Generating and characterizing primate iPSCs for evolutionary analyses

Dissertation an der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

Raissa Johanna Andrea Geuder
München 2022

Generating and characterizing primate iPSCs for evolutionary analyses

Dissertation an der Fakultät für Biologie

der Ludwig-Maximilians-Universität München

Raissa Johanna Andrea Geuder

München 2022

Diese Dissertation wurde angefertigt
unter der Leitung von Professor Dr. Wolfgang Enard
an der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

1. Gutachter:	Professor Dr. Wolfgang Enard
2. Ggutachter:	Professor Dr. Micha Drukker
Tag der Abgabe:	30. Mai 2022
Tag der mündlichen Prüfung:	14. September 2022

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt ist.

München, den 30.5.2022 Johanna Geuder

(Unterschrift)

Erklärung

Hiermit erkläre ich, *

☑ dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist.

☑ dass ich mich anderweitig einer Doktorprüfung ohne Erfolg **nicht** unterzogen habe.

☐ dass ich mich mit Erfolg der Doktorprüfung im Hauptfach ....

und in den Nebenfächern ....

bei der Fakultät für .... der ....

(Hochschule/Universität)

unterzogen habe.

☐ dass ich ohne Erfolg versucht habe, eine Dissertation einzureichen oder mich der Doktorprüfung zu unterziehen.

München, den 30.5.2022 Johanna Geuder

(Unterschrift)

*) Nichtzutreffendes streichen

Contents

Abbreviations **ix**

Publications **xi**

Declarations of contribution as a co-author **xv**

Summary **1**

1 Introduction **5**

1.1 The evolutionary perspective 5

1.2 Expression conservation as a means to infer function 6

1.3 Quantifying gene expression levels 8

1.3.1 RNA sequencing as the current gold standard 8

1.3.2 Single cell RNA sequencing to study heterogeneous cell populations 9

1.3.3 Optimizing RNA-seq methods 13

1.4 The cell bottleneck in comparative primate transcriptomics . . . 14

1.5 Primate embryonic stem cells 14

1.6 Primate induced pluripotent stem cells to study developmental processes 15

1.6.1 The need for a non-invasive somatic cell source 17

1.6.2 Non-integrating reprogramming methods are important for comparative studies 18

1.7 Generating comparable primate iPSCs 20

1.8 Applying primate iPSCs in a comparative approach 22

2 Results 25

2.1 A non-invasive method to generate induced pluripotent stem cells from primate urine 27

2.2 Prime-seq, efficient and powerful bulk RNA-sequencing 61

2.3 Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq 123

2.4 A comparative study of neural differentiation in primates 157

3 Discussion 191

3.1 Bulk RNA sequencing is an efficient way to characterize primate iPSCs 191

3.1.1 The classic way to characterize pluripotent cells 192

3.1.2 Peculiarities of non-human primate iPSCs 192

3.1.3 prime-seq is an efficient way to characterize primate iPSCs and their derivatives 193

3.2 Extension towards more primate species 195

3.2.1 The importance of a non-invasive somatic cell source . . 195

3.2.2 Human reprogramming factors for non-human primate reprogramming 196

3.2.3 Finding culture conditions that work for all primates . . 197

3.3 Leveraging the information from comparative approaches 199

4 Conclusion and Outlook **201**

Bibliography **203**

List of Figures **227**

Acknowledgements **229**

Curriculum Vitae **231**

Abbreviations

Abbreviation	Definition
DE	differential expression
DGE	differential gene expression
DNA	deoxyribonucleic acid
ESC	embryonic stem cell
FACS	fluorescence-activated cell sorting
hiPSC	human induced pluripotent stem cell
iPSC	induced pluripotent stem cell
mcSCRB-seq	molecular crowding single cell RNA barcoding and sequencing
MEF	mouse embryonic fibroblasts
mRNA	messenger RNA
NGS	Next generation sequencing
NHP	non-human primate
NPC	neural precursor cell
oriP	origin of viral replication
OSKL	OCT3/4, SOX2, KLF4, LIN28
OSKM	OCT3/4, SOX2, KLF4, c-MYC
PBMCs	peripheral mononuclear blood cells
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction
RNA	ribonucleic acid
RNA-seq	RNA-sequencing
rRNA	ribosomal RNA
SAGE	serial analysis of gene expression
scRNA-seq	single-cell RNA sequencing
SeV	sendai virus
TR	transcriptional regulator
UDSC	urine-derived stem cell

Chronological List of Publications

I. Bagnoli JW and Ziegenhain C and Janjic A, Wange LE , Vieth B , Parekh S, **Geuder J**, Hellmann I, Enard W
"Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq"
Nature Communications 9, 2937 (2018).
doi: https://doi.org/10.1038/s41467-018-05347-6

II. **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: https://doi.org/10.1038/s41598-021-82883-0

III. Janjic A and Wange LE, Bagnoli JW , **Geuder J**, Nguyen P, Richter D and Vieth B, Vick B, Jeremias I, Ziegenhain C, Hellmann I, Enard W
"Prime-seq, efficient and powerful bulk RNA sequencing"
Genome Biology 23, 88 (2022).
doi: https://doi.org/10.1186/s13059-022-02660-8

Other Publications

IV. Kliesmete Z and Wange LE, Vieth B, Esgleas B, Radmer J, Hülsmann M, **Geuder J**, Richter D, Ohnuki M, Götz M, Hellmann I, Enard W
"Regulatory and Coding Sequences of TRNP1 Co-Evolve With Cortical Folding in Mammals"
bioRxiv
doi: https://doi.org/10.1101/2021.02.05.429919

Declarations of contribution as a co-author

A non-invasive method to generate induced pluripotent stem cells from primate urine
Mari Ohnuki, Wolfgang Enard and I had the idea for this work. I established iPSC lines and conducted differentiation experiments. I performed EB differentiation and immunostaining experiments with help from Jessica Radmer. Lucas E. Wange, Aleksandar Janjic, Johannes W. Bagnoli and Philipp Janssen and I generated and analyzed RNA-seq data. I helped with karyotype analyses of iPSC lines. The manuscript was written by Wolfgang Enard and me.

Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq
This study was conceived by Christoph Ziegenhain and Wolfgang Enard. Johannes W. Bagnoli, Christoph Ziegenhain, Aleksandar Janjic and Lucas E. Wange prepared sequencing libraries. I and Johannes W. Bagnoli cultured and prepared mouse ES and iPSCs for experiments and performed FAC-sorting. The manuscript was written by Johannes W. Bagnoli, Christoph Ziegenhain, Aleksandar Janjic, Ines Hellmann and Wolfgang Enard.

Prime-seq, efficient and powerful bulk RNA sequencing
This study was conceived by Aleksandar Janjic, Lucas E. Wange, Christoph Ziegenhain and Wolfgang Enard. I prepared iPSCs and performed the differentiation to NPCs. I helped with HEK293T culture and tissue sample acquisition. The manuscript was written by Aleksandar Janjic, Lucas E. Wange, Johannes W. Bagnoli and Wolfgang Enard.

A comparative study of neural differentiation in primates
Wolfgang Enard, Ines Hellmann, Mari Ohnuki and I conceived the study. Mari Ohnuki and I cultured, differentiated and performed FAC-sorting of the cells. Aleksandar Janjic, Lucas E. Wange, Johannes W. Bagnoli and I prepared sequencing libraries. Ines Hellmann and Zane Kliesmete provided guidance in data analysis on all steps. I performed DE analyses, species comparisons and 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 Raissa Johanna Andrea Geuder to these publications.

Wolfgang Enard

Summary

The similarities and differences between us and our closest relatives, the primates, have fascinated researchers for decades and evoked various approaches to better understand the underlying genotype-phenotype relationship. Starting with early comparisons of protein sequences between humans and chimpanzees, substantial technological advances in genomics have led to a deeper understanding of the complexities in this relationship, ranging from cataloging genetic differences to modeling genetic differences in cellular and animal systems. Furthermore, the lack of genetic differences - sequence conservation - is crucial to annotate the human genome and interpret biomedically relevant variants within humans. Charting differences and similarities in molecular and cellular properties can take such a comparative approach to the next phenotypic level. In particular, similar to the information obtained from DNA conservation, expression conservation could help annotating and interpreting human gene expression patterns and thus also provide biomedically relevant information.

However, the major limiting factor in this venture is the availability of comparable samples of different primates, mainly due to ethical constraints. Induced pluripotent stem cells (iPSCs) are used in humans to overcome such limitations, as they can be propagated indefinitely and differentiated to many different cell types. Thus, they can provide a valuable and unique resource for functional primate genomics.

In this context, I established a method to generate iPSCs from primates. One of the major challenges in generating iPSCs from non-model organisms is the acquisition of the somatic cells for reprogramming. Therefore, I focused on urine as a non-invasive cell source and could show that cells can be isolated from very small amounts of primate urine samples, which were collected in an unsterile manner. These cells can be efficiently reprogrammed into iPSCs using the footprint-free Sendai Virus reprogramming method. Utilizing this approach, we generated four iPSC lines from two orangutans, three iPSC lines from one gorilla and nine lines from five humans. We validated the pluripotecy of these lines using immunocytochemistry, differentiation assays and also classified the cells as pluripotent using bulk RNA-sequencing. We further showed that expression differences among clones are comparable to those among individuals and considerably larger than technical sources of variation, suggesting that these cells are a suitable resource for functional primate genomics.

As RNA-sequncing (RNA-seq) is a decisive assay to classify cells and to study gene expression in a comparative context, a robust and affordable method to quantify RNA expression levels is indispensable. I contributed to develop prime-seq, a sensitive bulk RNA-seq protocol that we showed to perform equivalently to standard bulk RNA-seq methods, but at a fourfold higher efficiency due to almost 50-fold cheaper library costs. This is highly useful to e.g. classify generated iPSCs as described above. However, to compare heterogenous cell populations, as they arise for example during the differentiation of iPSCs, RNA-seq with single-cell resolution (scRNA-seq) is crucial. I contributed to develop mcSCRB-seq, a sensitive, powerful and efficient single cell RNA-seq method, that is plate-based and hence, can be used for scRNA-seq on sorted single cells.

Finally, I utilized mcSCRB-seq to compare gene expression trajectories during differentiation of our primate iPSCs towards neural precursor cells (NPCs). We sampled single cells of nine different clones from three species at six different time points during early neural differentiation and thus generated a comprehensive dataset to study this process in a comparable manner. We identify genes with a conserved constant up-regulation throughout the trajectory and find that these genes have a higher probability of being mutation intolerant and a higher probability to be associated with neurodevelopmental disorders. This strengthens the hypothesis that identifying conserved expression patterns in primate iPSCs could carry unique functional information to annotate and interpret the human genome.

In summary, within my thesis I describe the basis for comparative research settings, by providing a non-invasive and footprint-free method to generate iPSCs from various primates. Additionally, I contributed to efficient methods to characterize these cells and showcase in an encompassing study how expression conservation can help to better understand the human genome.

1 | Introduction

1.1 The evolutionary perspective

Our genomes carry the entire heritable information of our complex phenotypes. While we share the vast majority of our genetic information with our closest living relatives (Chimpanzee Sequencing and Analysis Consortium 2005), the phenotypic differences seem - at least to us humans - striking (Figure 1).

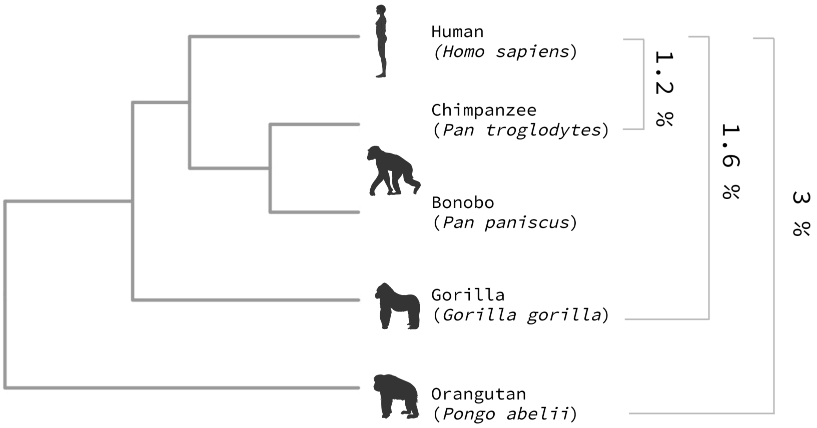


Figure 1. The phylogeny of great apes
The phylogeny of great apes (Bininda-Emonds et al. 2007) and their nucleotide divergence (Chen and Li 2001). Created with BioRender.com

Already by 1975, King and Wilson compared several protein sequences between humans and chimpanzees and reported that the observed phenotypic differences are encoded by less than 1% sequence differences (King and Wilson 1975). When the human chimp consortium in 2005 presented a draft genome sequence of the chimpanzee, they found that single-nucleotide substitutions occur at a rate of 1.23% but also point out that other events like insertions or deletions and chromosomal rearrangements have to be taken into account for a full picture (Chimpanzee Sequencing and Analysis Consortium 2005). Since then, technological advances in molecular biology and computational methods have enabled researchers to study these differences and similarities in more depth (Enard 2016). While examples of potential genetic causes for human-specific phenotypes like vocal learning (Enard et al. 2002), brain size (Florio et al. 2015; Heide et al. 2020) and synaptic plasticity (Charrier et al. 2012) are fascinating, investigating phylogenetic conservation is practically more relevant. For example, it allows to identify functional genetic elements or predict the deleteriousness of variants found within humans (Siepel et al. 2005; Kircher et al. 2014; Goode et al. 2010; Consortium and Zoonomia Consortium 2020; Alföldi and Lindblad-Toh 2013; Hubisz et al. 2011; Malhis et al. 2019). However, to what extend conservation of expression patterns can be used to infer function has not been studied, yet.

1.2 Expression conservation as a means to infer function

Although every cell in our body contains the same genetic material, cells can differ substantially in their morphology and function. For these specific differences to be possible, a complex interplay of chromatin state, DNA methylation, transcription factors, enhancers and regulators is needed in order to orchestrate gene expression, and by that define cellular identity. Even though all of this information is encoded in the genome, we are so far unable to compute all of the details of the phenotypes from the genotype. Measuring phenotypes like gene expression and thus considering the cellular context and dynamic cellular processes, might help explain the differences and similarities of specific cell types and processes in closely related species in more depth. Hence, it is relevant to further deepen our understanding of these genotype-phenotype relationships, by studying the regulation and function of the relevant genes during biological processes of interest.

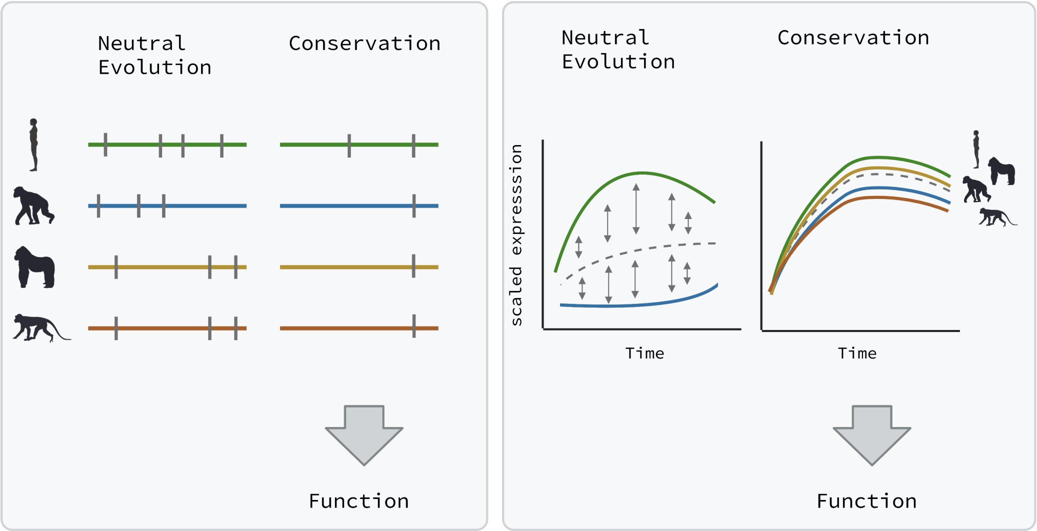


Figure 2. Comparative data and functional inference
Schematic sequence evolution at a neutral and a conserved, i.e. putative functional, locus. Grey bars indicate DNA positions that differ among species (left panel). Schematic expression evolution of a neutral and a conserved, i.e. putative functional gene expression profile across a developmental trajectory (right panel).

In 1994 Duboule proposed the hourglass model of embryogenesis, describing the phenomenon that the mid-embryonic developmental stage within a phylum is conserved, while the earliest and latest stages are more divergent (Duboule 1994). Since then many groups have tried to refine and explain this hypothesis and the underlying mechanisms. While early approaches to assess divergence and conservation were usually based on morphological observations, sequence based studies have opened up new possibilities in this field (Yanai et al. 2011; Kalinka et al. 2010; Irie and Kuratani 2011; Levin et al. 2012; Hu et al. 2017; Liu et al. 2021). In line with this notion, identifying conserved expression patterns during developmental processes could be highly informative and helpful for a better understanding of human biology and disease (Enard 2012). In recent years, the field of comparative genomics is growing and more and more cross-species studies investigate expression conservation between species. Blake et al. for example investigated endoderm differentiation in humans and chimpanzees, and showed that almost all known endoderm developmental markers have similar trajectories between chimp and human (Blake et al. 2018). Ward et al. found a conserved response to hypoxia in cardiomyocytes from humans and chimpanzees after exposing the cells to varying oxygen levels (Ward and Gilad 2019). Applying similar strategies to more developmental processes can help to create a better functional annotation of the human genome. Assessing genome wide expression patterns across species can, in a similar manner to DNA conservation, help identifying functional elements in specific biological systems and tissues (Figure 2). While many aspects of such an approach need further investigation to clarify its validity and its similarity to sequence comparisons, it definitely requires quantifying expression levels in comparable cells across species.

1.3 Quantifying gene expression levels

In order to investigate dynamic changes during developmental processes we need a reliable and quantitative measurement of mRNA levels for a given cell at a given time point. While initially it was only possible to measure an *a priori* defined, specific set of genes, via northern blot (Alwine et al. 1977), later qPCR, microarrays (Schena et al. 1995) or SAGE (Velculescu et al. 1995), this limitation was overcome by the advent of RNA sequencing (RNA-seq) (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Marioni et al. 2008). RNA-seq employs high throughput sequencing of cDNA libraries, generating global gene expression datasets. In contrast to previous methods, no prior knowledge of the sequences is required, making it possible to study non-model organisms and *de novo* transcripts (Shendure 2008; Vera et al. 2008; Zhao et al. 2014).

