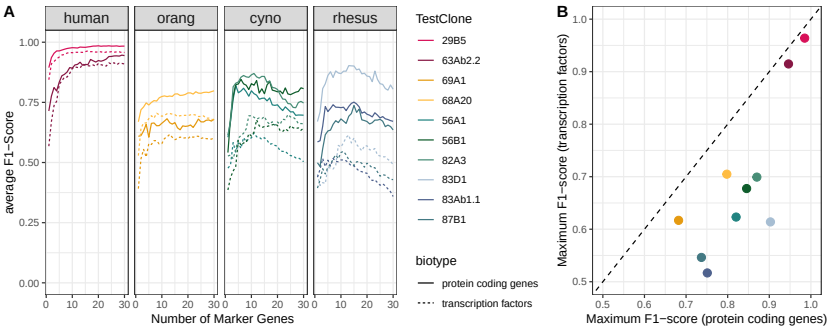
Supplementary Figure S7. Characteristics of genes with different levels of cell type-specific expression. A) Stacked bar plot of the number of genes per cell type specificity level for different species. B) Boxplot of expression conservation of genes with different levels of cell type specificity in orangutan, cynomolgus and rhesus. C) Boxplot of gene-level constraint based on primate phastCons scores [34] 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 S8. 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 S9. 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 S10. 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	[56]	[57]
iPSCs	NANOG	[56]	[58]
iPSCs	L1TD1	[59]	[59]
early ectoderm	SOX2	[60]	[61]
early ectoderm	HES5	[62]	[63]
early ectoderm	RFX4	[62]	[64]
granule precursor cells	NFIA	[65]	[66]
granule precursor cells	ZIC1	[67]	[68]
granule precursor cells	ZIC4	[67]	[69]
neural crest	SOX10	[32]	[32, 70]
neural crest	FOXD3	[71]	[72]
neural crest	S100B	[73]	[74]
neurons	STMN2	[75]	[76, 77]
neurons	TAGLN3 (NP25)	[78]	[77]
neurons	DCX	[79]	[79]
smooth muscle cells	COL8A1	[80]	[81]
smooth muscle cells	ACTG2	[82]	[81]
smooth muscle cells	ACTA2	[80]	[81]
cardiac fibroblasts	TNNT2	[83]	[84]
cardiac fibroblasts	DCN	[85]	[86]
cardiac fibroblasts	HAND2	[83]	[87]
epithelial cells	CDH1	[88]	[89]
epithelial cells	EPCAM	[90]	[91]
epithelial cells	CLDN7	[92]	[93]
hepatocytes	TTR	[94]	[95]
hepatocytes	APOA1	[96]	[97]
hepatocytes	APOA2	[96]	[98]

3 | Discussion

Among its many applications, scRNA-seq has significant potential for comparative and evolutionary transcriptomics. However, meaningful cross-species analyses require more than just collecting single-cell data from different organisms; they depend on robust data quality, comparable cellular systems and computational tools to align data between species. This thesis contributes to establishing a foundation for cross-species analysis in primates by addressing some of these challenges. First, I examined data quality and technical artifacts in single-cell and single-nucleus RNA-seq by characterizing background noise and assessing correction methods to improve data reliability (Janssen et al. 2023). Additionally, I contributed to expanding available cellular resources and datasets by characterizing newly generated iPSC lines from multiple non-human primates (Geuder et al. 2021; Jocher et al. 2024a; Jocher et al. 2024b; Jocher et al. 2024c), as well as by curating a reference dataset for early primate differentiation in embryoid bodies (Jocher et al. 2025). Lastly, I addressed the challenge of assigning comparable cell types across species by developing a pipeline for orthologous cell type annotation and assessing the transferability of marker genes (Jocher et al. 2025). Beyond these broader contributions, this work has also led to specific insights that emerged across multiple projects. In the following sections, I will discuss these findings in more detail.

3.1 The power of genetic variants in transcriptomic experiments

In bulk or single-cell RNA-seq experiments the primary representation of the data is counts, summarizing RNA expression levels per gene and sample or per gene and cell. While sequencing is a crucial step in generating these data, the actual sequence of the reads is typically only used to determine their genomic position and convert them into count data, if they overlap with an annotated gene. This takes place early in data pre-processing, after which downstream analyses is centered around the resulting count matrix. So why should we still care about the sequence itself? On one hand, it can be used to examine transcript structure (Haas et al. 2017; Shen et al. 2014) and transcriptional dynamics (Qi and Battle 2024; Larsson et al. 2019). On the other hand, sequence differences reflect the genetic background of the species or individual, allowing reads to be traced back to their origin. As demonstrated in this thesis, this possibility to distinguish species and individuals based on sequence differences can have useful applications in the analysis of (single-cell) RNA-seq data.

3.1.1 Enhancing multiplexing in cross-species studies

First of all, it enables multiplexed experimental design with minimal experimental manipulation. Multiplexing, i.e. combining multiple samples in the same experiment, is a key strategy in scRNA-seq which makes it possible to increase throughput and have a nested study design in which batch effects between technical replicates can be accounted for (Tung et al. 2017). Typically, multiplexing relies on experimental labeling strategies to introduce a sample-specific barcode that can later be used for in silico demultiplexing (Cheng et al. 2021). These methods often involve additional handling steps, such as tagging common surface proteins with barcoded antibodies (Stoeckius et al. 2018) or labeling the cell membrane with lipid tags (McGinnis et al. 2019b). However, an alternative approach eliminates the need for experimental manipulation: when samples are genetically distinct, their inherent sequence

variation can serve as a natural barcode for demultiplexing.

Several computational tools have been developed to leverage genetic variants for demultiplexing scRNA-seq data, mainly with the goal to distinguish individuals from the same species (Kang et al. 2018; Heaton et al. 2020; Huang et al. 2019; Xu et al. 2019). These methods have proven to be highly effective, requiring only a relatively small number of variants for accurate classification. As species are even more biologically distinct than individuals, the same principle can readily be applied in cross-species studies. In this case, an alternative demultiplexing strategy involves mapping sequencing reads to a combined reference genome containing sequences from all species in the experiment and then assigning each cell based on alignment preference. This method is routinely used in experiments involving mixed human and mouse cells, for which a combined reference is readily available from Cell Ranger and commonly applied (Cheloni et al. 2021). However, particularly for closely related species demultiplexing based single-nucleotide variants mapped onto a single reference genome remains valuable, as demonstrated for instance in salamander species (Cardiello et al. 2023).

The choice of demultiplexing strategy thus depends on the specific combination of species involved and can become more complex when multiple species with varying phylogenetic distances, or even species and multiple individuals from the same species, are pooled together in the same experiment. For the four primate species that we multiplexed in each scRNA-seq experiment for our embryoid body study (Jocher et al. 2025), we therefore developed a stepwise approach, moving from broader distinctions to finer resolution. Initially, we distinguished human, orangutan and macaque cells (combining cynomolgus and rhesus macaques at this stage due to their high genetic similarity) by mapping the data to the human genome and identifying distinctive variants. Subsequently, we separated cells from different human individuals using a candidate set of variants and re-aligned macaque cells to the rhesus genome for further demultiplexing. To distinguish cynomolgus and rhesus macaques we first performed unsupervised demultiplexing based on genetic variants and then assigned the resulting donor groups to species based on the fraction of informative variants that aligned more strongly to either the rhesus or cynomolgus genome.

Although the exact demultiplexing strategy may need adjustment based on the exper-

imental design, in practice it is usually possible to confidently assign cells back to their original individual or species. An important part of this process is assigning labels to the demultiplexed cell groups. This can be made easier by using reference panels of informative variants, for example curated from earlier transcriptomic datasets. These panels help with interpretation and make the process more efficient, since only a subset of relevant variants needs to be considered. A practical side effect of genotype-based demultiplexing is that cross-species doublets can be readily identified. Multiplexing therefore represents a valuable approach for designing multi-species single-cell experiments, as it reduces bias and batch effects in downstream analyses without compromising data quality and requiring only minimal additional experimental effort for pooling samples.

3.1.2 Authentication of cell lines

We also used the identification of genetic variants from bulk and single-cell RNA-seq data for the authentication of non-human primate iPS cell lines (Jocher et al. 2024a; Jocher et al. 2024b; Jocher et al. 2024c). Being able to confirm the identity of a cell line in subsequent experiments is crucial, since cross-contamination and misidentification are not rare occurrences (Capes-Davis et al. 2010). The most widely used approaches for authentication include STR analysis, which is the standard for human pluripotent stem cell lines (Ludwig et al. 2023), and SNP profiling. STR analysis is well established for human and mouse cell lines, but much less so for other species, making SNP profiling an attractive alternative for cell line authentication (Fasterius et al. 2017). Moreover, creating a suitable STR panel requires species-specific marker selection and assay design, which is less automated than sequencing-based approaches. The most comprehensive methods for SNP genotyping are whole-genome sequencing (WGS) and whole-exome sequencing (WES), but both options remain relatively expensive. RNA-seq provides a more cost-effective alternative that still allows detection of up to 70% of expressed coding variants (Piskol et al. 2013). This has facilitated the use of RNA-seq based approaches for cell line authentication (Fasterius et al. 2017; Mohammad et al. 2019).

A major advantage of RNA-seq-based authentication is that transcriptomic data for cell lines are often available or generated as part of standard characterization efforts, making it

possible to use the same dataset to evaluate multiple aspects of cell identity. For example, we used scRNA-seq data from three rhesus iPSC lines not only for SNP-based authentication but also to assess their differentiation potential by analyzing cell-type composition in embryoid bodies (Jocher et al. 2024a). Similarly, we performed SNP profiling using bulk RNA-seq data from two baboon iPSC lines (Jocher et al. 2024b) and two vervet iPSC lines (Jocher et al. 2024c). If these lines are used in future transcriptomic studies, the lists of informative expressed variants we generated can be directly employed to validate their identity.

It should be noted here that both the prime-seq bulk RNA-seq protocol (Janjic et al. 2022) and the 10x Genomics scRNA-seq protocol used in these studies have a strong enrichment of reads towards the 3' end of transcripts. As a result, SNP detection is largely restricted to regions with high coverage near transcript ends. While this limits the breadth of genotyping compared to whole-transcript RNA-seq, it also offers practical advantages: concentrating coverage at the transcript ends improves the reliability of variant calls in those regions and ensures strong overlap between samples and similarly biased reference panels. Therefore, the power to perform simultaneous transcriptomic characterization and authentication makes cost-efficient bulk RNA-seq methods like prime-seq a valuable tool for the validation of cell lines, especially for non-model organisms for which more targeted authentication methods may not yet be established.

3.1.3 Genetic variants as natural barcodes in cell-mixing experiments

Finally, genetic variants can be used as natural barcodes in cell-mixing experiments to generate ground-truth datasets for the benchmarking of technical performance and computational methods. We applied this principle to study background noise in single-cell and single-nucleus RNA-seq experiments (Janssen et al. 2023).

When creating experimental ground-truth datasets, it is essential to strike a balance between simplification to isolate specific effects and maintaining a level of complexity that is representative for real-world scenarios. While in silico simulations can also be used to

produce ground-truth datasets, they are known to introduce artificial biases, especially when modeling complex data such as those from scRNA-seq experiments (Crowell et al. 2023). Thus, experimentally generated datasets remain indispensable.

In the case of background noise, the challenge of obtaining suitable experimental data becomes evident. The frequently used mixture of a human and mouse cell line presents a very simple and straightforward dataset to estimate the magnitude of background noise in the data (Macosko et al. 2015; Goldstein et al. 2017; Fleming et al. 2023). However, a major limitation of this setup is that the species difference largely prevents ambient RNA from one species (e.g. mouse) from contaminating the transcriptome of another (e.g. human). Since reads are typically aligned to a combined reference genome and assigned to the species with the best alignment, mouse-derived background RNA is not counted toward human gene expression, and vice versa. This separation reduces the realism of the noise, as it avoids the cross-cell-type contamination seen in same-species experiments. Additionally, the lack of cell type diversity in this dataset makes it difficult to draw conclusions about the impact on downstream analysis. This oversimplification also prevents a comprehensive evaluation of the performance of background noise removal methods. To address this, these methods are often additionally tested on more complex datasets with well-characterized cell types, such as peripheral blood mononuclear cells (PBMCs) (Yang et al. 2020; Young and Behjati 2020). In these datasets, exclusive marker genes are well-defined, and their expression in unrelated cell types is assumed to result from background noise. However, this approach does not provide a true ground truth, as it relies on assumptions about marker gene exclusivity rather than direct measurement of background noise. Addressing these limitations requires a dataset that not only captures the complexity of real scRNA-seq data but also provides a controlled framework for assessing noise levels and the effectiveness of computational correction methods.

To fill this gap, we developed a more nuanced experimental dataset by pooling kidney cells from three mouse strains representing two distinct subspecies (Janssen et al. 2023). Using subspecies instead of species made it possible to align the data to a single reference genome, providing a unified feature space. At the same time, the density of informative genetic variants was high enough to estimate background noise levels for individual cells,

rather than only in aggregate or across the entire dataset. Establishing a pipeline to identify and quantify background noise based on these variants enabled us to examine its sources and downstream impact in detail. Furthermore, the complex cell type structure and variable levels of background noise in this dataset allowed us to systematically evaluate three background noise removal methods, providing the first independent benchmark of this kind.

This analysis clearly demonstrated that more complex cell-mixing experiments have great potential for generating realistic benchmark datasets. While mixtures of human and mouse cell lines have been commonly used, pooling cells from more closely related species (or even subspecies in this case), combined with a diverse cell type composition, can avoid oversimplifications and better reflect real experimental scenarios. The dataset and analytical framework presented in Janssen et al. (2023) could be adapted for additional applications such as evaluating doublet detection and removal methods or examining feature-level effects of background noise. Moreover, similar experimental designs may be valuable for investigating technical artifacts across other single-cell modalities such as scATAC-seq.

3.2 Marker genes - fragile cornerstones of scRNA-seq analysis

The strength of scRNA-seq comes from the robustness of using the whole transcriptome to characterize single cells, however the use of individual marker genes remains indispensable for two main reasons: 1) Making high-level analysis interpretable and visualizable and 2) incorporating prior knowledge into the analysis that may be crucial for example to annotate cell types. At the same time, the dependency on a few selected genes makes the analysis prone to artifacts and misinterpretation. Their reliability is affected by both technical and biological factors.

3.2.1 Susceptibility to background noise

We found that marker gene detection is particularly strongly affected by background noise compared to other lines of downstream analysis like clustering or classification which are based on the whole transcriptome (Janssen et al. 2023). At high levels of background noise, we observed a strong decrease in the specificity of marker genes, reflected in lower log2-fold changes and higher detection rates in other cell types. Markers of abundant cell types are particularly susceptible to being detected widely across unrelated cell populations. For instance, markers that should specifically label proximal tubule cells, the most abundant cell type in our kidney datasets, were also detected in virtually all other cell types in samples with high background noise levels. This can lead to artificial combinations of marker genes and may result in incorrect cell type annotations.

Caglayan et al. (2022) illustrated this issue in brain single-nucleus data. They re-analysed a previously annotated dataset in which one cell population was identified as 'immature oligodendrocytes', marked by high expression of neuronal genes. Upon closer inspection Caglayan et al. (2022) found that markers of immature oligodendrocytes significantly overlapped with the most abundant ambient RNA markers. They concluded that this population most likely represents glial cells with high levels of background noise originating from neuronal cells, which have been misannotated in several previous studies. Similarly, Zhang et al. (2023b) also detected neuronal markers such as *Syt1* and *Grin2b* in microglia and other non-neuronal populations in snRNA-seq data from mice. In both studies, computational background correction with CellBender successfully removed the contaminating signals and resolved the misannotations.

How much background noise is present—and how strongly it impacts analysis—depends on both technical and biological aspects of the experiment. Single-nucleus RNA-seq data are especially prone to background contamination, as reflected in our own comparisons and noted in previous studies (Fleming et al. 2023; Caglayan et al. 2022; Zhang et al. 2023b). The extent of this impact also depends on the biological system. In our kidney dataset, where cell types are well separated, classification remained reliable even in the presence of moderate background noise. But in developmental datasets, where differences between cell

types are more subtle, background RNA tends to blur cell identities more strongly and can significantly affect annotation accuracy.

Taken together, these results underscore the importance of interpreting conclusions based on individual genes or marker combinations with caution. Wherever possible, such analyses should be complemented by transcriptome-wide approaches, larger gene sets or experimental validation to reduce the risk of misinterpreting technical artifacts. Furthermore, our benchmarking showed that computational tools for the removal of background noise can be highly effective in improving marker gene detection, with CellBender (Fleming et al. 2023) performing best in this task. Thus, including the extra step of background noise correction can substantially improve the reliability of cell type annotation.

3.2.2 Limited transferability across species

Marker genes are not only affected by technical noise, but also by biological variation. Our study on primate EBs (Jocher et al. 2025) shows that the transferability of marker genes decreases with increasing evolutionary distance between species. The specificity and ranking of marker genes may differ strongly between species for the same cell types and their discriminatory power is also reduced in the cross-species setting.

This is in line with previous observations that the expression breadth of a gene is linked to its conservation. Genes that are broadly expressed across multiple tissues tend to be more evolutionarily constrained and also show more conserved expression patterns between species (Cardoso-Moreira et al. 2019; Brawand et al. 2011). A similar pattern emerged in our analysis of primate EBs: genes with broad expression tended to show conserved patterns across species, whereas those with more cell type–specific expression were often more species-specific as well (Jocher et al. 2025). Marker genes, which are by definition cell-type specific, fall on the more divergent end of this spectrum. This helps to explain their limited transferability.