1.3.1 RNA sequencing as the current gold standard

By 2015 RNA-seq was the dominant transcriptomic method on the market (Lowe et al. 2017), due to the high throughput and sensitivity but also comparably low costs. Many different protocols were generated optimizing this procedure further, but the general steps for most

RNA-seq library preparation methods and analysis workflows are comparable (Figure 3). The cells to be investigated are lysed and RNA is isolated, classically using a column based or bead-based method (Tavares et al. 2011; Oberacker et al. 2019). Enrichment of messenger RNA (mRNA) can be performed either via ribosomal RNA (rRNA) depletion or enrichment of polyadenylated transcripts (Zhao et al. 2018; Yi et al. 2011). Reverse transcription is performed and after a fragmentation step which, depending on the protocol, can also be performed before the cDNA synthesis step (Mortazavi et al. 2008; Picelli et al. 2014; Head et al. 2014; Adey et al. 2010), sequencing adapters are added via PCR. Sequencing can be performed on different platforms. However, Illumina's sequencing-by-synthesis approach dominates the market (Buermans and Dunnen 2014; Greenleaf and Sidow 2014). Subsequent pre-processing steps are necessary in order to ensure good quality data and to be able to to draw robust conclusions. At first reads are demultiplexed using the Illumina indices and/or cell barcodes and a quality filtering step can be implemented to ensure that only reads with high base call quality are used. A central challenge, especially when using non-model organisms, is the mapping of the reads to the reference genome. Subsequently, the number of reads assigned to each gene or transcript is counted and a count matrix is generated. The count matrix contains one column per analyzed sample and one row per detected gene, the values of the matrix are the counts of how many times a feature/gene was detected in a particular sample. Lowly expressed genes and low quality samples are excluded from the analysis and the data is normalized to account for differences in sequencing depth. As this step highly impacts all further analysis it is crucial to choose a good fitting normalization method for the data at hand (Vieth et al. 2019).

1.3.2 Single cell RNA sequencing to study heterogeneous cell populations

For some applications an investigation of the transcriptome with single-cell resolution is indispensable, especially when heterogeneous cell populations, i.e. differentiation processes

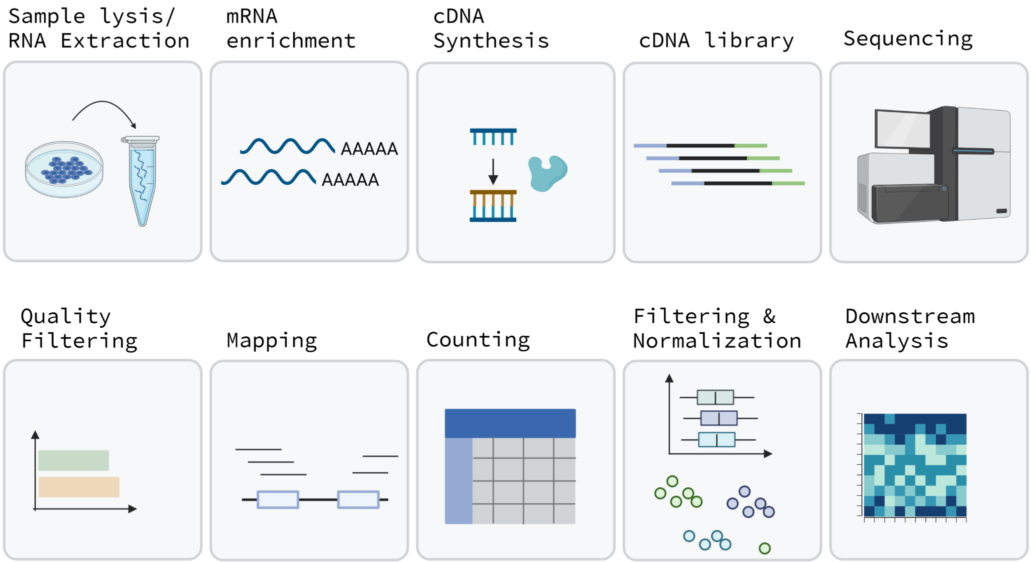


Figure 3. General RNA-seq workflow
After cell lysis and RNA extraction, mRNA is enriched by either rRNA depletion or oligo(dT) primers. The RNA is then reverse transcribed into cDNA, sequencing adapters are added and the library is sequenced on a high throughput machine, e.g. Illumina HiSeq. Reads are filtered for low quality and mapped to an annotated reference genome to generate a count matrix in which reads for each gene are counted for each sample. The data can then be normalized and lowly expressed genes as well as bad quality samples can be removed. Finally, downstream analyses can be performed, e.g. differential gene expression. Created with BioRender.com

or rare cell types are to be investigated (Volpato et al. 2018; Volpato and Webber 2020; Wen and Tang 2016; Ziegenhain et al. 2018). While the first single-cell methods (Tang et al. 2009) were expensive and time consuming, many different protocols have been established and optimized with regards to efficiency and throughput, so that it is a widely used technique to date (Svensson et al. 2018).

Plate-based versus droplet-based scRNA-seq

In essence, single cell RNA-seq (scRNA-seq) methods follow the same workflow as bulk RNA-seq methods. However, they pose different challenges. Naturally, a major challenge lies in the isolation and capture of the very small amount of RNA present in a single cell. Hence, scRNA-seq methods need to be especially sensitive (Bagnoli et al. 2018; Picelli et al. 2014). Furthermore at the beginning of every scRNA-seq experiment the cells to be investigated need to be dissociated into a single cell suspension. While this might be more straightforward for monolayer cellcultures, this can be rather challenging for 3D structures like tissues, organoids or whole organs and can already introduce biases (Brink et al. 2017). Subsequently the cells need to be captured. The most popular protocols either make use of flourescent-activated cell sorting (FACS) for capturing the cells (Soumillon et al. 2014; Picelli et al. 2013; Picelli et al. 2014; Bagnoli et al. 2018), or encapsulate single cells in microdroplets (Zheng et al. 2017; Klein et al. 2015; Macosko et al. 2015) (Figure 4).

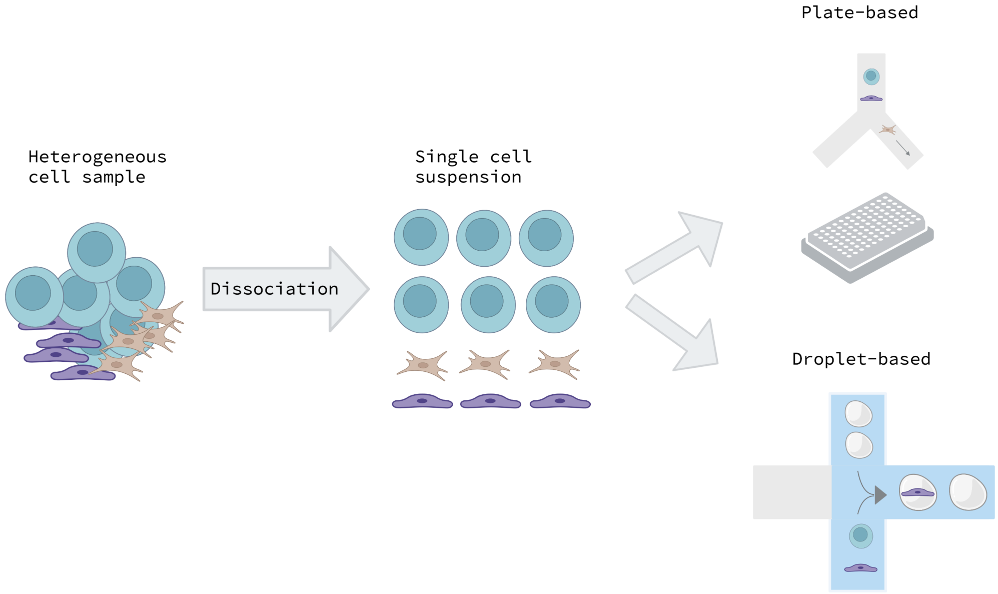


Figure 4. Isolation and capturing of single cells
For scRNA-seq, cells first need to be dissociated and then be captured using plate-based or droplet-based methods. A single cell suspension comprised of living and healthy cells is crucial. Created with BioRender.com

While droplet-based platforms like 10X Genomics are able to capture thousands of cells and therefore identify more rare sub populations, they detect fewer genes in comparison to very sensitive methods like Smart-seq2 (Picelli et al. 2013; Wang et al. 2021). These high-throughput methods, like 10X Genomics, being the current standard (Zheng et al. 2017; Svensson et al. 2018) make large-scale projects like the human cell atlas possible (Rozenblatt-Rosen et al. 2017; Regev et al. 2017). In contrast, plate-based methods are limited in the number of cells (Ziegenhain et al. 2017; Wang et al. 2021) but they offer more flexibility in study design and sample type. Samples can be sorted according to their surface markers and therefore the fluorescence signal of each cell can be associated with the position in the plate (Hayashi et al. 2010). Moreover, plates with lysed cells can be stored, which makes it especially appealing for time-series experiments. Hence, which method is best suited, should be decided depending on the specific scientific question (Ziegenhain et al. 2017; Wang et al. 2021).

Comparative analysis of RNA-seq data during dynamic processes

As more scRNA-seq methods were developed it became clear just how crucial a well-fitting computational analysis pipeline is for the success of the experiment. However, although rough standard workflows can be defined (Lun et al. 2016), all of the subsequent steps have an important impact on the outcome and choices should be made based on the research question and experimental setup (Vieth et al. 2019). Classic goals of scRNA-seq experiments could be differential expression (DE) analysis (Love et al. 2014; Ritchie et al. 2015), identification of cell types or states (Aran et al. 2019) or trajectory inference (Saelens et al. 2019). Comparative analyses between different species come with specific requirements and challenges, so that some steps need further, special and careful considerations. As mentioned before, mapping of the reads to a reference genome, especially when working with non-model organisms, can be challenging, however it is a crucial and central part (Parekh et al. 2018). The quality of the reference genomes is not the same across species and less studied organisms often have missing or truncated gene model annotations, which can lead to huge biases in cross species comparisons. Therefore, strategies are developed to deal with these difficulties. Extending annotations to recapture reads that would otherwise not be counted or cross-mapping for closely related species are essential strategies to deal with these complications (Parekh et al. 2018; Derr et al. 2016).

Furthermore, comparing absolute changes between species can be challenging and the results can be hard to interpret. In contrast, comparing species differences and similarities across a dynamic process, i.e. looking at relative expression changes across time, can lead to more robust results than comparing absolute values of steady cell states. Especially when the cells can be aligned along a common trajectory to account for species specific differences in differentiation speed and homologous cell states among species can be defined (Cannoodt et al. 2016). As for most of the computational tools it is crucial to choose the method based on the scientific question, conditions of the experiment as well as the topology of the data (Saelens et al. 2019). In conclusion, it is crucial to be aware of the experimental setup, limitations and challenges of the study, as well as the impact of choices for analysis tools along the scRNA-seq pipeline (Vieth et al. 2019).

1.3.3 Optimizing RNA-seq methods

In order to be able to study heterogeneous cell populations for example during early developmental processes, I contributed to establishing an optimized version of single cell RNA barcoding and sequencing (SCRB-seq) (Soumillon et al. 2014). A systematic optimization of the different steps of the method by comparing the impact of different reaction enhancers, RT and PCR enzymes led to mcSCRB-seq, a highly sensitive, cost efficient and flexible protocol (Bagnoli et al. 2018).

Based on these findings, we additionally set out to develop an optimized bulk RNA-seq method, as bulk RNA-seq is a still widely used and valuable method, especially when homogeneous cell populations are investigated. I helped developing and benchmarking prime-seq, a sensitive, affordable and robust method as shown for over 6000 samples across 17 species to date (Janjic et al. 2022).

1.4 The cell bottleneck in comparative primate transcriptomics

A major obstacle when attempting comparative approaches during dynamic processes, especially when primates are to be investigated, is the acquisition of the material, i.e. the cells. Generally, cells of early developmental stages are difficult to obtain. As with all vertebrates, experimental research with non-human primates in Germany is strictly regulated by law and precisely defined in the national Animal Welfare Act. While it is possible, with informed consent, to isolate fibroblasts via skin biopsy from humans, this is only hardly possible or even impossible for many primate species. Other cell types are even more critical, e.g. neural cells, unless post mortem tissue is used. However, the availability of post mortem tissue is also very limited and therefore most of the time there is no option to control for covariates like age, viability or individual effects. Although the investigation of post mortem tissue has led to many important insights (Romero et al. 2012), for further comparative approaches especially in a dynamic developmental context, it is crucial to start with comparable, matching cell states and molecular properties and for this renewable sample resources are of pivotal importance.

1.5 Primate embryonic stem cells

Embryonic stem cells (ESCs) have the distinctive advantage that they are pluripotent, meaning that are able to differentiate into any cell type of the adult body, and that they can proliferate indefinitely in culture, making it possible to investigate dynamic processes, rare cell types or transient developmental cell states (Evans and Kaufman 1981; Martin 1981). On the other hand, Human ESCs (hESCs) impose ethical concerns, as they involve the destruction of an embryo and therefore the amount of cell lines available is very limited. In Germany the generation of new hESC lines as well as the import and utilization of hESCs is in principle prohibited and only permissible when the imported human embryonic stem cell lines have been derived before 2007 and are governed by other strict ethical conditions approved by the German parliament (§ 1 StZG and § 4 Abs 1 StZG - Stammzellgesetz). Similarly, experiments on great apes are strictly prohibited and also for monkeys strict rules exist, so naturally very few non-human primate (NHP) ES cell lines are available. Available lines include for example rhesus and cynomolgus (Thomson et al. 1995; Mitalipov et al. 2006; Navara et al. 2007; Suemori and Nakatsuji 2006; Watanabe et al. 2019) or marmoset (Thomson et al. 1996; Sasaki et al. 2005; Debowski et al. 2016).

1.6 Primate induced pluripotent stem cells to study developmental processes

The possibilities of ESCs, but also concerns regarding their generation and utilization, made the finding of Takahashi and Yamanka in 2006 even more groundbreaking (Takahashi and Yamanaka 2006). They discovered that terminally differentiated, adult mouse fibroblasts could be reprogrammed into a pluripotent state, by the ectopic expression of the so-called Yamanaka factors Oct3/4, Sox2, Klf4, and c-Myc, using retroviral tranduction. They called this celltype "induced pluripotent stem cells" (iPSCs) (Takahashi and Yamanaka 2006). Only one year later the same group announced the generation of induced pluripotent stem cells (hiPSCs) from human fibroblasts using the same strategy (Takahashi et al. 2007). Simultaneously another research group succeeded to generate hiPSCs using a different set of factors (OCT3/4, SOX2, NANOG, and LIN28) (Yu et al. 2007). From this discovery on, a multitude of research projects were based on and around iPSCs. To date, hundreds of reprogramming protocols exist, utilizing different somatic cells as starting material, varying reprogramming factors, different ways to introduce the factors into the cells and different culture systems.

The basic workflow (Figure 5) starts with the isolation and culture of primary cells. The reprogramming process is initiated by the forced expression of the reprogramming factors. These factors can be introduced using different strategies which vary in efficiency and safety and should be chosen based on the cells to be reprogrammed and the conditions of the

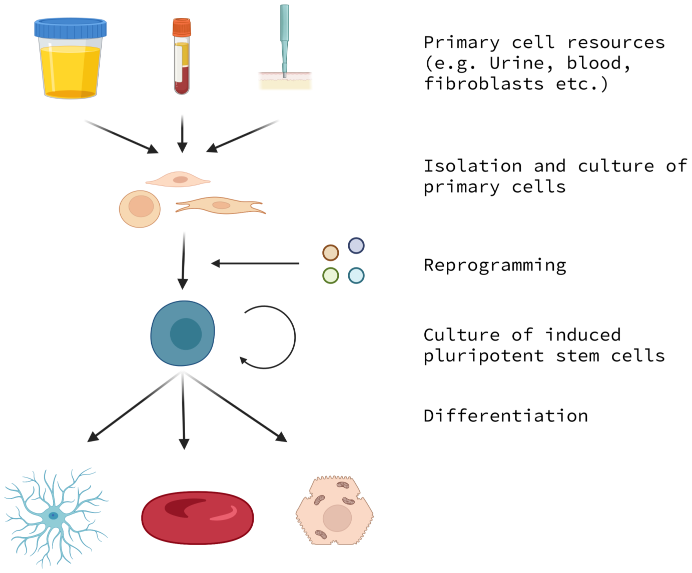

Primary cell resources (e.g. Urine, blood, fibroblasts etc.)
Isolation and culture of primary cells
Reprogramming
Culture of induced pluripotent stem cells
Differentiation

Figure 5. Generation of iPSCs
Primary somatic cells can be isolated from different sources, e.g. urine samples, blood samples or skin biopsies. The cells can then be reprogrammed to iPSCs by the ectopic expression of pluripotency associated transcription factors, classically OCT3/4, SOX2, KLF4, c-MYC - referred to as the Yamanaka factors. The delivery can be performed using various methods. When the cells are successfully reprogrammed they have the ability to self-renew and to differentiate into almost any desired celltype. Created with BioRender.com

facility. Between several days to some weeks post factor delivery to the cells, the first colonies appear, this timing is highly dependent on celltype and reprogramming strategy used. The clones are usually manually picked, transferred to a new plate, expanded and have to pass quality-control measures. Classically, verification assays for genomic stability, pluripotent cell-specific marker gene expression, and the ability to form the three germlayers are the minimum requirements necessary to prove their pluripotency. Similarly to ES cells, the first iPSCs were dependent on the co-culture with mouse embryonic fibroblasts (MEF)-derived feeder cells (Thomson et al. 1998; Takahashi et al. 2007). However, this was overcome in recent years by the invention of various specific media and matrices promoting pluripotency (Xu et al. 2005a; Xu et al. 2005b; Chen et al. 2011; Nakagawa et al. 2014), broadening the spectrum of application areas.

The invention of iPSCs made it possible to generate pluripotent cells specifically for patients and to differentiate them to the desired cell type to be investigated. Not only did this innovative discovery revolutionize the clinical research world, it also opened the door to obtaining any cell types from non-model organisms, like primates.