However, not all marker genes are equally affected, as differences in conservation also emerge between gene types. For example, genes encoding long non-coding RNAs (lncRNAs) are often very cell type-specific in their expression (Johnsson et al. 2014; Mattick et al. 2023) which makes them in principle good candidates as marker genes. Indeed, lncRNAs like

ESRG or *LINC00678* are among the most specific marker genes for human iPSCs (Lemmens et al. 2023). However, their low sequence conservation (Johnsson et al. 2014) and divergent expression patterns (Jocher et al. 2025) make lncRNAs less suitable as cross-species markers. In contrast, transcription factor (TF) markers show the highest concordance across species. This is in line with the concept of core regulatory complexes (CoRCs) proposed by Arendt et al. (2016), which define cell identities through conserved sets of TFs and are thought to underlie homologous cell types across species.

Overall, the transferability issue of marker genes has important implications for cross-species analysis. First of all, most resources for previously characterized marker genes are restricted to a few well-established model organisms. Large marker databases for single-cell studies like PanglaoDB (Franzén et al. 2019) and CellMarker (Zhang et al. 2019; Hu et al. 2023) focus on mouse and human data. As a result, it remains unclear which of these markers can be reliably used for other species. In such cases, markers shared between human and mouse may offer the most reliable option, but this narrows the pool of usable candidates considerably. Another important implication concerns cell type assignment across species. When marker genes vary across species, relying too heavily on them for annotation can lead to incorrect or inconsistent labels. This is particularly problematic when markers behave differently for orthologous cell types. This highlights the need for alternative strategies for identifying corresponding cell types across species that rely less on individual genes and more on transcriptome-wide expression patterns.

3.3 Cell type assignment across species

Not least because of the limited reliability of marker genes in cross-species contexts, comparing cell types between organisms is a delicate challenge. Both bulk and single-cell RNA-seq data can be used to characterize the transcriptome and assign cell types across species.

3.3.1 Classification of bulk RNA-seq data

Despite lacking single-cell resolution, cross-species cell type annotation can still be performed on bulk RNA-seq samples when relatively homogeneous cell populations are profiled. In our study on the generation of iPSCs from urinary cells of different primates (Geuder et al. 2021), we used reference-based classification to explore the identity of the primary cells and assess the outcome of reprogramming. By correlating the bulk RNA-seq data with a human reference using SingleR (Aran et al. 2019), both human and non-human primate urinary cells were classified as mesenchymal stem cells, epithelial cells, or smooth muscle cells, while all iPSC samples mapped clearly to iPSC or ESC profiles. This supported the successful reprogramming of the cells. Notably, the consistent results across species also showed that the classification worked reliably for non-human primates with a human reference.

In cases where the bulk samples are more heterogeneous, deconvolution methods using single-cell references make it possible to get an estimate of the cell type composition. In this context it can also be valuable to use reference data from another species to benefit from the availability of comprehensive data sets for well characterized model organisms. The cross-species performance of deconvolution depends on the phylogenetic distance. For example, one study found that deconvolution of rat kidney data using a mouse reference still produced good results (Wang et al. 2019). In contrast, a benchmark study showed that using a mouse reference to deconvolve human brain data consistently led to reduced performance across several methods (Sutton et al. 2022). This might again relate to the limited transferability of marker genes, as many deconvolution methods rely on cell type–specific genes to build reference signatures. If these markers do not behave consistently across species, deconvolution accuracy is likely to suffer.

3.3.2 Orthologous cell type assignment from scRNA-seq data

Single-cell transcriptomics offers the highest resolution to distinguish fine-grained cell types, but in cross-species studies this resolution also poses a challenge: corresponding cell types

must be carefully matched across species, despite differences in expression profiles and cellular composition. This is particularly relevant for datasets with substantial variation in cell type proportions and even the absence of certain populations in some species, as seen in our primate EB study (Jocher et al. 2025). In such heterogeneous settings, some approaches based on classification or data integration are not ideal as they either assume full overlap in cell types to one reference species or risk overcorrecting biological differences.

To address this, we developed a cluster-matching pipeline. Instead of aligning data at the single-cell level, we first clustered cells within each species and then used a classification-based approach to score the similarity between clusters across species. This approach has the advantage of preserving within-species structure and avoiding the strong assumptions required for integration or label transfer methods. It also accommodates cases where certain cell types are absent from one species, reducing the risk of forced or incorrect matches. A similar strategy has been applied in other recent studies facing complex cell type compositions across species (Jorstad et al. 2023; Suresh et al. 2023).

One limitation of this cluster-based matching approach is that it starts with discrete populations, which may lead to some ambiguity in continuous differentiation trajectories. Another important consideration is gene mapping across species. In our case, human gene annotations were transferred to the other primates, which is relatively straightforward due to close evolutionary proximity and helped compensate for gaps in the NHP genome annotations. However, this approach is less suitable for more distantly related species, where accounting for more complex gene correspondences is necessary (Tarashansky et al. 2021; Rosen et al. 2024).

Finally, while automated pipelines can support the annotation process, fully automatic cell type matching across species is not yet feasible. Manual curation remains essential for refining annotations and resolve ambiguities. To support this process, we developed an interactive Shiny app for our EB dataset (Jocher et al. 2025) that enables parameter tuning for the annotation workflow and exploration of marker gene expression. Such interactive tools, alongside careful expert evaluation and refinement, will continue to be a necessary part of cell type annotation workflows within and across species.

4 | Conclusion and Outlook

The synergy of scRNA-seq technologies with cross-species analysis holds great potential, but also comes with unique challenges. In this thesis, I aimed to address some of these, ranging from the validation of comparable primate cellular systems and the evaluation of scRNA-seq data quality to the refinement of methodological frameworks for comparative analysis. By systematically assessing background RNA contamination—a foundational yet often overlooked technical issue affecting all single-cell transcriptomic analyses—I highlighted its substantial impact on data interpretation and underscored the necessity for correction strategies. Additionally, I demonstrated that marker genes commonly used for cell type identification often lack robustness when transferred across species, emphasizing the need for cautious validation. I also addressed the methodological challenge of assigning cell types across species. Collectively, these findings help lay the groundwork for future evolutionary single-cell studies, particularly for comparative primate transcriptomics.

While this thesis focused on the technical and methodological foundations of cross-species single-cell analysis, the full potential of this approach lies in its application to evolutionary questions. Comparative studies allow us not only to detect differences, but to understand how cell types emerged, diversified, and specialized over time. Recent efforts to build comparative primate brain cell atlases have already uncovered both human-specific features and conserved patterns of cellular organization across species (Bakken et al. 2021; Jorstad et al. 2023; Suresh et al. 2023). Extending such studies to additional tissues and a wider range of species will provide a more comprehensive view of cell type evolution.

In this context, iPSCs and their derivatives have become increasingly valuable tools for comparative studies, especially for modeling early development across species. Even relatively simple systems, such as the embryoid bodies characterized in this work, provide

access to a diversity of cell types and species that would be difficult to study in vivo. As organoid models continue to increase in complexity and reproducibility, they will play an increasingly central role in modeling organ development and addressing functional questions in an experimentally controlled cross-species framework (Juan et al. 2023).

In developmental contexts, where cellular identities change gradually rather than discretely, aligning differentiation trajectories across species presents an additional challenge for analysis. While methods exist for aligning cell trajectories between different conditions (Alpert et al. 2018; Sugihara et al. 2022), cross-species comparisons are more complex due to asynchronous timing and the need to compare more than two trajectories simultaneously. A second major difficulty lies in comparing gene expression dynamics between species along these trajectories. Most current approaches to study dynamic gene expression focus on changes within a trajectory or contrast different trajectories (Van den Berge et al. 2020; Song and Li 2021), but are less suited for comparing how the same trajectory unfolds under different conditions—such as between species. Recent work by Sumanaweera et al. (2025) represents a step in this direction, as their Genes2Genes framework enables gene-level alignment of pseudotime trajectories. Moving forward, continued development of trajectory-based methods will be essential to enable meaningful cross-species comparisons of developmental programs.

Finally, when characterizing cell types across species, transcriptomic data alone provide only a partial view. Understanding how cell types evolve requires looking beyond gene expression to the regulatory mechanisms that control it by profiling chromatin accessibility, transcription factor activity, and epigenetic modifications. With the growing availability of single-cell assays for chromatin accessibility and other modalities, it is now possible to investigate these regulatory layers in a comparative cross-species framework. Integrating transcriptomic data with other layers through multimodal analysis will be essential for uncovering how regulatory programs are rewired across species.

Bibliography

10xgenomics.com (n.d.). *What is the maximum number of cells that can be profiled?* https://kb.10xgenomics.com/hc/en-us/articles/360001378811-What-is-the-maximum-number-of-cells-that-can-be-profiled. Accessed: 2025-4-6.

Abdelaal, Tamim, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J T Reinders, and Ahmed Mahfouz (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* 20.1, 194.

Aibar, Sara, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14.11, 1083–1086.

Almeida, Jamie L, Carolyn R Hill, and Kenneth D Cole (2011). Authentication of African green monkey cell lines using human short tandem repeat markers. *BMC Biotechnol.* 11.1, 102.