1.6.1 The need for a non-invasive somatic cell source

It is thought that most somatic cells of the body can in principle be reprogrammed into iPSCs (Ray et al. 2021). The most commonly used primary cells are fibroblasts, as they are well investigated, easy to handle and have low demands on culture conditions (Raab et al. 2014). Also many other somatic cell sources like PBMCs or keratinocytes were shown to be possible sources for proliferating, reprogrammable cells (Staerk et al. 2010; Aasen et al. 2008; Aasen and Izpisúa Belmonte 2010; Ray et al. 2021). The reported efficiency of reprogramming between these celltypes varies depending on the original source, lab and reprogramming strategy (Vidal et al. 2014; Liebau et al. 2013; Sacco et al. 2019; Schlaeger et al. 2015). From primates, skin biopsies or PBMCs can in principle be obtained during planned surgeries or after the death of an animal. However, this still practically constraints the number of individuals and species that can be obtained and thus, a non-invasive somatic cell source would decisively speed up the generation of a variety of primate iPSCs.

A promising celltype that has been widely used in the past years are keratinocytes. Keratinocytes can be isolated from plucked hair and reprogrammed using many different techniques (Klingenstein et al. 2020), they show high reprogramming efficiencies (Aasen et al. 2008; Linta et al. 2012; Petit et al. 2012) and a fast reprogramming process (Piao et al. 2014). Furthermore, keratinocytes from plucked hair are a minimal-invasive cell source, as the acquisition procedure is easy and does not require any special training or any other precautions (Raab et al. 2014). While this holds true in the case of humans, naturally this is not the case for great apes and other primates.

From all possibilities for primary cell isolation described so far, the only totally non-invasive source that is also applicable to other primates are urine samples (Zhou et al. 2011; Zhou et al. 2012; Geuder et al. 2021). Even low volumes of unsterile urine can be sufficient to isolate proliferating urine-derived stem cells (UDSCs) (Geuder et al. 2021). Moreover, the stem cell/progenitor cell properties of UDSCs (Zhang et al. 2008) make them a valuable cell source, as they can proliferate in culture for many passages and reprogram fast and efficiently (Zhou et al. 2011; Bharadwaj et al. 2013; Liu et al. 2020). Primate sampling related issues like contamination and low sampling volumes are negligible due to the low cost per sample and the minimal hands on time required through the process of isolation (Zhou et al. 2011; Zhou et al. 2012; Liu et al. 2020; Geuder et al. 2021).

1.6.2 Non-integrating reprogramming methods are important for comparative studies

Not only different somatic cell sources have been widely investigated, but also the reprogramming methods were subject to change and optimization. The different methods can roughly be grouped into four categories based on the mode of delivery, viral or non-viral, and the integration into the genome (Figure 6). Central for a successful reprogramming method is the ability to sustain the expression of the reprogramming factors for a sufficient time at high enough levels, which highly depends on the cells to be reprogrammed.

Integrating viral methods were the first to be applied during the early phase of induced pluripotent stem cells. Yamanaka and Takahashi used retroviral transduction to reprogram human fibroblasts to iPSCs in 2007 (Takahashi et al. 2007). γ-Retroviral and lentiviral vectors both have the ability to sustain the expression of reprogramming factors at high enough levels during multiple cell divisions, enabling the first iPSC reprogramming based on the information of a large screen for transcription factors. Although, the transgenes are ultimately methylated and silenced in iPSCs (Yao et al. 2004; Stadtfeld et al. 2008), due to their random genomic integration, they can be mutagenic, can obstruct cellular processes and due to residual expression or re-activation of reprogramming factors can interfere with later differentiation protocols (Nakagawa et al. 2008; Hu 2014).

PiggyBac transposons, as an example for the group of integrating, non-viral methods, integrate into the host genome, however in a second step the transient expression of a transposase can catalyze the excision of the transgenes (Fraser et al. 1996). Woltjen et al. used this method and induced the expression of reprogramming factors via a doxycyline

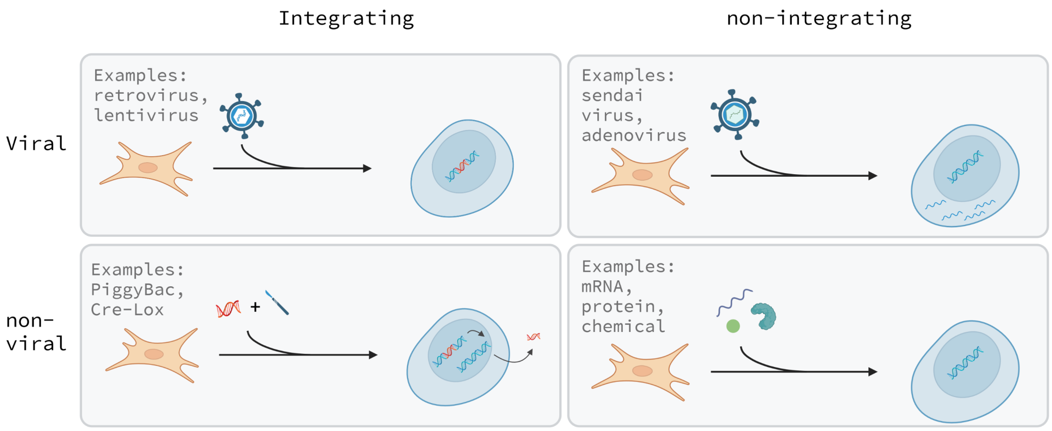

Integrating
non-integrating
Viral
non-viral
Examples: retrovirus, lentivirus
Examples: sendai virus, adenovirus
Examples: PiggyBac, Cre-Lox
+
Examples: mRNA, protein, chemical

Figure 6. Overview of iPSC reprogramming methods
The available reprogramming methods can be grouped into four categories. Integrating viral methods like retroviral (including lentiviral) vectors were the first methods to be used during the dawn of iPSCs. Non-integrating viral methods like the sendai viral vectors are also commonly used, the efficiency is high and they do not enter the nucleus of the host cell, due to their solely RNA-based lifecycle. Non-viral methods like the PiggyBac system involve the insertion of the transgene into the host genome but can subsequently be excised and are therefore potentially also footprint-free. The non-integrating and non-viral methods like mRNA or episomal vector transfection comprise the most diverse group. Many of these methods are widely and successfully used. Created with BioRender.com

inducible PiggyBac system. After successful reprogramming, when the generated clones became dox-independent, transient transposase expression led to complete removal of the transgenes which are flanked by inverted terminal repeats (Woltjen et al. 2009). However, these excisable methods of course also come with the need to verify that the excision did not introduce mutations itself.

Non-integrating viral methods like the Sendai virus (SeV) based reprogramming also show a high reprogramming efficiency but with their completely RNA-based life cycle, they do not enter the nucleus of the cell to be transduced (Bernloehr et al. 2004). A temperature sensitive variant leads to a faster clearing of the viral particles after around 10 passages (Ban et al. 2011; Fusaki et al. 2009). As for all viral methods, the virus absence needs to be proven after successful reprogramming. A major advantage of the Sendai virus system is the small amount of cells needed for the reprogramming procedure and morphological changes are observed comparably early after the transduction (Beers et al. 2015; Geuder et al. 2021).

Many different methods fall into the group of non-integrating and non viral reprogramming strategies. Episomal vectors for example divide extrachromosomally and get diluted out from the cell at a rate of 5 % per cell cycle (Yu et al. 2009). Derived from the Eppstein-Barr virus, the vectors contain the EBNA1 gene and virus origin of viral replication (oriP) they are described to efficiently reprogram cells with only one round of transfection needed, in contrast to previous vector-based methods (Okita et al. 2011; Okita et al. 2013). Another method and, due to its unambiguously footprint free nature, probably the most safe for clinical applications is mRNA reprogramming. mRNAs are synthesized by *in vitro* transcription and modified with 3' and 5' UTR elements and the incorporation of nucleosides to increase their stability (Karikó et al. 2005; Karikó and Weissman 2007; Karikó et al. 2008; Steinle et al. 2017). The first reprogramming protocols were very time consuming with the need for 17 consecutive transfections (Warren et al. 2010), however, rapid progress has been made to increase efficiency and decrease hands on time (Yakubov et al. 2010; Tavernier et al. 2012), so that today it is an efficient and widely used method.

Many different protocols have been described and compared for different types of cells and downstream applications in recent years (Rao and Malik 2012; Malik and Rao 2013; Al Abbar et al. 2020). Non-integrating methods are currently preferred, as their efficiency and practicability have improved and the advantage of not genetically altering cells during reprogramming is decisive, not only but especially for safety aspects in biomedical applications.

1.7 Generating comparable primate iPSCs

The possibility to generate iPSCs from non-model organisms like primates opened a new chapter in the field of comparative primate genomics. Protocols were adjusted and new methods developed to generate a wide range of primate iPSCs, like chimpanzees (Marchetto et al. 2013; Wunderlich et al. 2014; Fujie et al. 2014; Gallego Romero et al. 2015; Blake et al. 2018; Kanton et al. 2019; Pollen et al. 2019; Field et al. 2019), other great apes (Geuder et al. 2021; Ramaswamy et al. 2015) and more distant species like drill (Ben-Nun et al. 2011), baboon (Navara et al. 2018), marmoset (Hemmi et al. 2017; Yoshimatsu et al. 2021) or different macaques (Nakai et al. 2018; Wunderlich et al. 2012; Yada et al. 2017). The majority of the available protocols are optimized for human iPSCs and then applied to non-human primates. In some cases, however, the human optimized conditions seem not to be sufficient to keep the cells in a pluripotent state, for example feeder cells turned out to be essential for the maintenance of japanese macaque iPSCs (Nakai et al. 2018). Nevertheless, many of the primate cells can be reprogrammed feeder-free, cultured in human stem cell medium and show similar characteristics like hiPSCs during and after reprogramming (Geuder et al. 2021; Wunderlich et al. 2014; Gallego Romero et al. 2015).

Albeit the difficulties that come with the generation of primate iPSCs, the optimized protocols of somatic cell isolation as well as iPSC culture in combination with advances in functional genomic technologies make them a promising tool to study evolution within and between species (see also Dannemann and Gallego Romero (2021) for a recent review).

To this end, I established a method that uses urine as the only completely non-invasive cell source in combination with the footprint free Sendai virus (SeV) mediated reprogramming method. The method is based on previously described protocols for urine isolation (Zhou et al. 2011; Zhou et al. 2012) with optimizations specifically to issues that arise during the sampling of primate urine. Briefly, we show that volumes as little as five milliliters are sufficient to isolate reprogrammable cells. Additionally, storing the sample for at least four hours and the addition of a broad-spectrum antibacterial agent make the isolation process from unsterile collected NHP urine possible. Cells from human, gorilla and orangutan reprogrammed quickly and efficiently. Furthermore, expression distances within a species were similar, independent of the individual and donor cell type, highlighting the usefulness of these cells to further expand the zoo of species available for comparative evolutionary analyses (Geuder et al. 2021).

1.8 Applying primate iPSCs in a comparative approach

Using primate iPSCs in combination with efficient bulk and single-cell RNA-seq methodology, is a powerful means for evolutionary studies. To demonstrate the validity of this idea, I used our generated iPSCs in a comparative differentiation approach during which we profiled the transciptomes of single-cells using mcSCRB-seq (Bagnoli et al. 2018). We differentiated the cells towards neural precursor cells (NPCs) and sampled single cells of nine different clones from three species at six different time points, resulting in a comprehensive dataset of more than 4000 cells. We compared differentiation trajectories, identified a set of genes with conserved expression up-regulation during cell-state transition and further characterized these genes as to their functional relevance.

2 | Results

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

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

"A non-invasive method to generate induced pluripotent stem cells from primate urine" (2021)
Scientific Reports 11, 3516 (2021).
doi: https://doi.org/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#Sec26`

Abstract

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.

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

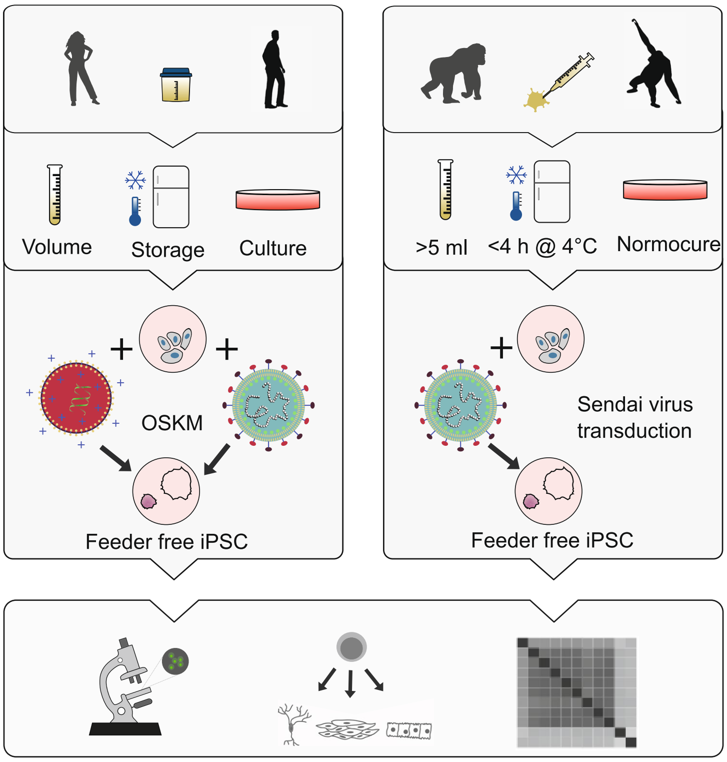


Figure 1. Workflow overview for establishing iPSCs from primate urine. We established the protocol for iPSC generation from human urine based on a previously described protocol[16]. We tested volume, storage and culture conditions for primary cells and compared reprogramming by overexpression of OCT3/4, 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,

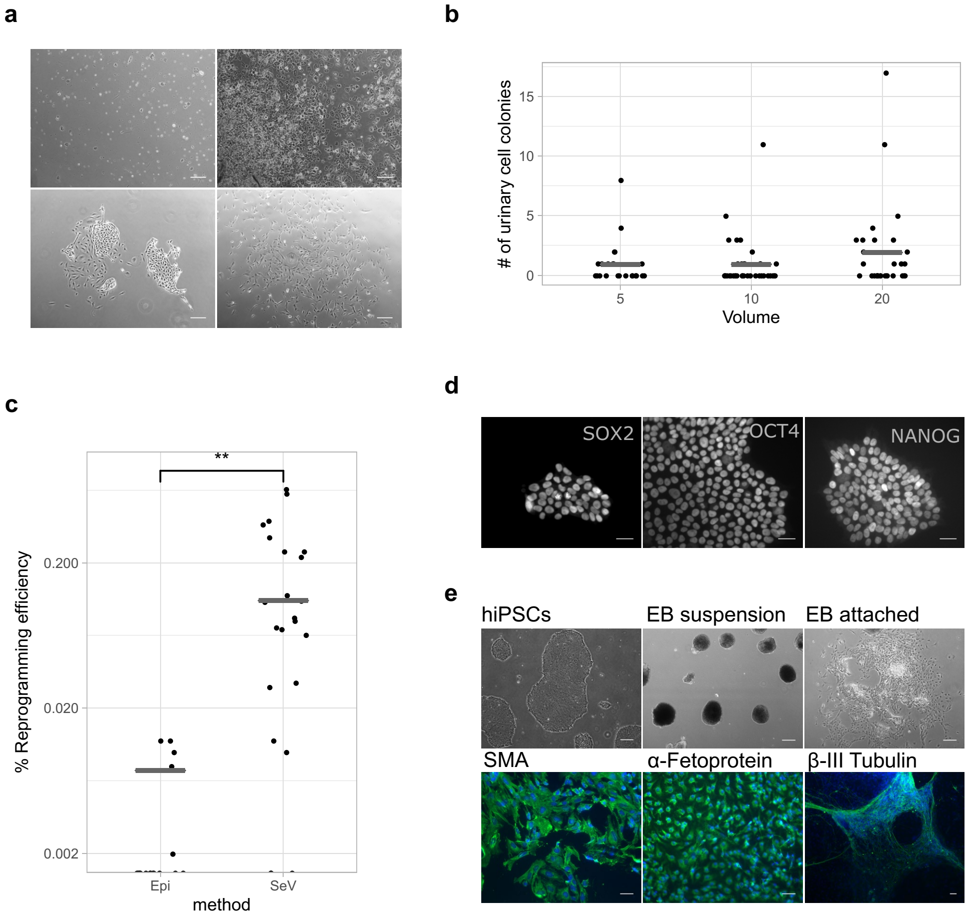


Figure 2. Establishing urinary cell isolation and reprogramming to iPSCs in human samples. (**a**) Human urine mainly consists of squamous cells and other differentiated cells that are not able to attach and proliferate (upper row). After ~ 5 days, the first colonies become visible and two types of colonies 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

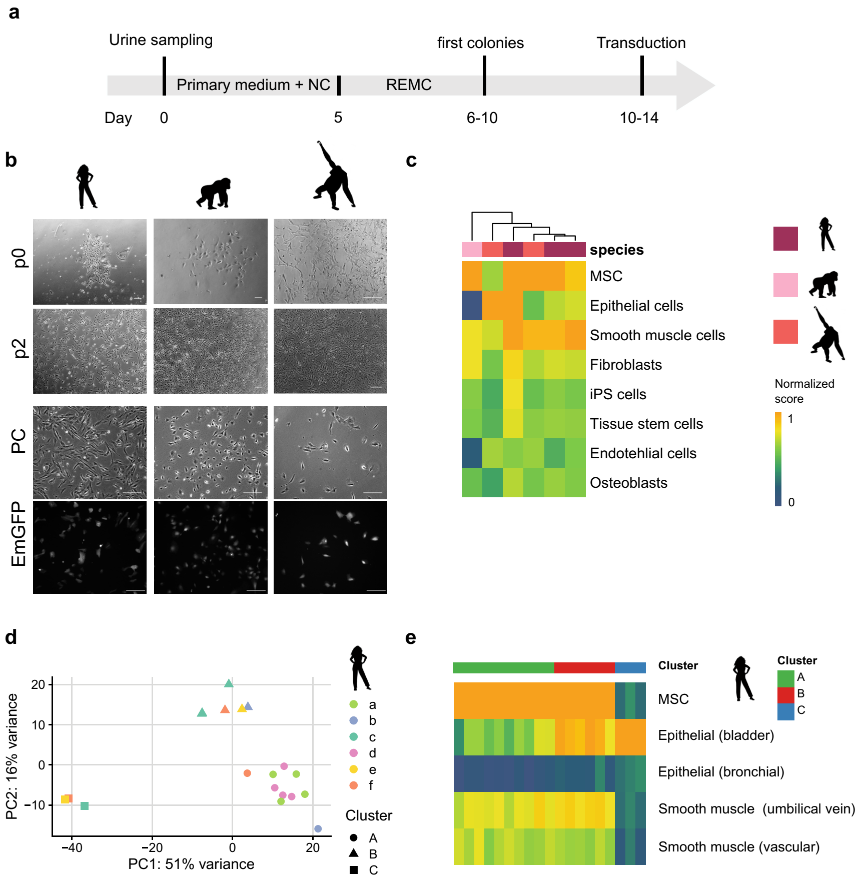


Figure 3. Isolation and characterization of primate urinary cells. (**a**) Workflow of cell isolation from primate urine samples. *NC* Normocure, *REMC* renal epithelial mesenchymal cell medium. (**b**) Primary cells obtained from human, gorilla and orangutan samples are morphologically indistinguishable 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

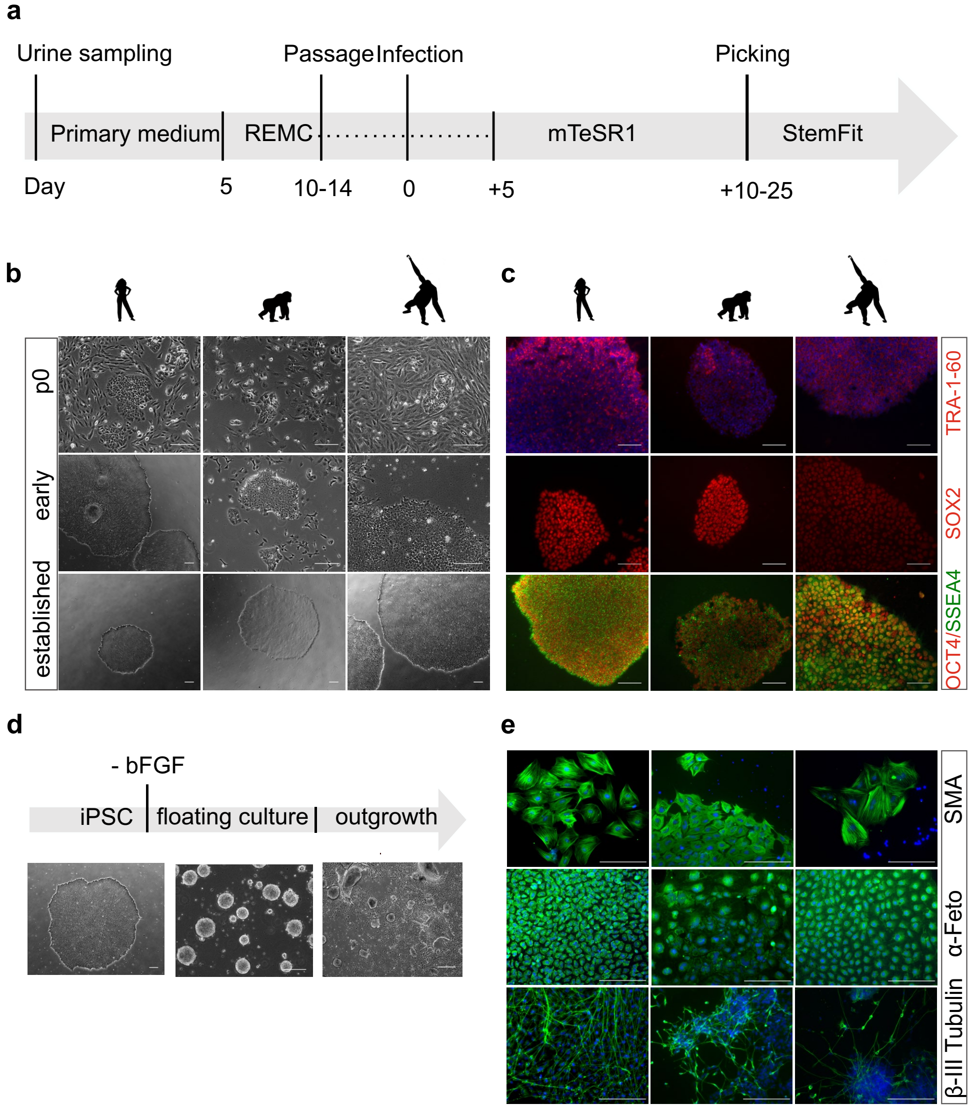


Figure 4. Generation and characterization of primate iPSCs. (**a**) Workflow for reprogramming of primate urinary cells. Urine collection and cell seeding is carried out in primary medium, then after 5 days changed to REMC medium, and only passaged for the first time after 10–14 days. When the cells 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.

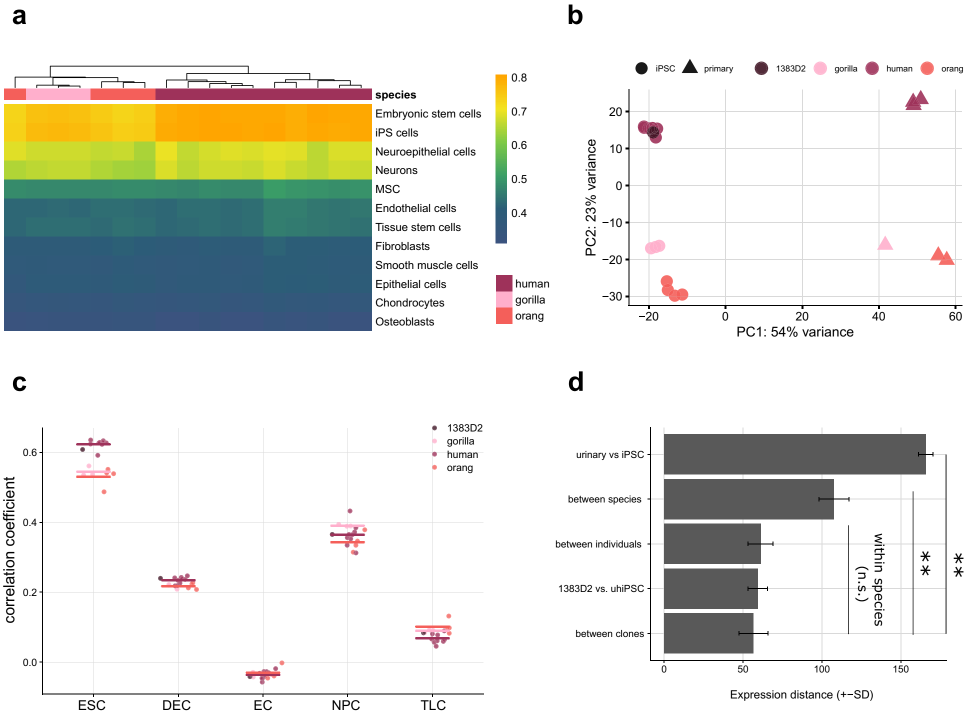


Figure 5. Characterization of primate iPSCs by expression profiling. (**a**) The package SingleR was used to correlate the expression profiles from seventeen samples of primate iPSCs (passage 1–3) to a reference set of 38 human cell types. The twelve cell types with the highest correlations are shown (*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/SinglR/) 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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

Supplementary Figures and Tables

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+]

[1]Anthropology & 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

[+] **Corresponding author, Lead contact:**

Wolfgang Enard and Mari Ohnuki

Anthropology and Human Genomics

Department of Biology II

Ludwig-Maximilians University

Großhaderner Str. 2

82152 Martinsried, Germany

Phone: +49 (0)89 / 2180 - 74 339

Fax: +49 (0)89 / 2180 - 74 331

E-Mail: enard@bio.lmu.de, ohnuki@biologie.uni-muenchen.de

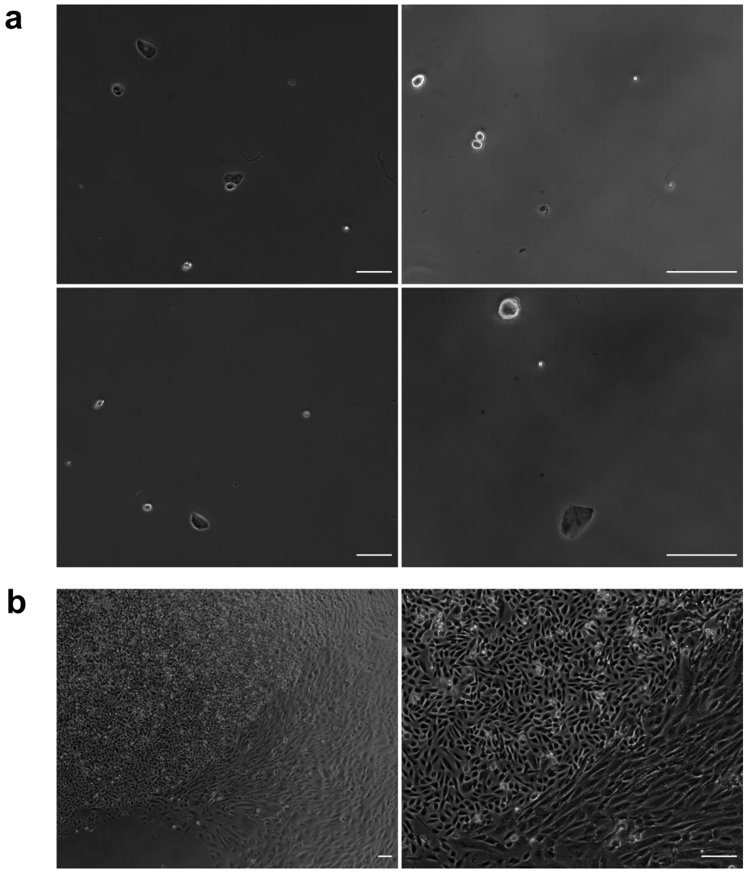


Figure S1. Cell types found in human urine samples
Different types of cells can be found in urine samples directly after collection and after proliferation. (**a**) Different cells found in human samples after centrifugation. Squamous cells as well as various smaller round cells can be found. (**b**) Two different types of cells can be distinguished after one week of culture.

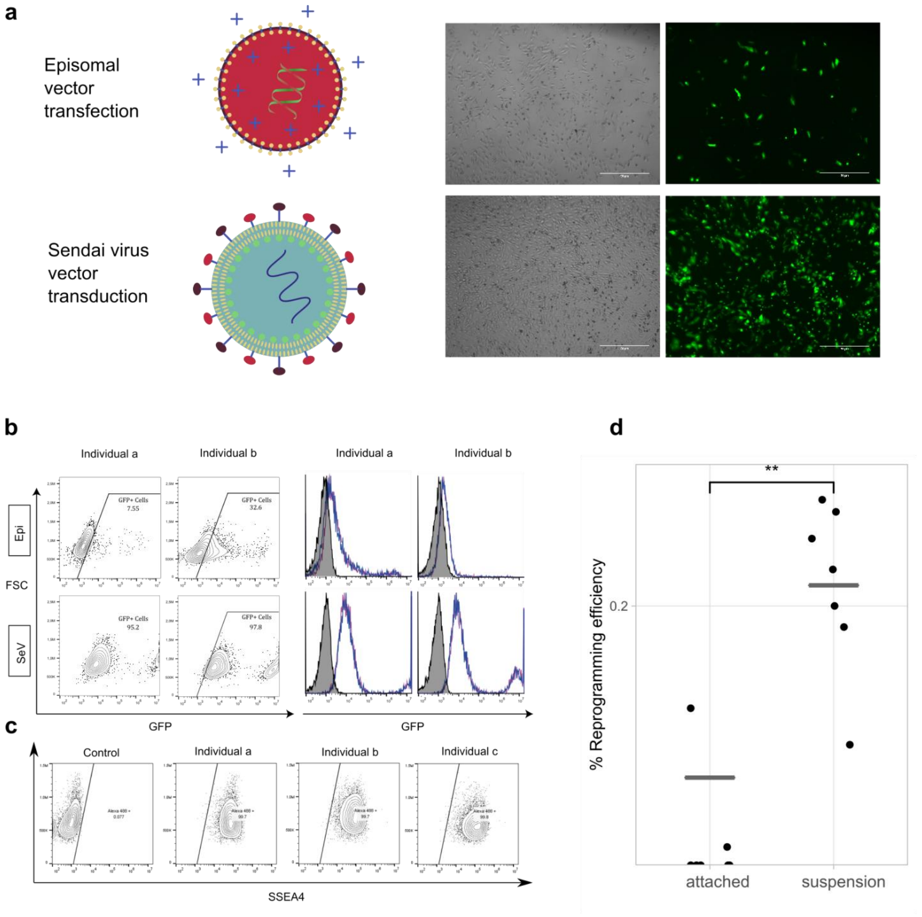


Figure S2. Transfection/Transduction efficiency of urinary cells

(a) GFP expression of urinary cells transfected with pcxle-EGFP episomal plasmids or CytoTune EmGFP transduced after 5 days **(b)** FACS analysis of GFP expressing cells 5 days post transfection/transduction **(c)** SSEA4 expression of urinary cells (**d**) Reprogramming efficiency comparison between attached and suspension reprogramming (suspension reprogramming efficiency: 0.2371%, N=7; attached reprogramming efficiency: 0.09%, N=7; Wilcoxon rank sum test: p=0.00265)

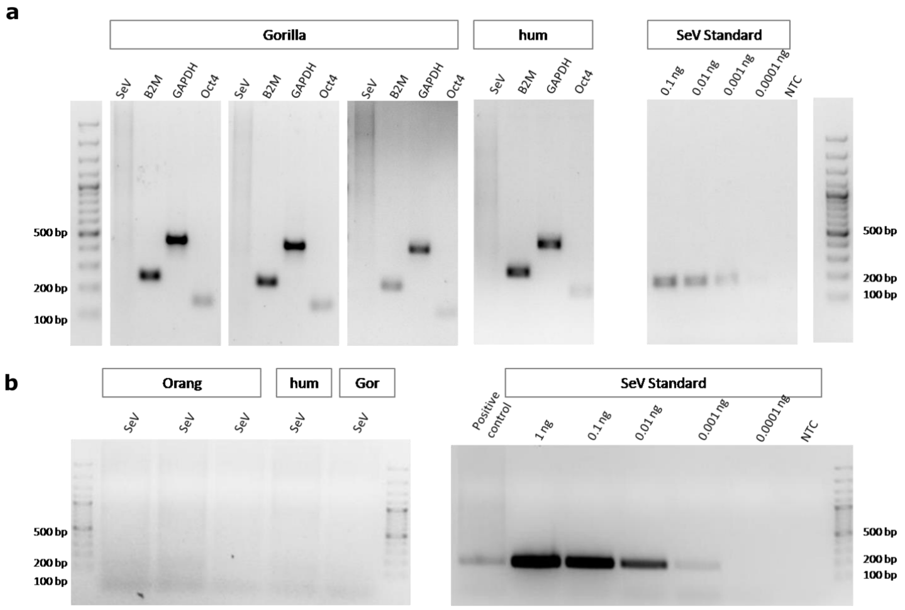


Figure S3. SeV absence verification of primate iPSC lines
Exemplary SeV absence PCR of human and nonhuman primate iPSCs. **(a)** Exemplary gorilla and human PCR targeting the SeV genome and B2M, GAPDH and OCT4 as controls. A standard dilution of the SeV product shows the sensitivity of this assay. **(b)** SEV detection PCR showing human and both primate species have no trace of SeV. The positive control are passage 1 EmGFP transduced fibroblasts.

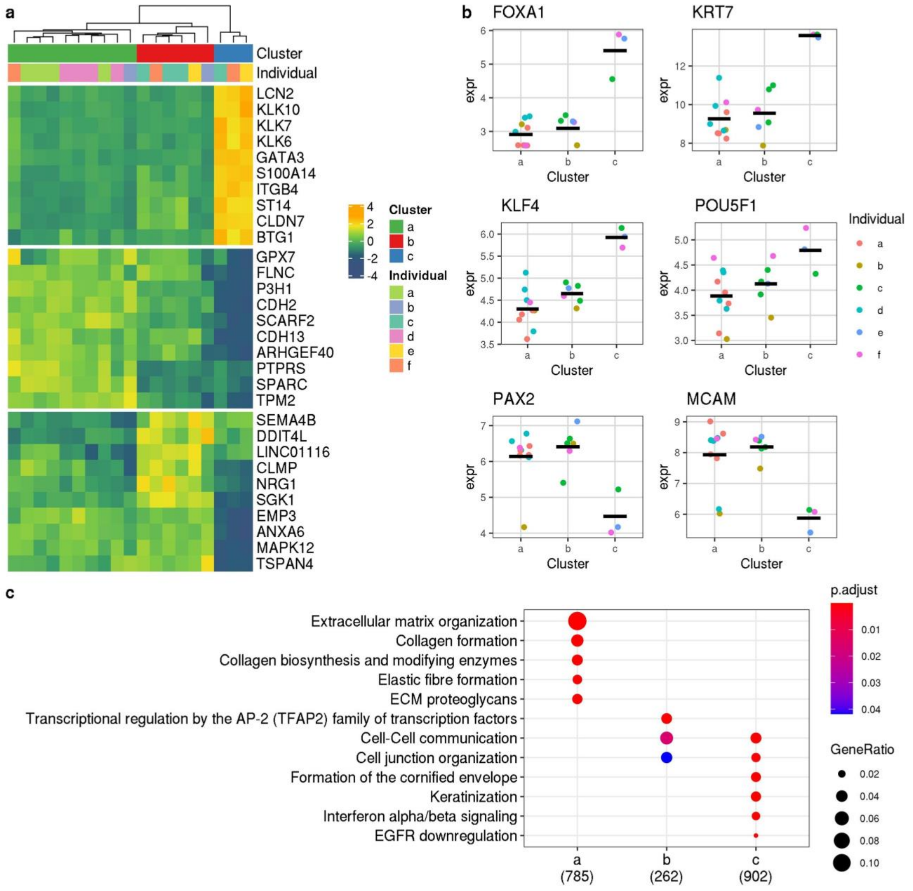


Figure S4. Characterization of human UDSCs originating from single colonies

Expression profiles of single colonies from human urine samples were subjected to further analysis. (**a**) Heatmap of top differentially expressed genes between the clusters. (**b**) Marker gene expression of different cell clusters. Cells in cluster c express urothelial cell markers (FOXA1 and KRT7). Pluripotency markers (KLF4 and POU5F1) are expressed in all clusters. PAX2 and MCAM expression is higher in cluster A and B. (**c**) Top 5 Reactome pathways enriched in the set of genes differentially expressed between one group and both other groups.