Alpert, Ayelet, Lindsay S Moore, Tania Dubovik, and Shai S Shen-Orr (2018). Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. Methods* 15.4, 267.

Alwine, J C, D J Kemp, and G R Stark (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. U. S. A.* 74.12, 5350–5354.

Amezquita, Robert A, Aaron T L Lun, Etienne Becht, Vince J Carey, Lindsay N Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte

Soneson, Levi Waldron, Hervé Pagès, Mike L Smith, Wolfgang Huber, Martin Morgan, Raphael Gottardo, and Stephanie C Hicks (2020). Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* 17.2, 137–145.

Anwised, Preeyanan, Ratree Moorawong, Worawalan Samruan, Sirilak Somredngan, Jittanun Srisutush, Chuti Laowtammathron, Irene Aksoy, Rangsun Parnpai, and Pierre Savatier (2023). An expedition in the jungle of pluripotent stem cells of non-human primates. *Stem Cell Reports* 18.11, 2016–2037.

Aran, Dvir, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, Atul J Butte, and Mallar Bhattacharya (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* 20.2, 163–172.

Arendt, Detlev, Paola Yanina Bertucci, Kaia Achim, and Jacob M Musser (2019). Evolution of neuronal types and families. *Curr. Opin. Neurobiol.* 56, 144–152.

Arendt, Detlev, Jacob M Musser, Clare V H Baker, Aviv Bergman, Connie Cepko, Douglas H Erwin, Mihaela Pavlicev, Gerhard Schlosser, Stefanie Widder, Manfred D Laubichler, and Günter P Wagner (2016). The origin and evolution of cell types. *Nat. Rev. Genet.* 17.12, 744–757.

Bais, Abha S and Dennis Kostka (2020). scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* 36.4, 1150–1158.

Bakken, Trygve E, Cindy Tj van Velthoven, Vilas Menon, Rebecca D Hodge, Zizhen Yao, Thuc Nghi Nguyen, Lucas T Graybuck, Gregory D Horwitz, Darren Bertagnolli, Jeff Goldy, Anna Marie Yanny, Emma Garren, Sheana Parry, Tamara Casper, Soraya I Shehata, Eliza R Barkan, Aaron Szafer, Boaz P Levi, Nick Dee, Kimberly A Smith, et al. (2021). Single-cell and single-nucleus RNA-seq uncovers shared and distinct axes of variation in dorsal LGN neurons in mice, non-human primates, and humans. *Elife* 10.

Barr, Kenneth A, Katherine L Rhodes, and Yoav Gilad (2023). The relationship between regulatory changes in cis and trans and the evolution of gene expression in humans and chimpanzees. *Genome Biol.* 24.1, 207.

Bernstein, Nicholas J, Nicole L Fong, Irene Lam, Margaret A Roy, David G Hendrickson, and David R Kelley (2020). Solo: Doublet identification in single-cell RNA-seq via semi-supervised deep learning. *Cell Syst.* 11.1, 95–101.e5.

Biharie, Kirti, Lieke Michielsen, Marcel J T Reinders, and Ahmed Mahfouz (2023). Cell type matching across species using protein embeddings and transfer learning. *Bioinformatics* 39.39 Suppl 1, i404–i412.

Bjornson-Hooper, Zachary B, Gabriela K Fragiadakis, Matthew H Spitzer, Han Chen, Deepthi Madhireddy, Kevin Hu, Kelly Lundsten, David R McIlwain, and Garry P Nolan (2022). A comprehensive atlas of immunological differences between humans, mice, and non-human primates. *Front. Immunol.* 13, 867015.

Bloom, Jesse D (2018). Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ* 6, e5578.

Brawand, David, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann (2011). The evolution of gene expression levels in mammalian organs. *Nature* 478.7369, 343–348.

Bridges, Kate and Kathryn Miller-Jensen (2022). Mapping and validation of scRNA-seq-derived cell-cell communication networks in the tumor microenvironment. *Front. Immunol.* 13, 885267.

Caglayan, Emre, Yuxiang Liu, and Genevieve Konopka (2022). Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron.*

Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, Cole Trapnell, and Jay Shendure (2019a). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 1.

Cao, Yue, Yingxin Lin, John T Ormerod, Pengyi Yang, Jean Y H Yang, and Kitty K Lo (2019b). scDC: single cell differential composition analysis. *BMC Bioinformatics* 20.Suppl 19, 721.

Capes-Davis, Amanda, George Theodosopoulos, Isobel Atkin, Hans G Drexler, Arihiro Kohara, Roderick A F MacLeod, John R Masters, Yukio Nakamura, Yvonne A Reid, Roger R Reddel, and R Ian Freshney (2010). Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int. J. Cancer* 127.1, 1–8.

Cardiello, Joseph F, Alberto Joven Araus, Sarantis Giatrellis, Clement Helsens, András Simon, and Nicholas D Leigh (2023). Evaluation of genetic demultiplexing of single-cell sequencing data from model species. *Life Sci. Alliance* 6.8, e202301979.

Cardoso-Moreira, Margarida, Jean Halbert, Delphine Valloton, Britta Velten, Chunyan Chen, Yi Shao, Angélica Liechti, Kelly Ascenção, Coralie Rummel, Svetlana Ovchinnikova, Pavel V Mazin, Ioannis Xenarios, Keith Harshman, Matthew Mort, David N Cooper, Carmen Sandi, Michael J Soares, Paula G Ferreira, Sandra Afonso, Miguel Carneiro, et al. (2019). Gene expression across mammalian organ development. *Nature* 571.7766, 505–509.

Cheloni, Stefano, Roman Hillje, Lucilla Luzi, Pier Giuseppe Pelicci, and Elena Gatti (2021). XenoCell: classification of cellular barcodes in single cell experiments from xenograft samples. *BMC Med. Genomics* 14.1, 34.

Cheng, Junyun, Jie Liao, Xin Shao, Xiaoyan Lu, and Xiaohui Fan (2021). Multiplexing methods for simultaneous large-scale transcriptomic profiling of samples at single-cell resolution. *Adv. Sci. (Weinh.)* 8.17, e2101229.

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437.7055, 69–87.

Clarke, Zoe A, Tallulah S Andrews, Jawairia Atif, Delaram Pouyabahar, Brendan T Innes, Sonya A MacParland, and Gary D Bader (2021). Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat. Protoc.* 16.6, 2749–2764.

Clevers, Hans (2016). Modeling development and disease with organoids. *Cell* 165.7, 1586–1597.

Crow, Megan, Anirban Paul, Sara Ballouz, Z Josh Huang, and Jesse Gillis (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* 9.1, 884.

Crowell, Helena L, Sarah X Morillo Leonardo, Charlotte Soneson, and Mark D Robinson (2023). The shaky foundations of simulating single-cell RNA sequencing data. *Genome Biol.* 24.1, 62.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, et al. (2025). CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.* 53.D1, D886–D900.

DePasquale, Erica A K, Daniel J Schnell, Pieter-Jan Van Camp, Íñigo Valiente-Alandí, Burns C Blaxall, H Leighton Grimes, Harinder Singh, and Nathan Salomonis (2019). DoubletDecon: Deconvoluting doublets from single-cell RNA-sequencing data. *Cell Rep.* 29.6, 1718–1727.e8.

Ding, Jiarui, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T Nguyen, John Y H Kwon, Boaz Barak, William Ge, Amanda J Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K Shalek, Alexandra-Chloé Villani, et al. (2020). Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* 38.6, 737–746.

Dixit, Atray (2016). Correcting chimeric crosstalk in single cell RNA-seq experiments. *bioRxiv*, 093237.

Efremova, Mirjana, Miquel Vento-Tormo, Sarah A Teichmann, and Roser Vento-Tormo (2020). CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* 15.4, 1484–1506.

Enard, Wolfgang (2012). Functional primate genomics–leveraging the medical potential. *J. Mol. Med.* 90.5, 471–480.

Ergen, Can, Galen Xing, Chenling Xu, Martin Kim, Michael Jayasuriya, Erin McGeever, Angela Oliveira Pisco, Aaron Streets, and Nir Yosef (2024). Consensus prediction of cell type labels in single-cell data with popV. *Nat. Genet.* 56.12, 2731–2738.

Fasterius, Erik, Cinzia Raso, Susan Kennedy, Nora Rauch, Pär Lundin, Walter Kolch, Mathias Uhlén, and Cristina Al-Khalili Szigyarto (2017). A novel RNA sequencing data analysis method for cell line authentication. *PLoS One* 12.2, e0171435.

Finak, Greg, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, Peter S Linsley, and Raphael Gottardo (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16.1, 1–13.

Fischer, Jan, Eduardo Fernández Ortuño, Fabio Marsoner, Annasara Artioli, Jula Peters, Takashi Namba, Christina Eugster Oegema, Wieland B Huttner, Julia Ladewig, and Michael Heide (2022). Human-specific ARHGAP11B ensures human-like basal progenitor levels in hominid cerebral organoids. *EMBO Rep.* 23.11, e54728.

Fischer, Stephan and Jesse Gillis (2021). How many markers are needed to robustly determine a cell's type? *iScience* 24.11, 103292.

Fleming, Stephen J, Mark D Chaffin, Alessandro Arduini, Amer-Denis Akkad, Eric Banks, John C Marioni, Anthony A Philippakis, Patrick T Ellinor, and Mehrtash Babadi (2023). Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat. Methods* 20.9, 1323–1335.

Franzén, Oscar, Li-Ming Gan, and Johan L M Björkegren (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* 2019, baz046.

Gao, Xin, Deqing Hu, Madelaine Gogol, and Hua Li (2019). ClusterMap: compare multiple single cell RNA-Seq datasets across different experimental conditions. *Bioinformatics* 35.17, 3038–3045.

Germain, Pierre-Luc, Aaron Lun, Carlos Garcia Meixide, Will Macnair, and Mark D Robinson (2021). Doublet identification in single-cell sequencing data using scDblFinder. *F1000Res.* 10, 979.

Geuder, Johanna, Lucas E Wange, Aleksandar Janjic, Jessica Radmer, Philipp Janssen, Johannes W Bagnoli, Stefan Müller, Artur Kaul, Mari Ohnuki, and Wolfgang Enard

(2021). A non-invasive method to generate induced pluripotent stem cells from primate urine. *Sci. Rep.* 11.1, 3516.

Goldstein, Leonard D, Ying-Jiun Jasmine Chen, Jude Dunne, Alain Mir, Hermann Hubschle, Joseph Guillory, Wenlin Yuan, Jingli Zhang, Jeremy Stinson, Bijay Jaiswal, Kanika Bajaj Pahuja, Ishminder Mann, Thomas Schaal, Leo Chan, Sangeetha Anandakrishnan, Chun-Wah Lin, Patricio Espinoza, Syed Husain, Harris Shapiro, Karthikeyan Swaminathan, et al. (2017). Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics* 18.1, 519.

González-Silva, Laura, Laura Quevedo, and Ignacio Varela (2020). Tumor functional heterogeneity unraveled by scRNA-seq technologies. *Trends Cancer* 6.1, 13–19.

González-Velasco, Óscar, Malte Simon, Rüstem Yilmaz, Rosanna Parlato, Jochen Weishaupt, Charles D Imbusch, and Benedikt Brors (2024). Identifying similar populations across independent single cell studies without data integration. *bioRxiv*, 2024.09.27.615367.

Haas, Brian J, Alex Dobin, Nicolas Stransky, Bo Li, Xiao Yang, Timothy Tickle, Asma Bankapur, Carrie Ganote, Thomas G Doak, Nathalie Pochet, Jing Sun, Catherine J Wu, Thomas R Gingeras, and Aviv Regev (2017). STAR-fusion: Fast and accurate fusion transcript detection from RNA-Seq. *bioRxiv*, 120295.

Haghverdi, Laleh, Aaron T L Lun, Michael D Morgan, and John C Marioni (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36.5, 421–427.

Han, Xiaoping, Haide Chen, Daosheng Huang, Huidong Chen, Lijiang Fei, Chen Cheng, He Huang, Guo-Cheng Yuan, and Guoji Guo (2018). Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biol.* 19.1, 47.

Heaton, Haynes, Arthur M Talman, Andrew Knights, Maria Imaz, Daniel J Gaffney, Richard Durbin, Martin Hemberg, and Mara K N Lawniczak (2020). Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* 17.6, 615–620.

Heid, C A, J Stevens, K J Livak, and P M Williams (1996). Real time quantitative PCR. *Genome Res.* 6.10, 986–994.

Hicks, Stephanie C, F William Townes, Mingxiang Teng, and Rafael A Irizarry (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19.4, 562–578.

Housman, Genevieve, Emilie Briscoe, and Yoav Gilad (2022). Evolutionary insights into primate skeletal gene regulation using a comparative cell culture model. *PLoS Genet.* 18.3, e1010073.

Housman, Genevieve and Yoav Gilad (2020). Prime time for primate functional genomics. *Curr. Opin. Genet. Dev.* 62, 1–7.

Hu, Congxue, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, Xia Li, Peng Wang, and Yunpeng Zhang (2023). CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res.* 51.D1, D870–D876.

Huang, Yuanhua, Davis J McCarthy, and Oliver Stegle (2019). Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* 20.1, 273.

Ilicic, Tomislav, Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C Marioni, and Sarah A Teichmann (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 17.1, 29.

Islam, Saiful, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21.7, 1160–1167.

Islam, Saiful, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11.2, 163–166.

Janjic, Aleksandar, Lucas E Wange, Johannes W Bagnoli, Johanna Geuder, Phong Nguyen, Daniel Richter, Beate Vieth, Binje Vick, Irmela Jeremias, Christoph Ziegenhain, Ines Hellmann, and Wolfgang Enard (2022). Prime-seq, efficient and powerful bulk RNA sequencing. *Genome Biol.* 23.1, 88.

Janssen, Philipp, Zane Kliesmete, Beate Vieth, Xian Adiconis, Sean Simmons, Jamie Marshall, Cristin McCabe, Holger Heyn, Joshua Z Levin, Wolfgang Enard, and Ines Hellmann

(2023). The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biol.* 24.1, 140.

Jocher, Jessica, Fiona C Edenhofer, Philipp Janssen, Stefan Müller, Dana C Lopez-Parra, Johanna Geuder, and Wolfgang Enard (2024a). Generation and characterization of three fibroblast-derived Rhesus Macaque induced pluripotent stem cell lines. *Stem Cell Res.* 74.103277, 103277.

Jocher, Jessica, Fiona C Edenhofer, Stefan Müller, Philipp Janssen, Eva Briem, Johanna Geuder, and Wolfgang Enard (2024b). Generation and characterization of two fibroblast-derived Baboon induced pluripotent stem cell lines. *Stem Cell Res.* 75.103316, 103316.

— (2024c). Generation and characterization of two Vervet monkey induced pluripotent stem cell lines derived from fibroblasts. *Stem Cell Res.* 75.103315, 103315.

Jocher, Jessica, Philipp Janssen, Beate Vieth, Fiona C Edenhofer, Tamina Dietl, Anita Térmeg, Paulina Spurk, Johanna Geuder, Wolfgang Enard, and Ines Hellmann (2025). Identification and comparison of orthologous cell types from primate embryoid bodies shows limits of marker gene transferability.

Johnsson, Per, Leonard Lipovich, Dan Grandér, and Kevin V Morris (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim. Biophys. Acta* 1840.3, 1063–1071.

Jorstad, Nikolas L, Janet H T Song, David Exposito-Alonso, Hamsini Suresh, Nathan Castro-Pacheco, Fenna M Krienen, Anna Marie Yanny, Jennie Close, Emily Gelfand, Brian Long, Stephanie C Seeman, Kyle J Travaglini, Soumyadeep Basu, Marc Beaudin, Darren Bertagnolli, Megan Crow, Song-Lin Ding, Jeroen Eggermont, Alexandra Glandon, Jeff Goldy, et al. (2023). Comparative transcriptomics reveals human-specific cortical features. *Science* 382.6667, eade9516.

Jovic, Dragomirka, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.* 12.3, e694.

Juan, David, Gabriel Santpere, Joanna L Kelley, Omar E Cornejo, and Tomas Marques-Bonet (2023). Current advances in primate genomics: novel approaches for understanding evolution and disease. *Nat. Rev. Genet.*

Kang, Hyun Min, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell, and Chun Jimmie Ye (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36.1, 89–94.

Kanthaswamy, Sreetharan, Andrea von Dollen, Jennifer D Kurushima, Ona Alminas, Jeffrey Rogers, Betsy Ferguson, Nicholas W Lerche, Philip C Allen, and David Glenn Smith (2006). Microsatellite markers for standardized genetic management of captive colonies of rhesus macaques (Macaca mulatta). *Am. J. Primatol.* 68.1, 73–95.

Kanton, Sabina, Michael James Boyle, Zhisong He, Malgorzata Santel, Anne Weigert, Fátima Sanchís-Calleja, Patricia Guijarro, Leila Sidow, Jonas Simon Fleck, Dingding Han, Zhengzong Qian, Michael Heide, Wieland B Huttner, Philipp Khaitovich, Svante Pääbo, Barbara Treutlein, and J Gray Camp (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* 574.7778, 418–422.

Khaitovich, Philipp, Wolfgang Enard, Michael Lachmann, and Svante Pääbo (2006). Evolution of primate gene expression. *Nat. Rev. Genet.* 7.9, 693.

Kim, Ik Soo, Jingyi Wu, Gilbert J Rahme, Sofia Battaglia, Atray Dixit, Elizabeth Gaskell, Huidong Chen, Luca Pinello, and Bradley E Bernstein (2020). Parallel single-cell RNA-seq and genetic recording reveals lineage decisions in developing embryoid bodies. *Cell Rep.* 33.1, 108222.