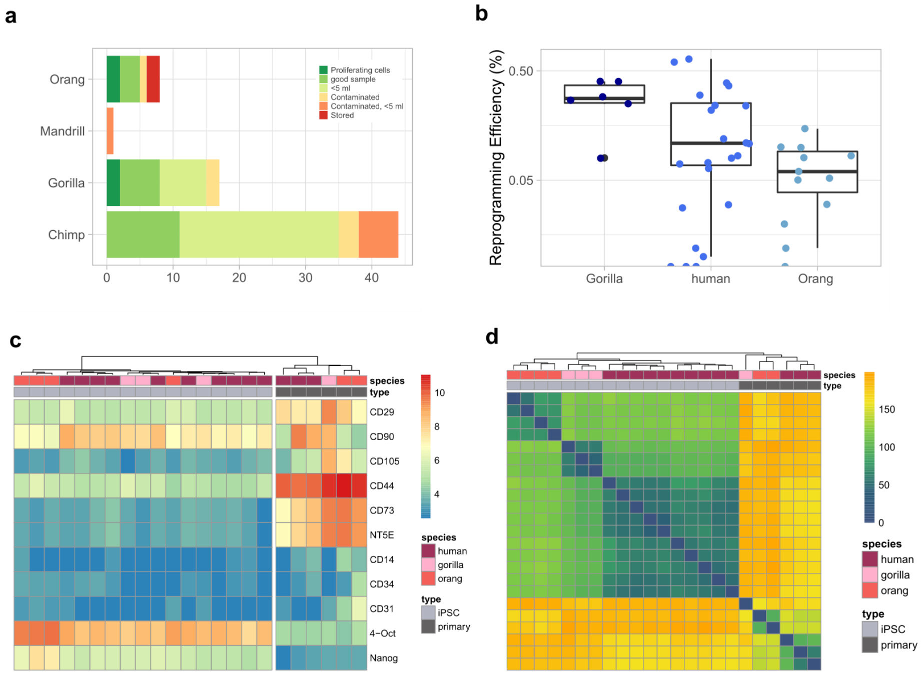


Figure S5. UDSCs and corresponding iPSC characteristics

(a) Overview of collected urine samples and properties of the samples, associated with successful isolation of proliferating cells. **(b)** Reprogramming efficiency shown as colonies per number of seeded cells between species. **(c)** Heatmap of mesenchymal stem cell and iPSC marker expression. **(d)** Euclidean distance between samples.

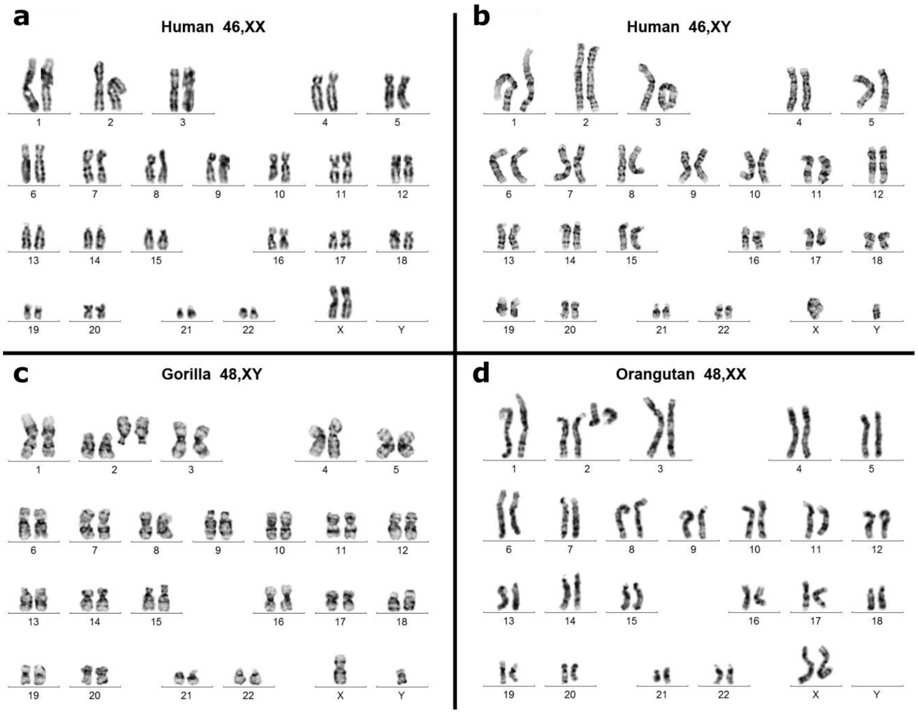


Figure S6. Karyograms of primate iPSC lines

Exemplary karyotyping analysis of human and nonhuman primate iPSCs. **(a)** human female, 46,XX **(b)** human male, 46,XY **(c)** gorilla male, 48,XY and **(d)** orangutan female, 48,XX. All karyotyped iPSC lines showed normal karyotypes without recurrent numerical or structural chromosomal alterations. Note: Ape chromosomes were ordered according to their homologies with human chromosomes and accordingly, human chromosome 2 corresponds to each two gorilla and orangutan chromosomes with homology to the long and the short arm, respectively.

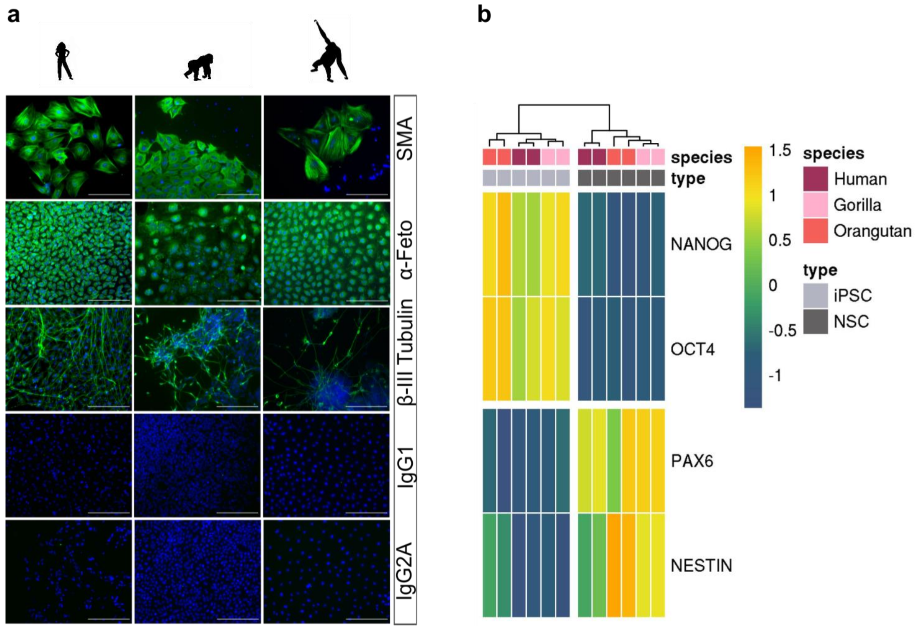
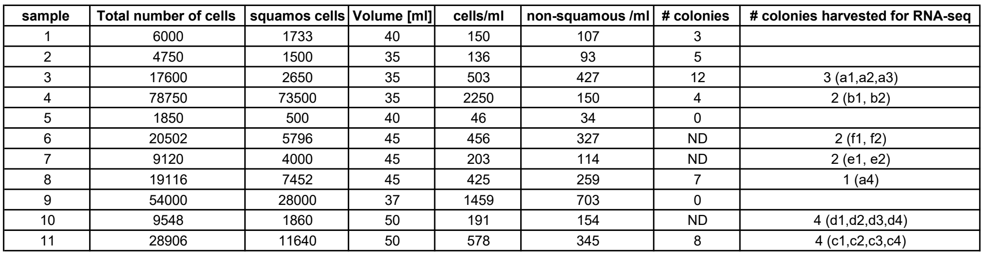


Figure S7. Differentiation capacity of iPSCs
(**a**) Immunofluorescence analyses of ectoderm (β-III Tubulin), mesoderm (α-SMA) and endoderm markers (α-Feto) after EB outgrowth. Nuclei were counterstained with DAPI. Upper 3 panels are taken from Figure 4, lower 2 panels show isotype controls for above antibodies.Nuclei are stained with DAPI in all panels; Scale bars represent 400µm. (**b**) Dual-SMAD inhibition leads to the formation of neurospheres in floating culture, confirmed by neural stem cell marker expression (NESTIN+, PAX6+) using qRT-PCR.

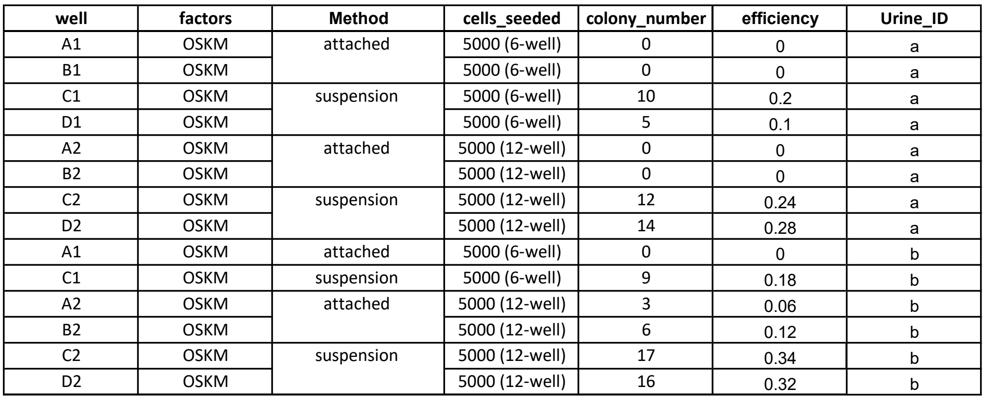


sample	Total number of cells	squamos cells	Volume [ml]	cells/ml	non-squamous /ml	# colonies	# colonies harvested for RNA-seq
1	6000	1733	40	150	107	3	
2	4750	1500	35	136	93	5	
3	17600	2650	35	503	427	12	3 (a1,a2,a3)
4	78750	73500	35	2250	150	4	2 (b1, b2)
5	1850	500	40	46	34	0	
6	20502	5796	45	456	327	ND	2 (f1, f2)
7	9120	4000	45	203	114	ND	2 (e1, e2)
8	19116	7452	45	425	259	7	1 (a4)
9	54000	28000	37	1459	703	0	
10	9548	1860	50	191	154	ND	4 (d1,d2,d3,d4)
11	28906	11640	50	578	345	8	4 (c1,c2,c3,c4)

Experiment	Individual	storage	sex	Normocure	volume [ml]	No_urinary cell colonies	corresponding iPSC line
#1	a	no	M	no	150	8	11C2
	b	no	F	no	180	0	
	c	no	F	no	150	2	12C2
	a	1 hr	M	no	150	12	
#2	a	no	F	no	180	2	64AB1, 64A1
#3	a	no	M	no	90	1	
	a	no	M	no	45	1	
#4	a	2.5	F	no	45	0	
	a	2.5	F	no	45	0	
	a	2.5	F	no	45	0	
	b	1hr on ice	M	no	45	5	
	b	1hr on ice	M	no	45	5	
	b	1hr on ice	M	no	45	4	
	b	1hr on ice	M	no	45	>10	
	c	no	F	no	70	1	
	c	no	F	no	70	1	
	a	no	F	no	45	0	
	a	no	F	no	45	0	
	a	no	F	no	45	1	65B4, 65A3
#5	a	no	M	no	40	4	29B5
	a	no	M	no	30	2	
	a	no	M	no	20	4	
	a	no	M	no	12.5	0	
	b	no	F	no	40	1	
	b	no	F	no	32.5	3	
	b	no	F	no	30	0	
	b	no	F	no	20	0	
	b	no	F	no	12.5	0	
#6	a	no	F	no	40	0	
	a	no	F	no	30	2	
	a	no	F	no	30	1	
	a	no	F	no	20	0	
	a	no	F	no	10	5	
#7	a	13 hrs	M	no	42.5	0	
#8	a	no	M	no	42.6	1	63Ab1.2, 63AB1
	a	no	M	no	10	0	
	a	no	M	no	20	1	
	a	1hr RT	M	no	10	2	
	a	1hr RT	M	no	20	1	
	a	1hr in F	M	no	10	>10	
	a	1hr in F	M	no	20	>10	
	a	1hr on Ice	M	no	10	1	
	a	1hr on Ice	M	no	20	3	
	a	2hrs RT	M	no	10	0	
	a	2hrs RT	M	no	20	2	
	a	2hrs in F	M	no	10	3	
	a	2hrs in F	M	no	20	1	
	a	2hrs on ice	M	no	10	1	
	a	2hrs on ice	M	no	20	3	
	a	3hrs RT	M	no	10	3	
	a	3hrs RT	M	no	20	1	
	a	3hrs in F	M	no	10	1	
	a	3hrs in F	M	no	20	0	
	a	3hrs on ice	M	no	10	0	

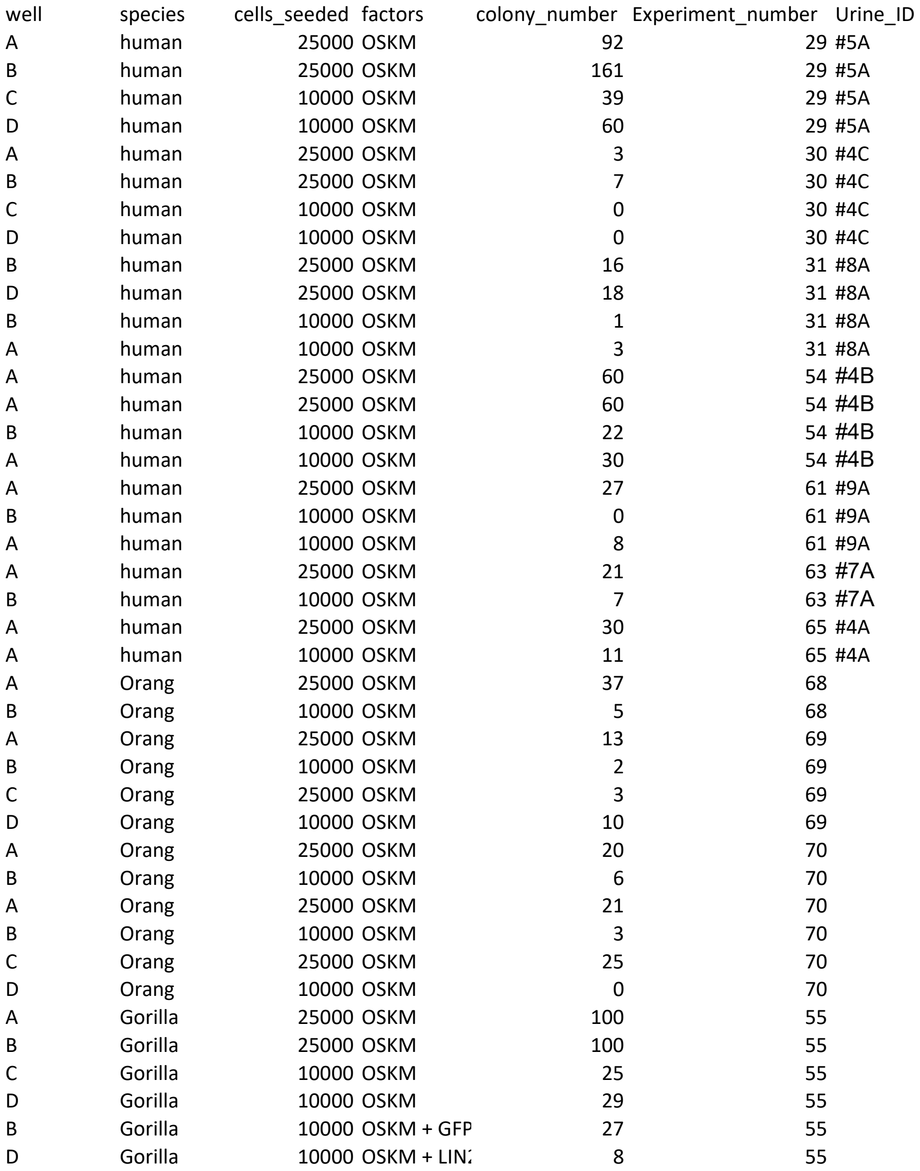
#9	a	3hrs on ice	M	no	20	3	
	b	no	F	no	10	0	
	b	no	F	no	20	0	
	b	1hr RT	F	no	10	0	
	b	1hr RT	F	no	20	0	
	b	1hr in F	F	no	10	0	
	b	1hr in F	F	no	20	0	
	b	1hr on Ice	F	no	10	0	
	b	1hr on Ice	F	no	20	0	
	b	2hrs RT	F	no	10	0	
	b	2hrs RT	F	no	20	0	
	b	2hrs in F	F	no	10	0	
	b	2hrs in F	F	no	20	0	
	b	2hrs on ice	F	no	10	0	
	b	2hrs on ice	F	no	20	0	
	b	3hrs RT	F	no	10	0	
	b	3hrs RT	F	no	20	0	
	b	3hrs in F	F	no	10	0	
	b	3hrs in F	F	no	20	0	
	b	3hrs on ice	F	no	10	0	
#10	b	3hrs on ice	F	no	20	0	
	a	no	F	no	10	0	
	a	no	F	no	20	5	
	a	1hr RT	F	no	10	0	
	a	1hr RT	F	no	20	0	
	a	1hr in F	F	no	10	0	
	a	1hr in F	F	no	20	2	
	a	2hrs RT	F	no	10	0	
#11	a	2hrs RT	F	no	20	0	
	a	0	F	F	5	0	
	a	0	F	T	5	1	
	a	0	F	F	5	0	
	a	0	F	T	5	4	
	a	0	F	F	10	0	
	a	0	F	T	10	0	
	a	4	F	F	5	1	
	a	4	F	T	5	2	
	a	4	F	F	5	1	
	a	4	F	T	5	1	
	a	4	F	F	10	1	
	a	4	F	T	10	1	
#12	a	0	F	F	20	0	
	a	0	M	F	5	0	
	a	0	M	T	5	0	
	a	0	M	F	5	0	
	a	0	M	T	5	0	
	a	0	M	F	10	0	
	a	0	M	T	10	0	
	a	4	M	F	5	1	
	a	4	M	T	5	0	
	a	4	M	F	5	0	
	a	4	M	T	5	0	
	a	4	M	F	10	1	
	a	4	M	T	10	0	
#13	a	0	M	F	20	3	
	a	0	F	F	5	8	
	a	0	F	T	5	0	
	a	0	F	F	5	0	
	a	0	F	T	5	2	
	a	0	F	F	10	3	

	a	0	F	T	10	1	
	a	4	F	F	5	0	
	a	4	F	T	5	0	
	a	4	F	F	5	0	
	a	4	F	T	5	1	
	a	4	F	F	10	0	
	a	4	F	T	10	0	
#14	a	0	F	F	20	17	