Kolodziejczyk, Aleksandra A, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58.4, 610–620.

Korsunsky, Ilya, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* 16.12, 1289–1296.

Krienen, Fenna M, Melissa Goldman, Qiangge Zhang, Ricardo C H Del Rosario, Marta Florio, Robert Machold, Arpiar Saunders, Kirsten Levandowski, Heather Zaniewski, Benjamin Schuman, Carolyn Wu, Alyssa Lutservitz, Christopher D Mullally, Nora Reed,

Elizabeth Bien, Laura Bortolin, Marian Fernandez-Otero, Jessica D Lin, Alec Wysoker, James Nemesh, et al. (2020). Innovations present in the primate interneuron repertoire. *Nature* 586.7828, 262–269.

Kuleshov, Maxim V, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, Michael G McDermott, Caroline D Monteiro, Gregory W Gundersen, and Avi Ma'ayan (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44.W1, W90–7.

Lake, Blue B, Rizi Ai, Gwendolyn E Kaeser, Neeraj S Salathia, Yun C Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Mostafa Ronaghi, Jian-Bing Fan, Wei Wang, Jerold Chun, et al. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* 352.6293, 1586–1590.

Larsson, Anton J M, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M Rivera, Bing Ren, and Rickard Sandberg (2019). Genomic encoding of transcriptional burst kinetics. *Nature* 565.7738, 251–254.

Lemmens, Myriam, Juliane Perner, Leon Potgeter, Michael Zogg, Sineha Thiruchelvam, Matthias Müller, Thierry Doll, Annick Werner, Yoann Gilbart, Philippe Couttet, Hans-Jörg Martus, and Silvana Libertini (2023). Identification of marker genes to monitor residual iPSCs in iPSC-derived products. *Cytotherapy* 25.1, 59–67.

Liao, Mingfeng, Yang Liu, Jing Yuan, Yanling Wen, Gang Xu, Juanjuan Zhao, Lin Cheng, Jinxiu Li, Xin Wang, Fuxiang Wang, Lei Liu, Ido Amit, Shuye Zhang, and Zheng Zhang (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* 26.6, 842–844.

Lindblad-Toh, Kerstin, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, Lucas D Ward, Craig B Lowe, Alisha K Holloway, Michele Clamp, Sante Gnerre, Jessica Alföldi, Kathryn Beal, Jean Chang, Hiram Clawson, James Cuff, et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478.7370, 476–482.

Liu, Angela, Beverly Peng, Ajith V Pankajam, Thu Elizabeth Duong, Gloria Pryhuber, Richard H Scheuermann, and Yun Zhang (2024). Discovery of optimal cell type classification marker genes from single cell RNA sequencing data. *BMC Methods* 1.1, 1–20.

Liu, Haisong, Fangfang Zhu, Jun Yong, Pengbo Zhang, Pingping Hou, Honggang Li, Wei Jiang, Jun Cai, Meng Liu, Kai Cui, Xiuxia Qu, Tingting Xiang, Danyu Lu, Xiaochun Chi, Ge Gao, Weizhi Ji, Mingxiao Ding, and Hongkui Deng (2008). Generation of induced pluripotent stem cells from adult rhesus monkey fibroblasts. *Cell Stem Cell* 3.6, 587–590.

Liu, Xingyan, Qunlun Shen, and Shihua Zhang (2023). Cross-species cell-type assignment from single-cell RNA-seq data by a heterogeneous graph neural network. *Genome Res.* 33.1, 96–111.

Lotfollahi, Mohammad, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, Sergei Rybakov, Alexander V Misharin, and Fabian J Theis (2022). Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* 40.1, 121–130.

Lowe, Rohan, Neil Shirley, Mark Bleackley, Stephen Dolan, and Thomas Shafee (2017). Transcriptomics technologies. *PLoS Comput. Biol.* 13.5, e1005457.

Ludwig, Tenneille E, Peter W Andrews, Ivana Barbaric, Nissim Benvenisty, Anita Bhattacharyya, Jeremy M Crook, Laurence M Daheron, Jonathan S Draper, Lyn E Healy, Meritxell Huch, Maneesha S Inamdar, Kim B Jensen, Andreas Kurtz, Madeline A Lancaster, Prisca Liberali, Matthias P Lutolf, Christine L Mummery, Martin F Pera, Yoji Sato, Noriko Shimasaki, et al. (2023). ISSCR standards for the use of human stem cells in basic research. *Stem Cell Reports* 18.9, 1744–1752.

Luecken, M D, M Büttner, K Chaichoompu, A Danese, M Interlandi, M F Mueller, D C Strobl, L Zappia, M Dugas, M Colomé-Tatché, and F J Theis (2020). Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv*, 2020.05.22.111161.

Luecken, Malte D and Fabian J Theis (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15.6, e8746.

Lyck, Lise, Ishar Dalmau, John Chemnitz, Bente Finsen, and Henrik Daa Schrøder (2008). Immunohistochemical markers for quantitative studies of neurons and glia in human neocortex. *J. Histochem. Cytochem.* 56.3, 201–221.

Ma, Shaojie, Mario Skarica, Qian Li, Chuan Xu, Ryan D Risgaard, Andrew T N Tebbenkamp, Xoel Mato-Blanco, Rothem Kovner, Željka Krsnik, Xabier de Martin, Victor Luria, Xavier Martí-Pérez, Dan Liang, Amir Karger, Danielle K Schmidt, Zachary Gomez-Sanchez, Cai Qi, Kevin T Gobeske, Sirisha Pochareddy, Ashwin Debnath, et al. (2022). Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. *Science* 377.6614, eabo7257.

Macosko, Evan Z, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161.5, 1202–1214.

Maecker, Holden T, J Philip McCoy, and Robert Nussenblatt (2012). Standardizing immunophenotyping for the Human Immunology Project. *Nat. Rev. Immunol.* 12.3, 191–200.

Mattick, John S, Paulo P Amaral, Piero Carninci, Susan Carpenter, Howard Y Chang, Ling-Ling Chen, Runsheng Chen, Caroline Dean, Marcel E Dinger, Katherine A Fitzgerald, Thomas R Gingeras, Mitchell Guttman, Tetsuro Hirose, Maite Huarte, Rory Johnson, Chandrasekhar Kanduri, Philipp Kapranov, Jeanne B Lawrence, Jeannie T Lee, Joshua T Mendell, et al. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.* 24.6, 430–447.

McGinnis, Christopher S, Lyndsay M Murrow, and Zev J Gartner (2019a). DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* 8.4, 329–337.e4.

McGinnis, Christopher S, David M Patterson, Juliane Winkler, Daniel N Conrad, Marco Y Hein, Vasudha Srivastava, Jennifer L Hu, Lyndsay M Murrow, Jonathan S Weissman, Zena Werb, Eric D Chow, and Zev J Gartner (2019b). MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* 16.7, 619–626.

Mohammad, Tabrez A, Yun S Tsai, Safwa Ameer, Hung-I Harry Chen, Yu-Chiao Chiu, and Yidong Chen (2019). CeL-ID: cell line identification using RNA-seq data. *BMC Genomics* 20.Suppl 1, 81.

Moon, Kevin R, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, Natalia B Ivanova, Guy Wolf, and Smita Krishnaswamy (2019). Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* 37.12, 1482–1492.

Mora-Bermúdez, Felipe, Farhath Badsha, Sabina Kanton, J Gray Camp, Benjamin Vernot, Kathrin Köhler, Birger Voigt, Keisuke Okita, Tomislav Maricic, Zhisong He, Robert Lachmann, Svante Pääbo, Barbara Treutlein, and Wieland B Huttner (2016). Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *Elife* 5.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5.7, 621–628.

Nelson, M E, S G Riva, and A Cvejic (2022). SMaSH: a scalable, general marker gene identification framework for single-cell RNA-sequencing. *BMC Bioinformatics* 23.1, 328.

Noureen, Nighat, Zhenqing Ye, Yidong Chen, Xiaojing Wang, and Siyuan Zheng (2022). Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. *Elife* 11.

Parekh, Swati, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann (2018). zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* 7.6, giy059.

Park, Youngjun, Nils P Muttray, and Anne-Christin Hauschild (2024). Species-agnostic transfer learning for cross-species transcriptomics data integration without gene orthology. *Brief. Bioinform.* 25.2, bbae004.

Pasquini, Giovanni, Jesus Eduardo Rojo Arias, Patrick Schäfer, and Volker Busskamp (2021). Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* 19, 961–969.

Picelli, Simone, Omid R Faridani, Asa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* 9.1, 171–181.

Piskol, Robert, Gokul Ramaswami, and Jin Billy Li (2013). Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* 93.4, 641–651.

Pliner, Hannah A, Jay Shendure, and Cole Trapnell (2019). Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* 16.10, 983–986.