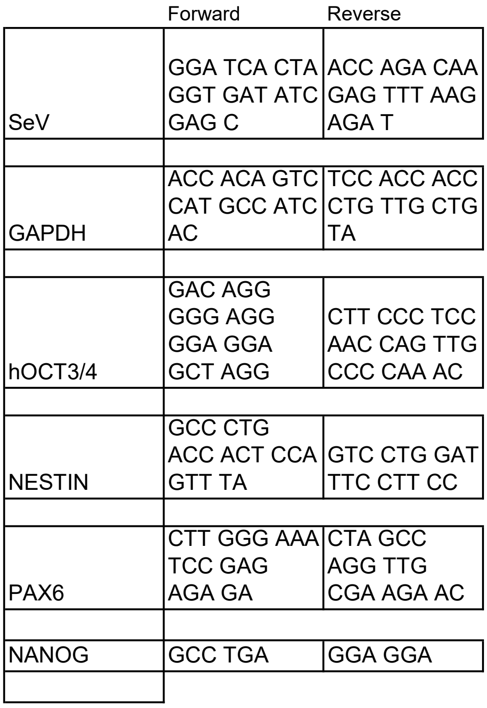


well	factors	Method	cells_seeded	colony_number	efficiency	Urine_ID
A1	OSKM	attached	5000 (6-well)	0	0	a
B1	OSKM		5000 (6-well)	0	0	a
C1	OSKM	suspension	5000 (6-well)	10	0.2	a
D1	OSKM		5000 (6-well)	5	0.1	a
A2	OSKM	attached	5000 (12-well)	0	0	a
B2	OSKM		5000 (12-well)	0	0	a
C2	OSKM	suspension	5000 (12-well)	12	0.24	a
D2	OSKM		5000 (12-well)	14	0.28	a
A1	OSKM	attached	5000 (6-well)	0	0	b
C1	OSKM	suspension	5000 (6-well)	9	0.18	b
A2	OSKM	attached	5000 (12-well)	3	0.06	b
B2	OSKM		5000 (12-well)	6	0.12	b
C2	OSKM	suspension	5000 (12-well)	17	0.34	b
D2	OSKM		5000 (12-well)	16	0.32	b

experiment	Individual	storage	sex	Normocure	species	Volume [ml]	Contaminated	#of proliferating colonies	Corresponding iPSC line
#1	Jahe		F	Y	Orang Utan	15	F		
	Tano		M	Y	Gorilla	10	F	4	55A1, 55C1, 55D1
	Tano		M	Y	Gorilla	5	F	1	
#2	Annemarie		F	Y	Chimpanzee	7	F		
	Sofie		F	Y	Chimpanzee	2	F		
	Walter		M	Y	Chimpanzee	8	F		
	Bagira		F	Y	Chimpanzee	4	F		
#3	Willi		M	Y	Chimpanzee	2	F		
#4	Drill		M	Y	Mandrill	2	T		
#5	Willi		M	Y	Chimpanzee	10	F		
	Willi		M	Y	Chimpanzee	10	F		
#6	Hannerl		F	Y	Chimpanzee	12	F		
	Hannerl		F	Y	Chimpanzee	12	F		
#7	Isahi		F	Y	Orang	6	F		
	Nafi		F	Y	Gorilla	11,5	F		
	Bagira		F	Y	Gorilla	12	T		
	Neema		F	Y	Gorilla	12	F		
	Sonja		F	Y	Gorilla	8	F		
	Bagira		F	Y	Gorilla	6	T		
#8	Bruno		M	Y	Orang Utan	6,5	F		
	Sonja		F	Y	Gorilla	11,5	F		
	Bagira		F	Y	Gorilla	5,5	F		
#9	Bagira		F	Y	Gorilla	2	F		
	Neema		F	Y	Gorilla	3	F		
	zenta		F	Y	Chimp	1,5	F		
	Sophie		F	Y	Chimp	1	F		
	Annemarie		F	Y	Chimp	1	F		
	Walter		M	Y	Chimp	1	F		
#10	Bagira		F	Y in primary med	Gorilla	1,2	F		
	Hanni		F	Y in primary med	chimp	1,3	F		
	Willi		M	Y in primary med	chimp	1,3	F		
	Walter		M	Y in primary med	chimp	0,5	F		
	Sophie		F	Y in primary med	chimp	2,1	F		
	Zenta		F	Y in primary med	chimp	0,9	T		
#11	Hannerl		F	Y in primary med	chimp	2	F		
	Zenta		F	Y in primary med	chimp	2	F		
	Willi		M	Y in primary med	chimp	2	F		
	Hanni		F	Y in primary med	chimp	2	F		
	Walter		M	Y in primary med	chimp	3	F		
	Sophie		F	Y in primary med	chimp	2	F		
#12	Annemarie		F	Y in primary med	chimp	3	F		
	Willi		M	Y in primary med	chimp	3	F		
	Neema		F	Y in primary med	Gorilla	1,2	F		
	Sonja		F	Y in primary med	Gorilla	4	F		
	Bagira		F	Y in primary med	chimp	1,9	F		
	Hanni		F	Y in primary med	chimp	1,2	F		
#13	Annemarie		F	Y in primary med	chimp	3	F		
	Willi		M	Y in primary med	chimp	3	F		
	Sophie		F	Y in primary med	chimp	3	T		
#14	Willi		M	Y in primary med	Gorilla	3	F		
	Zenta		F	Y in primary med	Gorilla	2	F		
	Walter		M	Y in primary med	chimp	3	F		
	Sophie		F	Y in primary med	chimp	4	F		
#15	Neema		F	Y in primary med	Gorilla	5	F		
	Walter		M	Y in primary med	chimp	5	F		
	Sophie		F	Y in primary med	chimp	5	T		
	Sophie		F	Y in primary med	chimp	6	F		
	Willi		M	Y in primary med	chimp	6	F		
	Willi		M	Y in primary med	chimp	6	F		
	Willi		M	Y in primary med	chimp	6	T		
	Willi		M	Y in primary med	chimp	6	T		
	Isalie	>24 hrs	F	Y in primary med	Orang	5	F		
	Bruno	>24 hrs	M	Y in primary med	Orang	6	F		
	Bruno		M	Y in primary med	Orang	5	F	1	68A20, 69A1
#16	Walter		M	Y in primary med	chimp	3	F		
	Sophie		F	Y in primary med	chimp	4	T		
	Hannerl		F	Y in primary med	chimp	3	T		
	Bruno		M	Y in primary med	Orang	6	T		
	Jahe		F	Y in primary med	Orang	6	F	1	70Ab1, 70Af1
	Willi		M	Y in primary med	chimp	5	F		



well	species	cells_seeded	factors	colony_number	Experiment_number	Urine_ID
A	human	25000	OSKM	92	29	#5A
B	human	25000	OSKM	161	29	#5A
C	human	10000	OSKM	39	29	#5A
D	human	10000	OSKM	60	29	#5A
A	human	25000	OSKM	3	30	#4C
B	human	25000	OSKM	7	30	#4C
C	human	10000	OSKM	0	30	#4C
D	human	10000	OSKM	0	30	#4C
B	human	25000	OSKM	16	31	#8A
D	human	25000	OSKM	18	31	#8A
B	human	10000	OSKM	1	31	#8A
A	human	10000	OSKM	3	31	#8A
A	human	25000	OSKM	60	54	#4B
A	human	25000	OSKM	60	54	#4B
B	human	10000	OSKM	22	54	#4B
A	human	10000	OSKM	30	54	#4B
A	human	25000	OSKM	27	61	#9A
B	human	10000	OSKM	0	61	#9A
A	human	10000	OSKM	8	61	#9A
A	human	25000	OSKM	21	63	#7A
B	human	10000	OSKM	7	63	#7A
A	human	25000	OSKM	30	65	#4A
A	human	10000	OSKM	11	65	#4A
A	Orang	25000	OSKM	37	68	
B	Orang	10000	OSKM	5	68	
A	Orang	25000	OSKM	13	69	
B	Orang	10000	OSKM	2	69	
C	Orang	25000	OSKM	3	69	
D	Orang	10000	OSKM	10	69	
A	Orang	25000	OSKM	20	70	
B	Orang	10000	OSKM	6	70	
A	Orang	25000	OSKM	21	70	
B	Orang	10000	OSKM	3	70	
C	Orang	25000	OSKM	25	70	
D	Orang	10000	OSKM	0	70	
A	Gorilla	25000	OSKM	100	55	
B	Gorilla	25000	OSKM	100	55	
C	Gorilla	10000	OSKM	25	55	
D	Gorilla	10000	OSKM	29	55	
B	Gorilla	10000	OSKM + GFP	27	55	
D	Gorilla	10000	OSKM + LIN	8	55	



	Forward	Reverse
SeV	GGA TCA CTA GGT GAT ATC GAG C	ACC AGA CAA GAG TTT AAG AGA T
GAPDH	ACC ACA GTC CAT GCC ATC AC	TCC ACC ACC CTG TTG CTG TA
hOCT3/4	GAC AGG GGG AGG GGA GGA GCT AGG	CTT CCC TCC AAC CAG TTG CCC CAA AC
NESTIN	GCC CTG ACC ACT CCA GTT TA	GTC CTG GAT TTC CTT CC
PAX6	CTT GGG AAA TCC GAG AGA GA	CTA GCC AGG TTG CGA AGA AC
NANOG	GCC TGA	GGA GGA

2.2 Prime-seq, efficient and powerful bulk RNA-sequencing

Janjic, Aleksandar and Wange, Lucas E., Bagnoli, Johannes W., **Geuder, Johanna**, Nguyen, Phong, Richter, Daniel, Vieth, Beate, Vick, Binje, Jeremias, Irmela, Ziegenhain, Christoph, Hellmann, Ines, Enard, Wolfgang

"Prime-seq, efficient and powerful bulk RNA sequencing" (2022)
Genome Biology 23, 88 (2022).
doi: https://doi.org/10.1186/s13059-022-02660-8
Supplementary Information is freely available at the publisher's website:
`https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02660-8#Sec33`

Abstract

Cost-efficient library generation by early barcoding has been central in propelling single-cell RNA sequencing. Here, we optimize and validate prime-seq, an early barcoding bulk RNA-seq method. We show that it performs equivalently to TruSeq, a standard bulk RNA-seq method, but is fourfold more cost-efficient due to almost 50-fold cheaper library costs. We also validate a direct RNA isolation step, show that intronic reads are derived from RNA, and compare cost-efficiencies of available protocols. We conclude that prime-seq is currently one of the best options to set up an early barcoding bulk RNA-seq protocol from which many labs would profit.

Janjic *et al. Genome Biology* (2022) 23:88
https://doi.org/10.1186/s13059-022-02660-8

Genome Biology

METHOD **Open Access**

Prime-seq, efficient and powerful bulk RNA sequencing

Aleksandar Janjic[1,2†], Lucas E. Wange[1†], Johannes W. Bagnoli[1], Johanna Geuder[1], Phong Nguyen[1], Daniel Richter[1], Beate Vieth[1], Binje Vick[3,4], Irmela Jeremias[3,4,5], Christoph Ziegenhain[6], Ines Hellmann[1] and Wolfgang Enard[1*]

*Correspondence:
enard@bio.lmu.de
† Aleksandar Janjic and Lucas E. Wange contributed equally to this work.
[1] Anthropology & Human Genomics, Faculty of Biology, Ludwig-Maximilians University, Großhaderner Str. 2, 82152 Martinsried, Germany
Full list of author information is available at the end of the article

Abstract

Cost-efficient library generation by early barcoding has been central in propelling single-cell RNA sequencing. Here, we optimize and validate prime-seq, an early barcoding bulk RNA-seq method. We show that it performs equivalently to TruSeq, a standard bulk RNA-seq method, but is fourfold more cost-efficient due to almost 50-fold cheaper library costs. We also validate a direct RNA isolation step, show that intronic reads are derived from RNA, and compare cost-efficiencies of available protocols. We conclude that prime-seq is currently one of the best options to set up an early barcoding bulk RNA-seq protocol from which many labs would profit.

Keywords: RNA-seq, Transcriptomics, Genomics, Power analysis

Background

RNA sequencing (RNA-seq) has become a central method in biology and many technological variants exist that are adapted to different biological questions [1]. Its most frequent application is the quantification of gene expression levels to identify differentially expressed genes, infer regulatory networks, or identify cellular states. This is done on populations of cells (bulk RNA-seq) and increasingly with single-cell or single-nucleus resolution (scRNA-seq). Choosing a suitable RNA-seq method for a particular biological question depends on many aspects, but the number of samples that can be analyzed is almost always a crucial factor. Including more biological replicates increases the power to detect differences and including more sample conditions increases the generalizability of the study. As the limiting factor for the number of samples is often the budget, the costs of an RNA-seq method are an essential parameter for the biological insights that can be gained from a study. Of note, costs need to be viewed in the context of statistical power, i.e., in light of the true and false positive rate of a method [2, 3] and these "normalized" costs can be seen as cost efficiency. On top of reagent costs per sample, aspects like robustness, hands-on time, and setup investments of a method can also be seen as

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

cost factors. Other important factors less directly related to cost efficiency are the number and types of genes that can be detected (complexity), the amount of input material that is needed to detect them (sensitivity), and how well the measured signal reflects the actual transcript concentration (accuracy).

In recent years, technological developments have focused on scRNA-seq due to its exciting possibilities and due to the urgent need to improve its cost efficiency and sensitivity [4–6]. A decisive development for cost efficiency was "early barcoding", i.e., the integration of sample-specific DNA tags in the primers used during complementary DNA (cDNA) generation [7, 8]. This allows one to pool cDNA for all further library preparation steps, saving time and reagents. However, the cDNA and the barcode need to be sequenced from the same molecule and hence cDNA-tags and not full-length cDNA sequences are generated. An improvement in measurement noise is achieved by integrating a random DNA tag along with the sample barcode, a Unique Molecular Identifier (UMI), that allows identifying PCR duplicates and is especially relevant for the small starting amounts in scRNA-seq [2, 7, 9]. Optimizing reagents and reaction conditions (e.g., [10, 11]) and the efficient generation of small reaction chambers such as microdroplets [12–14], further improved cost efficiency and sensitivity and resulted in the current standard of scRNA-seq, commercialized by 10X Genomics [5].

Despite these exciting developments, bulk RNA-seq is still widely used and—more importantly—still widely useful as it allows for more flexibility in the experimental design that can be advantageous and complementary to scRNA-seq approaches. For example, investigated cell populations might be homogenous enough to justify averaging, single-cell or single-nuclei suspensions might be difficult or impossible to generate, or single-cell or single-nucleus suspension might be biased towards certain cell types. Most trivial, but maybe most crucial, the number of replicates and conditions is limited due to the high costs of scRNA-seq per sample. Furthermore, as more knowledge on cellular and spatial heterogeneity is acquired by scRNA-seq and spatial approaches, bulk RNA-seq profiles can be better interpreted, e.g., by computational deconvolution of the bulk profile [15]. Hence, bulk RNA-seq will remain a central method in biology, despite or even because of the impressive developments from scRNA-seq and spatial transcriptomics. However, bulk RNA-seq libraries are still largely made by isolating and fragmenting mRNA to generate random primed cDNA sequencing libraries. Commercial variants of such protocols, such as TruSeq and NEBNext, can be considered the current standard for bulk RNA-seq methods. This is partly because improvements of sensitivity and cost efficiency were less urgent for bulk RNA-seq as input amounts were often high, overall expenses were dominated by sequencing costs, and $n = 3$ experimental designs have a long tradition in experimental biology [16]. However, input amounts can be a limiting factor, sequencing costs have decreased and will further decrease, and low sample size is a central problem of reproducibility [17, 18]. To address these needs, several protocols have been developed, including targeted approaches [19–21] and genome-wide approaches that leverage the scRNA-seq developments described above [16, 22]. However, given the importance and costs of bulk RNA-seq, even seemingly small changes, e.g., in the sequencing design of libraries [16], the number of PCR cycles [9], or enzymatic reactions [22], can have relevant impacts on cost efficiency, complexity, accuracy, and sensitivity. Furthermore, protocols need to be available to many labs to be useful and insufficient documentation, limited validation, and/or setup costs can prevent their implementation. Accordingly, further developments of bulk RNA-seq protocols are still useful.

Here, we have optimized and validated a bulk RNA-seq method that combines several methodological developments from scRNA-seq to generate a very sensitive and cost-efficient bulk RNA-seq method we call prime-seq (Fig. 1, Additional file 1: Fig. S1). In particular, we have integrated and benchmarked a direct lysis and RNA purification step, validated that intronic reads are informative as they are not derived from genomic DNA, and show that prime-seq libraries are similar in complexity and statistical power to TruSeq libraries, but at least fourfold more cost-efficient due to almost 50-fold cheaper library costs. Prime-seq is also robust, as we have used variants of it in 22 publications [9, 23–43], 132 experiments, and in 17 different organisms (Additional file 2: Table S1, Additional file 1: Fig. S2). Additionally, it has low setup costs as it does not require specialized equipment and is well validated and documented. Hence, it will be a very useful protocol for many labs or core facilities that quantify gene expression levels on a regular basis and have no cost-efficient protocol available yet.