Pollen, Alex A, Aparna Bhaduri, Madeline G Andrews, Tomasz J Nowakowski, Olivia S Meyerson, Mohammed A Mostajo-Radji, Elizabeth Di Lullo, Beatriz Alvarado, Melanie Bedolli, Max L Dougherty, Ian T Fiddes, Zev N Kronenberg, Joe Shuga, Anne A Leyrat, Jay A West, Marina Bershteyn, Craig B Lowe, Bryan J Pavlovic, Sofie R Salama, David Haussler, et al. (2019). Establishing cerebral organoids as models of human-specific brain evolution. *Cell* 176.4, 743–756.e17.

Pollen, Alex A, Umut Kilik, Craig B Lowe, and J Gray Camp (2023). Human-specific genetics: new tools to explore the molecular and cellular basis of human evolution. *Nat. Rev. Genet.*, 1–25.

Prabhakar, Shyam, James P Noonan, Svante Pääbo, and Edward M Rubin (2006). Accelerated evolution of conserved noncoding sequences in humans. *Science* 314.5800, 786.

Pullin, Jeffrey M and Davis J McCarthy (2024). A comparison of marker gene selection methods for single-cell RNA sequencing data. *Genome Biol.* 25.1, 56.

Qi, Guanghao and Alexis Battle (2024). Computational methods for allele-specific expression in single cells. *Trends Genet.* 40.11, 939–949.

Raab, Stefanie, Moritz Klingenstein, Stefan Liebau, and Leonhard Linta (2014). A comparative view on human somatic cell sources for iPSC generation. *Stem Cells Int.* 2014.1, 768391.

Regev, Aviv, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, et al. (2017). The Human Cell Atlas. *Elife* 6.

Ren, Xianwen, Lei Zhang, Yuanyuan Zhang, Ziyi Li, Nathan Siemers, and Zemin Zhang (2021). Insights gained from single-cell analysis of immune cells in the tumor microenvironment. *Annu. Rev. Immunol.* 39.1, 583–609.

Rhodes, Katherine, Kenneth A Barr, Joshua M Popp, Benjamin J Strober, Alexis Battle, and Yoav Gilad (2022). Human embryoid bodies as a novel system for genomic studies of functionally diverse cell types. *Elife* 11.

Rogers, Jeffrey and Richard A Gibbs (2014). Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat. Rev. Genet.* 15.5, 347–359.

Rosen, Yanay, Maria Brbić, Yusuf Roohani, Kyle Swanson, Ziang Li, and Jure Leskovec (2024). Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nat. Methods* 21.8, 1492–1500.

Rosenberg, Alexander B, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, Suzie H Pun, Drew L Sellers, Bosiljka Tasic, and Georg Seelig (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360.6385, 176–182.

Schena, M, D Shalon, R W Davis, and P O Brown (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270.5235, 467–470.

Schiebout, Courtney, Hannah Lust, Yina Huang, and H Robert Frost (2023). Cell type-specific interaction analysis using doublets in scRNA-seq. *Bioinform. Adv.* 3.1, vbad120.

scRNA-tools.org (n.d.). *A catalogue of tools for scRNA-seq analysis*. `https://www.scrna-tools.org/`. Accessed: 2025-4-6.

Shafer, Maxwell E R (2019). Cross-species analysis of single-cell transcriptomic data. *Front. Cell Dev. Biol.* 7, 175.

Shen, Shihao, Juw Won Park, Zhi-Xiang Lu, Lan Lin, Michael D Henry, Ying Nian Wu, Qing Zhou, and Yi Xing (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* 111.51, E5593–601.

Singh, Abhishek, Mukesh Thakur, Vivek Sahajpal, Sujeet K Singh, Kailash Chandra, Arun Sharma, Nisha Devi, and Ausma Bernot (2019). Cross-species validation of human

specific STR system, SureID® 21G and SureID® 23comp (Health Gene Technologies) in Chimpanzee (Pan Troglodytes). *BMC Res. Notes* 12.1, 750.

Song, Dongyuan and Jingyi Jessica Li (2021). PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data. *Genome Biol.* 22.1, 124.

Song, Yuyao, Zhichao Miao, Alvis Brazma, and Irene Papatheodorou (2023). Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nat. Commun.* 14.1, 6495.

Spangler, Abby, Emily Y Su, April M Craft, and Patrick Cahan (2018). A single cell transcriptional portrait of embryoid body differentiation and comparison to progenitors of the developing embryo. *Stem Cell Res.* 31, 201–215.

Squair, Jordan W, Matthieu Gautier, Claudia Kathe, Mark A Anderson, Nicholas D James, Thomas H Hutson, Rémi Hudelle, Taha Qaiser, Kaya J E Matson, Quentin Barraud, Ariel J Levine, Gioele La Manno, Michael A Skinnider, and Grégoire Courtine (2021). Confronting false discoveries in single-cell differential expression. *Nat. Commun.* 12.1, 5692.

Stoeckius, Marlon, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14.9, 865–868.

Stoeckius, Marlon, Shiwei Zheng, Brian Houck-Loomis, Stephanie Hao, Bertrand Z Yeung, William M Mauck 3rd, Peter Smibert, and Rahul Satija (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* 19.1, 224.

Street, Kelly, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19.1, 477.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija (2019). Comprehensive Integration of Single-Cell Data. *Cell* 177.7, 1888–1902.e21.

Sugihara, Reiichi, Yuki Kato, Tomoya Mori, and Yukio Kawahara (2022). Alignment of single-cell trajectory trees with CAPITAL. *Nat. Commun.* 13.1, 5972.

Sumanaweera, Dinithi, Chenqu Suo, Ana-Maria Cujba, Daniele Muraro, Emma Dann, Krzysztof Polanski, Alexander S Steemers, Woochan Lee, Amanda J Oliver, Jong-Eun Park, Kerstin B Meyer, Bianca Dumitrascu, and Sarah A Teichmann (2025). Gene-level alignment of single-cell trajectories. *Nat. Methods* 22.1, 68–81.

Sun, Xiaobo, Xiaochu Lin, Ziyi Li, and Hao Wu (2022). A comprehensive comparison of supervised and unsupervised methods for cell type identification in single-cell RNA-seq. *Brief. Bioinform.* 23.2, bbab567.

Suresh, Hamsini, Megan Crow, Nikolas Jorstad, Rebecca Hodge, Ed Lein, Alexander Dobin, Trygve Bakken, and Jesse Gillis (2023). Comparative single-cell transcriptomic analysis of primate brains highlights human-specific regulatory evolution. *Nat Ecol Evol*.

Sutton, Gavin J, Daniel Poppe, Rebecca K Simmons, Kieran Walsh, Urwah Nawaz, Ryan Lister, Johann A Gagnon-Bartsch, and Irina Voineagu (2022). Comprehensive evaluation of deconvolution methods for human brain gene expression. *Nat. Commun.* 13.1, 1358.

Svensson, Valentine, Roser Vento-Tormo, and Sarah A Teichmann (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13.4, 599–604.

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, and Principal investigators (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562.7727, 367–372.

Takahashi, Kazutoshi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131.5, 861–872.

Takahashi, Kazutoshi and Shinya Yamanaka (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126.4, 663–676.

Tan, Yuqi and Patrick Cahan (2019). SingleCellNet: A computational tool to classify single cell RNA-seq data across platforms and across species. *Cell Syst.* 9.2, 207–213.e2.

Tanay, Amos and Arnau Sebé-Pedrós (2021). Evolutionary cell type mapping with single-cell genomics. *Trends Genet.* 37.10, 919–932.

Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6.5, 377–382.

Tarashansky, Alexander J, Jacob M Musser, Margarita Khariton, Pengyang Li, Detlev Arendt, Stephen R Quake, and Bo Wang (2021). Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* 10.

Tasic, Bosiljka (2018). Single cell transcriptomics in neuroscience: cell classification and beyond. *Curr. Opin. Neurobiol.* 50, 242–249.

Thomson, J A, J Itskovitz-Eldor, S S Shapiro, M A Waknitz, J J Swiergiel, V S Marshall, and J M Jones (1998). Embryonic stem cell lines derived from human blastocysts. *Science* 282.5391, 1145–1147.

Thomson, J A, J Kalishman, T G Golos, M Durning, C P Harris, R A Becker, and J P Hearn (1995). Isolation of a primate embryonic stem cell line. *Proc. Natl. Acad. Sci. U. S. A.* 92.17, 7844–7848.

Tian, Luyi, Xueyi Dong, Saskia Freytag, Kim-Anh Lê Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S Weber, Azadeh Seidi, Jafar S Jabbari, Shalin H Naik, and Matthew E Ritchie (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* 16.6, 479–487.

Trapnell, Cole (2015). Defining cell types and states with single-cell genomics. *Genome Res.* 25.10, 1491–1498.

Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32.4, 381–386.

Tung, Po-Yuan, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad (2017). Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* 7.1, 39921.

Van den Berge, Koen, Hector Roux de Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* 11.1, 1201.

Vargo, Alexander H S and Anna C Gilbert (2020). A rank-based marker selection method for high throughput scRNA-seq data. *BMC Bioinformatics* 21.1, 477.

Vieth, Beate, Swati Parekh, Christoph Ziegenhain, Wolfgang Enard, and Ines Hellmann (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* 10.1, 4667.