Results

Development of the prime-seq protocol

The prime-seq protocol is based on the scRNA-seq method SCRB-seq [44] and our optimized derivative mcSCRB-seq [11]. It uses the principles of poly(A) priming,

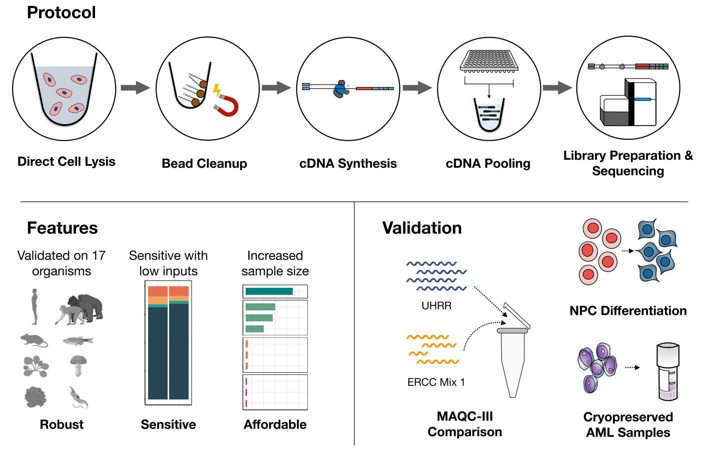


Fig. 1 Graphical overview of prime-seq, highlighting its robustness, sensitivity, affordability, and the validation experiments performed. Cells are first lysed, mRNA is then isolated using magnetic beads, and in turn reverse transcribed into cDNA. Following cDNA synthesis, all samples are pooled, libraries are made, and the samples are sequenced. The protocol has been validated on 17 organisms, including human, mouse, zebrafish, and arabidopsis. Additionally, prime-seq is sensitive and works with low inputs, and the affordability of the method allows one to increase sample size to gain more biological insight. To verify prime-seq's performance, we first compared prime-seq to TruSeq using the publicly available MAQC-III Study data. We then showed robust detection of marker genes in NPC differentiation and high-throughput analysis of AML-PDX patient samples without compromising the archived samples

template switching, early barcoding, and UMIs to generate 3′ tagged RNA-seq libraries (Fig. 1 and Additional file 1: Fig. S1). Compared to previous versions as described, e.g., in [32], we have optimized the workflow, switched from a Nextera library preparation protocol to an adjusted version of NEBNext Ultra II FS, and made the sequencing layout analogous to 10X Chromium v3 gene expression libraries to facilitate pooling of libraries on Illumina flow cells, which is of great practical importance [16]. A detailed step-by-step protocol of prime-seq, including all materials and expected results, is available on protocols.io (https://doi.org/10.17504/protocols.io.s9veh66). We have so far used this and previous versions of the protocol in 22 publications [9, 23–43] and have generated just within the last year over 24 billion reads from > 4800 RNA-seq libraries in 97 projects from vertebrates (mainly mouse and human), plants, and fungi (Additional file 2: Table S1 and Fig. 2A). From these experiences, we find that the protocol works robustly and detects per sample on average >20,000 genes with 6.7 million reads of which 90.0% map to the genome and 71.6% map to exons and introns (Additional file 2: Table S1). Notably, a large fraction (21%) of all UMIs map to introns with considerable variation among samples (Fig. 2A). Across all data sets, about 8000 genes are detected only by exonic reads, ~ 8000 by exonic and intronic reads, and ~ 4000 by intronic reads only (Additional file 1: Fig. S2B, Additional file 2: Table S1). Previous studies for scRNA-seq data showed that intronic reads can improve cluster identification [45] and allow to infer expression dynamics [46]. Also for bulk RNA-seq data, it has been shown that they are informative [47]. Nevertheless, it is an uncommon practice to use them. This might be due to concerns that

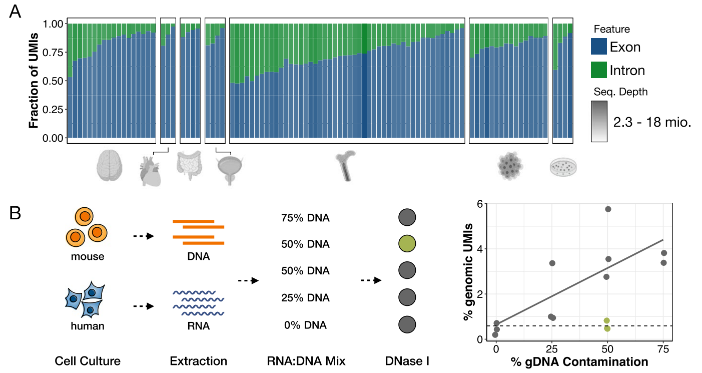


Fig. 2 Intronic reads account for a variable but substantial fraction of UMIs and stem from RNA. **A** Fraction of exonic and intronic UMIs from 97 primate and mouse experiments using various tissues (neural, cardiopulmonary, digestive, urinary, immune, cancer, induced pluripotent stem cells). Sequencing depth is indicated by shading of the individual bars. We observe an average of 21% intronic UMIs, with some level of tissue-specific deviations as, e.g., immune cells generally have higher fractions of intronic reads. **B** To determine if intronic reads stem from genomic DNA or mRNA, we extracted DNA from mouse embryonic stem cells (mESCs) and RNA from human-induced pluripotent stem cells (hiPSCs), pooled the two in various ratios (75, 50, 25, and 0% gDNA), and either treated the samples with DNase I (green) or left them untreated (gray). We then counted the percentage of genomic (=mouse-mapped) UMIs. This indicates that DNase I treatment in prime-seq is complete and that observed intronic reads are derived from RNA

intronic reads could at least partially be derived from genomic DNA as MMLV-type reverse transcriptases could prime DNA that escaped a DNase I digest. Therefore, we investigated the origin of the intronic reads in prime-seq.

Intronic reads are derived from RNA

First, we measured the amount of DNA yield generated from genomic DNA (gDNA). We lysed varying numbers of cultured human embryonic kidney 293T (HEK293T) cells and treated the samples with DNase I, RNase A, or neither prior to cDNA generation using the prime-seq protocol (up to and including the pre-amplification step). Per 1000 HEK cells, this resulted in ~5 ng of "cDNA" generated from gDNA in addition to the 12–32 ng of cDNA generated from RNA (Additional file 1: Fig. S3A). To test the efficiency of DNase I digestion and quantify the actual number of reads generated from gDNA, we mixed mouse DNA and human RNA in different ratios (Fig. 2B). Prime-seq libraries were generated and sequenced from untreated and DNase I-treated samples and reads were mapped to the mouse and human genome (Fig. 2B). In the sample that did not contain any mouse DNA, ~70% of reads mapped to exons or introns (Additional file 1: Fig. S3B) and ~0.5% of the exonic and intronic UMIs mapped to the mouse genome (Additional file 1: Fig. S3C), representing the background level due to mismapping. Importantly, the DNase I-treated sample had almost the same distribution and amount of mismapped UMIs (0.7%), strongly suggesting that the DNase I digest is nearly complete and that essentially all reads in the DNase I-treated sample are derived from RNA (Fig. 2B and Additional file 1: Fig. S3).

As expected, with increasing amounts of mouse DNA, the proportion of mouse-mapped UMIs increased (Fig. 2B), but even with 75% of the sample being mouse DNA, only 3.6% of the UMIs map to the mouse genome, suggesting that also for gDNA-containing samples (e.g., single cells) the impact of genomic reads on expression levels is likely small. Notably, with increasing amounts of gDNA, the fraction of unmapped reads also increased (Additional file 1: Fig. S3B), suggesting that the presence of gDNA does decrease the quality of RNA-seq libraries and does influence which molecules are generated during cDNA generation.

We also analyzed the properties of the intronic reads in DNase-digested prime-seq libraries from HEK cells (Additional file 1: Fig. S4). Intronic reads are enriched towards the 3′ end of genes albeit not as strongly as exonic reads, suggesting that they are derived from internal as well as poly(A)-tail priming events (Additional file 1: Fig. S4). The probability of obtaining an intronic read from a gene depends probably on many factors, such as splicing dynamics (~10% of all transcripts are thought to be pre-mRNAs [46]), expression levels, efficiency of poly(A)-tail priming, and presence of internal priming sites. But as long as these reads are derived from RNA molecules, it seems reasonable to use them for quantifying and comparing gene expression levels as has been laid out previously [47].

In summary, these results indicate that essentially all reads in prime-seq libraries are derived from RNA when samples are DNase I treated and hence that intronic reads can be used to quantify expression levels.

Prime-seq performs as well as TruSeq

Next, we quantitatively compared the performance of prime-seq to a standard bulk RNA-seq method with respect to library complexity, accuracy, and statistical power. A gold standard RNA-seq data set was generated in the third phase of the Microarray Quality Control (MAQC-III) study [48], consisting of deeply sequenced TruSeq RNA-seq libraries generated from five replicates of Universal Human Reference RNA (UHRR) and External RNA Controls Consortium (ERCC) spike-ins. As Illumina's TruSeq protocol can be considered a standard bulk RNA-seq method, and as the reference RNAs (UHRR and ERCCs) are commercially available, this is an ideal data set to benchmark our method. As in the MAQC-III design, we mixed UHRR and ERCCs (Additional file 1: Fig. S5) in the same ratio but at a 1000-fold lower input and generated eight prime-seq libraries, which were sequenced to a depth of at least 30 million reads. We processed and downsampled both data using the zUMIs pipeline [45] and compared the two methods with respect to their library complexity (number and expression levels of detected genes), accuracy (correlation of estimated expression level and actual number of spiked-in ERCCs), and statistical power (true positive and false positive rates in data simulated based on the mean-variance distribution of technical replicates of each method).

We found that prime-seq has a slightly lower fraction of exonic and intronic reads that can be used to quantify gene expression (78% vs. 85%; Fig. 3A, Additional file 1: Fig. S6A). But despite the slightly lower number of reads that can be used, prime-seq does detect at least as many genes as TruSeq (Fig. 3B). Of these, 33,230 genes are detected with both methods (76.2%) (Additional file 1: Fig. S6B). Pairwise sample comparisons between ($R^2 = 0.64$) the two methods are lower than within the methods ($R^2 = 0.94$ and 0.97), as one would expect (Additional file 1: Fig. S6C). Additionally, the comparison of normalized expression data between prime-seq and TruSeq shows stronger correlation in ERCC spike-in molecules ($R^2 = 0.95$) than endogenous molecules ($R^2 = 0.67$) (Additional file 1: Fig. S6D). This is likely explained by the biological variation of the samples, as the ERCC spike-ins are synthetically produced to exact specifications, and UHRR is extracted from a mixture of cell lines, which may have altered in composition or expression in the 7 years separating the two experiments. Both methods also show a similar distribution of gene expression levels (Fig. 3D), indicating that the complexity of generated libraries is generally very similar.

The accuracy of a method, i.e., how well estimated expression levels reflect actual concentrations of mRNAs, is relevant when expression levels are compared among genes. Here, TruSeq and prime-seq show the same correlation (Pearson's $R^2 = 0.94$) between observed expression levels and the known concentration of ERCC spike-ins, indicating that their accuracy is very similar (Fig. 3C).

However, for most RNA-seq experiments, a comparison among samples—e.g., to detect differentially expressed genes—is more relevant. Therefore, it matters how well genes are measured by a particular method, i.e., how much technical variation a method generates across genes. As we have 8 and 5 technical replicates of the same RNA for prime-seq and TruSeq, respectively, we can estimate for each method the mean and variance per gene. Note that UMIs are only available for prime-seq and hence only prime-seq can profit from removing technical variance by removing PCR duplicates (Fig. 3A). The empirical distribution shows the characteristic dependency of RNA-seq

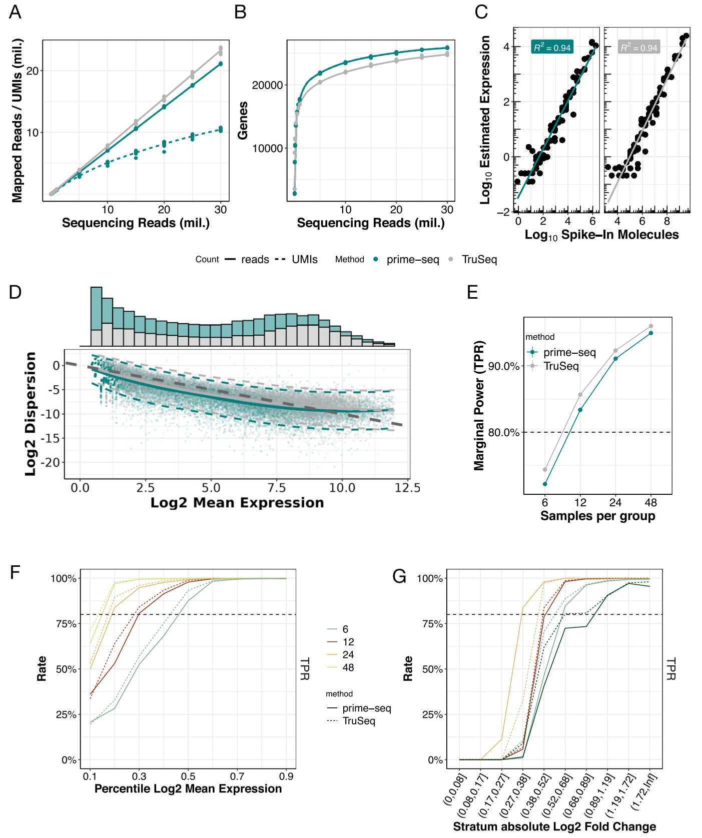


Fig. 3 Prime-seq has similar sensitivity and power compared to TruSeq (MAQC-III data). **A** Mapped reads, UMIs (dashed line, only prime-seq), and **B** detected genes (exonic + intronic reads) at varying sequencing depths between TruSeq data from the MAQC-III Study and matched prime-seq data, show prime-seq and TruSeq are similarly sensitive (filtering parameters: detected UMI $\geq$ 1, detected gene present in at least 25% of samples and is protein coding). **C** Accuracy, measured by spike-in molecules, is similarly high in both methods ($R^2 = 0.94$). **D** The distribution of genes across mean expression is similar for both methods, as well as the dispersion, which follows a Poisson distribution (dark gray dashed line) for lower expressed genes and then increases as technical variation increases for highly expressed genes. The local polynomial regression fit between mean and dispersion estimates per method is shown in solid lines with 95% variability band per gene shown in dashed lines. **E** Power analysis at a sequencing depth of 10 million reads shows almost identical power between prime-seq and TruSeq, and a similar increase at varying sample size for **F** mean expression and **G** absolute log2 fold change. Data filtering parameters: detected UMI $\geq$ 1, detected gene present in at least 25% of samples

data on sampling (Poisson expectation) at low expression levels and an increasing influence of the additional technical variation at higher expression levels (Fig. 3D). Prime-seq shows a slightly lower variance for medium expression levels where most genes are expressed (Fig. 3D). To quantify to what extent these differences in the mean-variance distribution actually matter, we used power simulations as implemented in powsimR [49]. We simulated that 10% of genes sampled from the estimated mean-variance relation of each method are differentially expressed between two groups of samples. The fold changes of these genes were drawn from a distribution similar to those we observed in actual data between two cell types (iPSCs and NPCs) or two types of acute myeloid leukemia (AML) (see below and Additional file 1: Fig. S7A). The comparison between this ground truth and the identified differentially expressed genes in a simulation allows us to estimate the true positive rate (TPR) and the false discovery rate (FDR) for a particular parameter setting. We stratified TPR and FDR across the number of replicates (Fig. 3E), the expression levels (Fig. 3F), and the fold changes (Fig. 3G) to illustrate the strong dependence of power on these parameters. At a given FDR level, a more powerful method reaches a TPR of 80% with fewer replicates, at a lower expression level, and/or for a lower fold change. We find that the power of the two methods is almost identical as FDR and TPR are very similar across conditions for both methods. The false discovery rates (FDR) are—as expected—generally below 5% for 12, 24, or 48 replicates per condition (Additional file 1: Fig. S7B-D) and the (marginal) TPR across all expression levels and fold changes is 80% for both methods at ~12 replicates per condition (Fig. 3E). The power increases for both methods in a similar manner with increasing expression levels (Fig. 3F) and increasing fold changes (Fig. 3G). This is also the case when using only exonic reads for the power analysis (Additional file 1: Fig. S7B and S7E-F). In summary, prime-seq and TruSeq perform very similarly in estimating gene expression levels with respect to library complexity, accuracy, and statistical power.

Bead-based RNA extraction increases cost efficiency and throughput

As library costs and sequencing costs drop, standard RNA isolation becomes a considerable factor for the cost efficiency of RNA-seq methods. RNA isolation using magnetic beads is an attractive alternative [50] and we have used it successfully in combination with our protocol before [11]. To investigate the effects of RNA extraction more systematically, we compared prime-seq libraries generated from RNA extracted via silica columns and via affordable carboxylated magnetic beads (for more information see Additional file 3. Supplemental Text). Libraries from cultured HEK293T cells, human peripheral blood mononuclear cells (PBMC), and mouse brain tissue showed a similar distribution of mapped reads, albeit with a slightly higher fraction of intronic reads in magnetic bead libraries (Fig. 4A and S8) and considerable differences in expression levels (Fig. 4B and S9).

To further explore these differences, we tested the influence of the Proteinase K digestion and its associated heat incubation (50 °C for 15 min and 75 °C for 10 min), which is part of the bead-based RNA isolation protocol. We prepared prime-seq libraries using HEK293T RNA extracted via silica columns ("Column"), magnetic beads with Proteinase K digestion ("Magnetic Beads"), magnetic beads without Proteinase K digestion ("No Incubation"), and magnetic beads with the same incubations but without the addition of the enzyme ("Incubation"). Interestingly, the shift to higher intronic fractions and the expression profile similarity is mainly due to the heat incubation, rather than the enzymatic digestion by Proteinase K (Additional file 1: Fig. S8A and B).

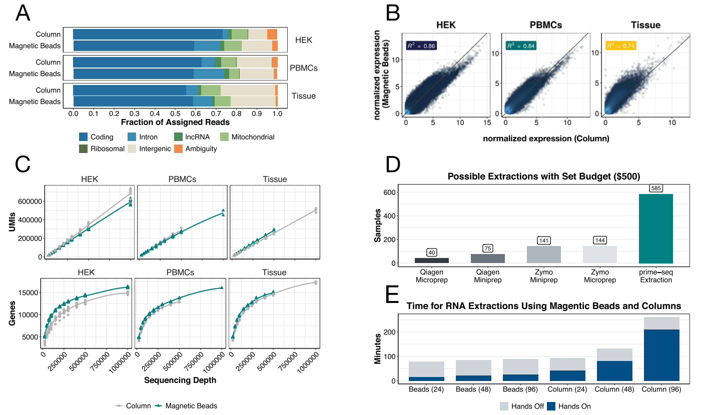


Fig. 4 RNA extraction with beads, rather than columns, provides similar sequencing data while increasing throughput capabilities. **A** Feature distributions of RNA isolated with a column-based kit and magnetic beads show that both RNA extraction protocols produce similar amounts of useable reads from cultured human embryonic kidney 293T (HEK293T) cells, peripheral blood mononuclear cells (PBMC), and harvested mouse brain tissue. **B** Gene expression between both bead and column extraction are also similar in all three tested inputs ($R^2 = 0.86$ HEK, 0.84 PBMCs, and 0.74 tissue). **C** Detected UMIs and detected genes for column and magnetic beads in HEK293T, PBMCs, and tissue are almost identical, with slightly more detected genes in the bead condition (filtering parameters: detected UMI $\geq$ 1, detected gene present in at least 25% of samples and is protein coding). Comparison of costs (**D**) and time (**E**) required for different RNA extractions

Hence, bead-based extraction does create a different expression profile than column-based extraction, especially due to the often necessary Proteinase K incubation step. This confirms the general influence of RNA extraction protocols on gene expression profiles [51]. Importantly, the complexity of the two types of libraries is similar, with a slightly higher number of genes detected in the bead-based isolation (Fig. 4C, Additional file 1: Fig. S8C and S8D), potentially due to a preference for longer transcripts with lower GC contents (Additional file 1: Fig. S9C).