Wang, Xuran, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* 10.1, 380.

Wang, Ziwei, Hui Ding, and Quan Zou (2020). Identifying cell types to interpret scRNA-seq data: how, why and more possibilities. *Brief. Funct. Genomics* 19.4, 286–291.

Weber, Leah L, P Sashittal, and M El-Kebir (2021). doubletD: detecting doublets in single-cell DNA sequencing data. *Bioinformatics* 37, i214–i221.

Wolock, Samuel L, Romain Lopez, and Allon M Klein (2019). Scrublet: Computational identification of cell Doublets in Single-cell transcriptomic data. *Cell Syst.* 8.4, 281–291.e9.

Wu, Fengying, Jue Fan, Yayi He, Anwen Xiong, Jia Yu, Yixin Li, Yan Zhang, Wencheng Zhao, Fei Zhou, Wei Li, Jie Zhang, Xiaosheng Zhang, Meng Qiao, Guanghui Gao, Shanhao Chen, Xiaoxia Chen, Xuefei Li, Likun Hou, Chunyan Wu, Chunxia Su, et al. (2021). Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat. Commun.* 12.1, 2540.

Wunderlich, Stephanie, Martin Kircher, Beate Vieth, Alexandra Haase, Sylvia Merkert, Jennifer Beier, Gudrun Göhring, Silke Glage, Axel Schambach, Eliza C Curnow, Svante Pääbo, Ulrich Martin, and Wolfgang Enard (2014). Primate iPS cells as tools for evolutionary analyses. *Stem Cell Res.* 12.3, 622–629.

Xi, Nan Miles and Jingyi Jessica Li (2021). Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell Syst.* 12.2, 176–194.e6.

Xu, Jun, Caitlin Falconer, Quan Nguyen, Joanna Crawford, Brett D McKinnon, Sally Mortlock, Anne Senabouth, Stacey Andersen, Han Sheng Chiu, Longda Jiang, Nathan J Palpant, Jian Yang, Michael D Mueller, Alex W Hewitt, Alice Pébay, Grant W Montgomery, Joseph E Powell, and Lachlan J M Coin (2019). Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biol.* 20.1, 290.

Yang, Guang, Hyenjong Hong, April Torres, Kristen E Malloy, Gourav R Choudhury, Jeffrey Kim, and Marcel M Daadi (2018). Standards for deriving nonhuman primate-induced pluripotent stem cells, neural stem cells and dopaminergic lineage. *Int. J. Mol. Sci.* 19.9, 2788.

Yang, Shiyi, Sean E Corbett, Yusuke Koga, Zhe Wang, W Evan Johnson, Masanao Yajima, and Joshua D Campbell (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* 21.1, 57.

Young, Matthew D and Sam Behjati (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* 9.12.

Yu, Junying, Maxim A Vodyanik, Kim Smuga-Otto, Jessica Antosiewicz-Bourget, Jennifer L Frane, Shulan Tian, Jeff Nie, Gudrun A Jonsdottir, Victor Ruotti, Ron Stewart, Igor I Slukvin, and James A Thomson (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318.5858, 1917–1920.

Yuan, Musu, Liang Chen, and Minghua Deng (2022). scMRA: a robust deep learning method to annotate scRNA-seq data with multiple reference datasets. *Bioinformatics* 38.3, 738–745.

Zhang, Hongning, Mingkun Lu, Gaole Lin, Lingyan Zheng, Wei Zhang, Zhijian Xu, and Feng Zhu (2023a). SoCube: an innovative end-to-end doublet detection algorithm for analyzing scRNA-seq data. *Brief. Bioinform.* 24.3, bbad104.

Zhang, Ran, Mu Yang, Jacob Schreiber, Diana R O'Day, James M A Turner, Jay Shendure, Christine M Disteche, Xinxian Deng, and William Stafford Noble (2024). Cross-species imputation and comparison of single-cell transcriptomic profiles. *bioRxivorg*, 2023.10.19.563173.

Zhang, Xinxin, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, Yanyan Ping, Feng Li, Aiai Shi, Jing Bai, Tingting Zhao,

Xia Li, and Yun Xiao (2019). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* 47.D1, D721–D728.

Zhang, Yuan, Jinyun Tan, Kai Yang, Weijian Fan, Bo Yu, and Weihao Shi (2023b). Ambient RNAs removal of cortex-specific snRNA-seq reveals Apoe+ microglia/macrophage after deeper cerebral hypoperfusion in mice. *J. Neuroinflammation* 20.1, 152.

Zhang, Yulong, Siwen Xu, Zebin Wen, Jinyu Gao, Shuang Li, Sherman M Weissman, and Xinghua Pan (2022). Sample-multiplexing approaches for single-cell sequencing. *Cell. Mol. Life Sci.* 79.8, 466.

Zheng, Grace X Y, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049.

Zhong, Huawen, Wenkai Han, David Gomez-Cabrero, Jesper Tegner, Xin Gao, Guoxin Cui, and Manuel Aranda (2025). Benchmarking cross-species single-cell RNA-seq data integration methods: towards a cell type tree of life. *Nucleic Acids Res.* 53.1, gkae1316.

Ziegenhain, Christoph, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* 65.4, 631–643.e4.

List of Figures

1 **Computational workflow for scRNA-seq analysis.** The analysis is broadly divided into three stages: (A) Raw data processing, including read quality control, alignment to a reference genome, and quantification to generate a count matrix; (B) Pre-processing of the count matrix, involving filtering of low-quality cells and genes, doublet detection, ambient RNA correction, normalization, and dimensionality reduction; (C) Downstream analysis, which includes both cell-level approaches such as clustering, cell type annotation, trajectory inference and cell-cell communication analysis, and gene-level analyses like marker gene identification, DE analysis, gene set enrichment analysis, and inference of gene regulatory networks. Created with BioRender.com . . . 8

2 **Visualization of marker genes to support cell type annotation.** UMAPs show the expression of individual markers across cells (left), while heatmaps summarize multiple markers across cell types (right). Adapted from Janssen et al. (2023). 13

3 **Species mixing experiments for scRNA-seq quality assessment.** Cells from two or more species are pooled prior to performing scRNA-seq. During analysis, reads can be attributed back to their species of origin using genome alignment differences or single-nucleotide variants. Summarizing the contributions of different species per cell barcode helps to quantify technical artifacts: balanced contributions suggest doublets, while low-level signals from a second species indicate background RNA contamination. Created with BioRender.com 17

4 **Main approaches for cross-species cell type assignment from scRNA-seq data.** A) Integration across species and annotation on a shared embedding. B) Classification or label transfer from one annotated reference species to the other. C) Independent grouping of cells within each species, followed by identification of correspondences across species. Created with BioRender.com 26

Acknowledgements

It has been a real pleasure to work in this group and that is mainly thanks to all the awesome people who've been part of this journey.

First of all, thank you, **Ines**, for your great guidance throughout the years, for giving me the opportunity to work on so many interesting projects, for many helpful discussions and everything you taught me about research, mentoring, and dogs. Big thanks also to **Wolfgang** for creating such an awesome group and work environment together with Ines, and for your inspiring attitude toward science.

Big thanks to **Zane** for being the most supportive colleague and friend I could have wished for. Since my very first day in the group, your constant encouragement, humor and friendship made the tough moments easier and the whole journey much more fun.

I would also like to thank **Jessy**, for being an excellent counterpart in the wet lab—providing consistently high-quality data and making sure our projects stay on track thanks to your impressive organizational skills. Thank you, **Beate**, for always having that one extra idea and for being such a reliable source of advice and help for all of my projects. And thank you, **Lu**, for your contagious energy—as a scientist and on the volleyball court, football pitch or headis table.

I also want to thank **Johanna, Daniel, Aleks, Ilse and Johannes**——the senior PhD students at the time—for making it so easy to find my place in the group when I first joined and showing me throughout how to navigate a PhD and also have fun along the way.

Thank you to **Fiona, Eva, Felix, Antonia, Manqi and Dana**—the junior PhD students who joined the lab after me, but quickly became a fantastic crew to share this phase with. Not only as scientists, but also as teammates in pub quizzes, marathon relays, parties, and countless lunches and coffee breaks.

Special thanks also to **Dana, Tamina, Paulina** and all the other students I've had the honor of supervising—I'm pretty sure I learned at least as much from you as you did from me.

And thanks to **Ines B., Sara, Karin, Frau Zhao** and all the other past and present members of the Enard/Hellmann group for making this place not just a workplace, but a genuinely good place to be.

To my parents, who have always stood by me—thank you for your incredible support, both emotionally and in all the practical ways that made this journey possible. And thank you to my family and friends for making life outside of science so joyful and full of laughter. A special thank you to **Tobi**, for turning so many vacations into truly memorable and energizing adventures.

And finally, special thank you to **Anita**—having you by my side, not only in the office but also in life, has meant more than I can put into words. You lifted me up during the lows, celebrated the highs, and made me feel safe and loved through it all. I could not have done this without your support and love.