So while bead-based RNA isolation and column-based RNA isolation create different but similarly complex expression profiles, bead-based RNA isolation has the advantage of being much more cost-efficient. At least four times more RNA samples can be processed for the same budget (Fig. 4D, Additional file 4: Table S2). In addition, RNA isolation using magnetic beads is twice as fast and without robotics more amenable to high-throughput experiments (Additional file 5: Table S3). Thus, we show that bead-based RNA isolation can make prime-seq considerably more cost-efficient without compromising library quality.

Prime-seq is sensitive and works well with 1000 cells

As prime-seq was developed from a scRNA-seq method [44], it is very sensitive, i.e., it generates complex libraries from one or very few cells. This makes it useful when input material is limited, e.g., when working with rare cell types isolated by FACS or when working with patient material. To validate a range of input amounts, we generated RNA-seq libraries from 1000 (low input, ~10–20 ng total RNA) and 10,000 (high input, ~100–200 ng) HEK293T cells. The complexity of the two types of libraries was very similar, with only a 2% decrease in the fraction of exonic and intronic reads and a 7.7% and 1.9% reduction in the number of UMIs and detected genes at the same sequencing depth (Additional file 1: Fig. S10A). The expression profiles were almost as similar between the two input conditions as within the input conditions (median r within = 0.94, median *r* between = 0.93; Additional file 1: Fig. S10B), indicating that expression profiles from 1000 and 10,000 cells are almost identical in prime-seq. Using a lower number of input cells is certainly possible and unproblematic as long as the number of cells is unbiased with respect to the variable of interest. Using higher amounts than 10,000 cells is certainly also possible, but it is noteworthy that we have observed a large fraction of intergenic reads in highly concentrated samples, potentially due to incomplete DNase I digestion (data not shown). In summary, we validate that an input amount of at least 1000 cells does not compromise the complexity of prime-seq libraries and hence that prime-seq is a very sensitive RNA-seq protocol.

Barcode swapping in prime-seq is low

One potential concern with early barcoding methods is the swapping of barcodes due to the formation of chimeric molecules during PCR, resulting in a “contamination” of a cell’s expression profile with transcripts from another cell. This has been discussed in the context of scRNA-seq library generation [52, 53], but it is not clear to what extent it is relevant in bulk RNA-seq methods. To quantify barcode swapping, we generated prime-seq libraries from isolated total RNA from mouse embryonic stem cells (mESCs) and human-induced pluripotent stem cells (iPSCs) either separately or pooled after reverse transcription (pooling) as it is normally done in the prime-seq protocol (Additional file 1: Fig. S11A). We find that less than 0.1% of the mapped UMIs in the ten separately amplified human libraries, map to mouse, representing a low background rate due to mismapping and index swapping during sequencing. In contrast, ~0.5% of the mapped UMIs in the five human libraries that were generated together with five mouse libraries map to mouse (Additional file 1: Fig. S11B). So barcode swapping does occur, but at a relatively low level, consistent with previous findings for single human and mouse cells for our related mcSCBR-seq method [11] (Additional file 1: Fig. S11C) and that the amount of swapped barcodes correlates strongly with the amount of transcripts in the pool (Additional file 1: Fig. S11D). Importantly, even 10% of barcode swapping has fairly little influence on power as shown in simulations (Additional file 1: Fig. S11E). In summary, we show that barcode swapping is present, but not a major issue for prime-seq as long as absolute expression levels, like the presence or absence of a gene, are interpreted accordingly. However, the amount of barcode swapping does depend on reaction conditions, specifically on the number of PCR cycles, but probably on more conditions such as types of polymerases [54], input amounts, library complexity, and sequence similarities. Hence, better controlling and understanding barcode swapping within and across methods might be important.

Two exemplary applications of prime-seq

To exemplify the advantages with respect to sensitivity and throughput in an actual setting, we used prime-seq to profile cryopreserved human acute myeloid leukemia (AML) cells from patient-derived xenograft (PDX) models [23, 55]. These consisted of different donors and AML subtypes and were stored in freezing medium at – 80 °C for up to 3.5 years (Fig. 5A). Due to the sensitivity of prime-seq, we could use a minimal fraction of the sample without thawing it by taking a 1-mm biopsy punch from the vial of cryopreserved cells and putting it directly into the lysis buffer. This allowed sampling of precious samples without compromising their amount or quality and resulted in 94 high-quality expression profiles that clustered mainly by AML subtype (Fig. 5B) as expected [56].

To further exemplify the performance of prime-seq, we investigated its ability to detect known differences in a well-established differentiation system [57]. We differentiated five human-induced pluripotent stem cell (iPSCs) lines [36] to neural progenitor cells (NPCs) and generated expression profiles using prime-seq (Fig. 5C). In a hierarchical clustering of well-known marker genes [58], the iPSCs and NPCs formed two distinct groups and the expression patterns were in agreement with their cellular identity. For example, the iPSC markers POU5F1, NANOG, and KLF4 showed an increased expression in the iPSCs and NES, SOX1, and FOXG1 in NPCs (Fig. 5D).

Prime-seq is cost-efficient

We have shown above that the power, accuracy, and library complexity is similar between prime-seq and TruSeq. The performance and robustness of the prime-seq protocol has been demonstrated by the two examples above as well as its many applications using this or previous versions of the protocol [9, 23–35, 42, 43, 59, 60]. In

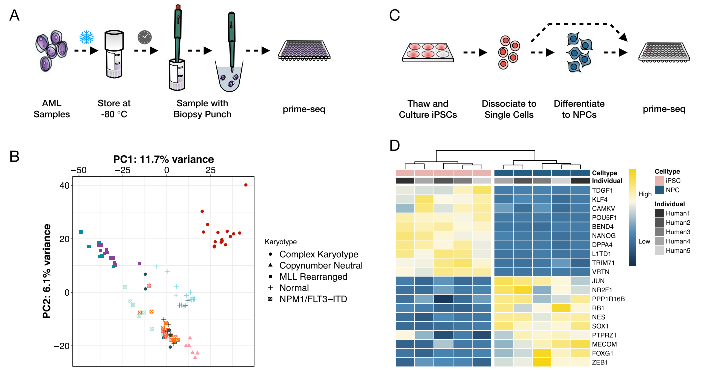


Fig. 5 Two exemplary applications of prime-seq. **A** Experimental design for an acute myeloid leukemia (AML) study, where a biopsy punch was used to collect a small fraction of a frozen patient-derived xenograft (PDX)-AML sample. **B** Prime-seq libraries were generated from 94 PDX samples, derived from 11 different AML-PDX lines (color-coded) from 5 different AML subtypes (symbol-coded) and cluster primarily by AML subtype. **C** Experimental design for studying the differentiation from five human-induced pluripotent stem cell lines (iPSCs) to neural progenitor cells (NPC). **D** Expression levels from 20 a priori known marker genes cluster iPSCs and NPCs as expected

summary, one could argue that prime-seq performs as well as TruSeq for quantifying gene expression levels. Other methods that generate tagged cDNA libraries using early barcoding have also been developed [16, 22, 61–64]. This includes BRB-seq that uses poly(A) priming and DNA-Pol I for second-strand synthesis and also performs similarly to TruSeq [22]. Decode-seq also uses poly(A) priming and template switching like prime-seq, but adds sample-specific barcodes and UMIs at the 5′ end [16]. In a direct comparison, Decode-seq performed slightly better than BRB-seq and due to a more flexible sequencing layout [16]. While slight differences in power, accuracy, and/or library complexity might exist among these protocols, cross-laboratory benchmarking on exactly the same samples as recently done, e.g., for scRNA-seq methods [5] or small RNA-seq methods [65], are probably needed to quantify such differences reliably. For now, it is probably fair to say that RNA-seq methods like BRB-seq, prime-seq, TruSeq, Smart-Seq, or Decode-seq all perform fairly equal with respect to quantifying gene expression levels. Hence, at a fixed budget, the cost per sample will determine to a large extent how many samples can be analyzed and hence how much biological insight can be gained.

To this end, we calculated the required reagent costs to generate a library from isolated RNA in a batch of 96 samples for the different commercial methods as well as for prime-seq, Decode-seq, and BRB-seq (Additional file 6 Table S4). With $2.53 per sample prime-seq is the most cost-efficient method, followed by BRB-seq ($4.05) and Decode-seq ($6.58). Commercial methods range from $60 (NEBNext) to $164 (SMARTer Stranded). This is illustrated by the number of libraries that can be generated by a fixed budget of $500 (Fig. 6A). Note that these costs include for all methods $1.39 per sample for two Bioanalyzer (Agilent) Chips (Additional file 6: Table S4) and do not consider the additional cost reduction that is associated with the direct bead-based RNA extraction of prime-seq (see above). The drastic advantage of prime-seq, Decode-seq, and BRB-seq also becomes apparent when power is plotted as a function of costs with and without sequencing (10 million reads per sample) (Fig. 6B, Additional file 1: Fig. S12A). For example, to reach an 80% TPR at a desired FDR of 5%, one needs to spend $715 including sequencing costs for prime-seq, $795 when using Decode-seq, $1625 when using Illumina Stranded, and $3485 when using TruSeq (Additional file 1: Fig. S12B).

Cost efficiency with respect to time can also matter and we calculated hands-on and hands-off time for the different methods (Additional file 7: Table S5). Hands-on times vary from 30 to 35 min for the non-commercial, early barcoding methods to 52–191 min for commercial methods. However, as all methods require essentially a full day of lab work, we consider the differences in required times not as decisive, at least not in a research lab setting where RNA-seq is not done on a daily or weekly basis. In summary, we find that prime-seq is the most cost-efficient bulk RNA-seq method currently available.

Discussion

In this paper, we present and validate prime-seq, a bulk RNA-seq protocol, and show that it is as powerful and accurate as TruSeq in quantifying gene expression levels, but more sensitive and much more cost-efficient. We validate the DNase I treatment and

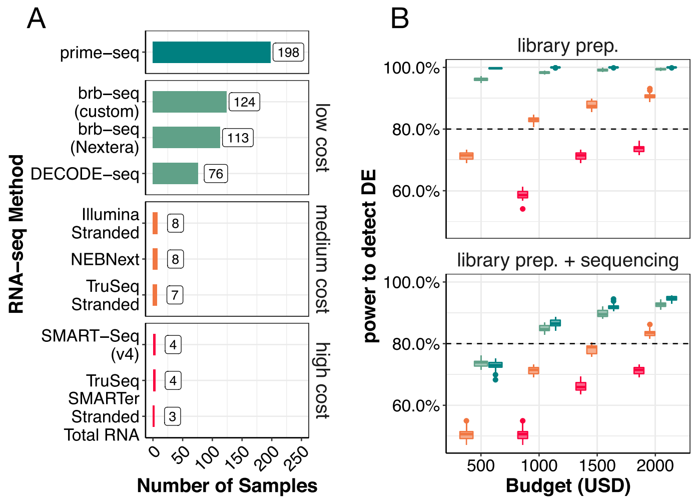


Fig. 6 Prime-seq is very cost-efficient. **A** With a set budget of $500, prime-seq allows one to process 198 samples, which is 1.6 times more samples than the next cost-efficient method. **B** The compared methods were grouped into low, middle, and high cost methods and the TruSeq MAQCII data was used as a basis for power analysis for all methods but prime-seq. The increase in sample size due to cost efficiency directly impacts the power to detect differentially expressed genes, as evident by the increased performance of prime-seq and other low cost methods (BRB-seq and Decode-seq), even when sequencing costs are included in the comparison (sequencing depth of 10 mio. reads at a cost of $3.40 per 1 mio. reads)

determine that intronic reads are derived from RNA and can be used in downstream analysis. We also validate input ranges and the direct lysis and bead-based RNA purification of tissue and cell culture samples. Finally, we exemplify the use of prime-seq by profiling AML samples and NPC differentiation and show that prime-seq is currently the most cost-efficient bulk RNA-seq method. In the following, we focus our discussion on advantages and drawbacks of prime-seq in comparison to other RNA-seq protocols. To this end, we distinguish protocols like TruSeq, Smart-Seq, or NEBNext that individually process RNA samples and generate full-length cDNA profiles ("full-length protocols") from protocols like prime-seq, Decode-seq, or BRB-seq that use early barcoding and generate 5′ or 3′ tagged cDNA libraries ("tag protocols").

Complexity, power, and accuracy are similar among most bulk RNA-seq protocols

Initially, early barcoding 3′ tagged protocols generated slightly less complex libraries (i.e., detected fewer genes for the same number of reads), especially due to a considerable fraction of unmapped reads [22, 66]. These reads are probably caused by PCR artifacts during cDNA generation and amplification. Protocol optimizations as shown for BRB-seq [22], Decode-seq [16], and here for prime-seq have reduced these artifacts and hence have improved library complexity to the level of standard full-length protocols. For prime-seq, we have shown quantitatively that its complexity, accuracy, and power is very similar to that of TruSeq. More comprehensive studies, ideally across laboratories [5, 48], would be needed to quantitatively compare protocols, also with respect to their robustness across laboratories and conditions and their biases for individual transcripts. For the context and methods discussed here, we would argue that there are no decisive differences in power, accuracy, and complexity among tag protocols and full-length protocols at least when performed under validated and optimized conditions.

Cost efficiency makes tag-protocols preferable when quantifying gene expression levels

As shown above (Fig. 6) and as argued before [16, 22, 66], the main advantage of tag protocols is their cost efficiency. Their most obvious drawback is that they cannot quantify expression levels of different isoforms. Smart-Seq2 [67] and Smart-Seq3 [10] are relatively cost-efficient full-length protocols that were developed for scRNA-seq. However, they have not been validated and optimized for bulk RNA-seq and would still be considerably more expensive than most tag protocols. Furthermore, as reconstructing transcripts from short-read data is difficult and requires deep sequencing, isoform detection and quantification is now probably more efficiently done by using long-read technologies [1]. However, from our experience, most RNA-seq projects quantify expression at the gene level not at the transcript level. This is probably because most projects use RNA-seq to identify affected biological processes or pathways by a factor of interest. As different genes are associated with different biological processes, but different isoforms are only very rarely associated with different biological processes, most projects do not profit much from quantifying isoforms. Hence, we would argue that quantifying expression levels of genes is the better option, as long as isoform quantification is not of explicit relevance for a project.

Another limitation is that all tag-protocols use poly(A) priming and hence do not capture mRNA from bacteria, organelles, or other non-polyadenylated transcripts. For full-length protocols like TruSeq, cDNA generation by random priming after rRNA depletion can be done. Another possibility is poly(A) tailing after rRNA depletion [68], but to our knowledge, this has not been adopted to tag-based protocols yet. How to efficiently combine profiling of polyadenylated, non-polyadenylated, and small RNA is certainly worth further investigating. However, it is also true that for eukaryotic cells, quantification of mRNAs contains most of the information. Hence, similar to the quantification of isoforms, we would argue that quantifying expression levels of genes by polyadenylated transcript is often sufficient, as long as non-polyadenylated transcripts are not explicitly relevant.

Furthermore, early barcoding and pooling necessitates calibrating input amounts. Input calibration is easy when starting with extracted RNA or when it is possible to count cells prior to direct lysis. When counting cells is not possible, we have also developed a protocol adaptation of prime-seq that allows for RNA quantification and normalization after bead-based RNA isolation and prior to reverse transcription (https://doi.org/10.17504/protocols.io.s9veh66).

Finally, early barcoding and pooling can lead to barcode swapping. We have shown that barcode swapping is not a major issue for prime-seq, but the amount of barcode swapping is unknown for most tag-protocols. However, even rather high levels of barcode swapping have a much smaller impact on power than a decrease in sample size (Additional file 1: Fig. S11E) and as long as the interpretation of absolute expression levels (e.g., presence/absence) is not crucial, the cost efficiency of tag-based protocols outweighs this drawback.

In summary, when quantification of isoforms and/or non-polyadenylated RNA is not necessary, a technically validated tag protocol has no drawbacks. Protocols that use poly(A) priming and template switching also have the advantage that they are very sensitive, and for prime-seq, we have validated that it still works optimally also with 1000 cells (~10–20 ng total RNA) as input. However, the decisive advantage of tag protocols is their drastically higher cost efficiency (Fig. 6), as this leads to drastically higher power and much more flexibility in the experimental design for a given budget. As repeated by biostatisticians over the decades, a good experimental design and a sufficient number of replicates is the most decisive factor for expression profiling. It is sobering how enduring the $n = 3$ tradition is, as is nicely shown in [16], although it is known that it is better to distribute the same number of reads across more biological replicates [17]. Cost-efficient tag protocols will hopefully make such experimental designs more common. While library costs are less notable for sequencing depths of 10 M reads or more (Fig. 6B), they may enable RNA-seq experiments that can be done with shallow sequencing, something which is less obvious and might be overlooked. Replacing qPCR has been advocated as one example by the authors of BRB-seq [22]. But also other applications, like characterizing cell type composition [36], quality control of libraries, or optimizing experimental procedures can profit considerably from low library costs.

In summary, tag protocols allow flexible designs of RNA-seq experiments that should be helpful for many biological questions and have a vast potential when readily accessible for many labs.

Validation, documentation, and cost efficiency make prime-seq a good option for setting up a tag protocol

We have argued above that adding a tag protocol to the standard method repertoire of a molecular biology lab is advantageous due to its cost efficiency. As the different tag protocols discussed here perform fairly similar with respect to complexity, power, accuracy, sensitivity, and cost efficiency, essentially any of them would suffice. If one has a validated, robust protocol running in a lab or core facility, it is probably not worth switching. That said, our results might still help to better validate existing protocols, integrate direct lysis, and make use of intronic reads. If one does not have a tag protocol running, we would argue that our results provide helpful information to decide on a protocol and that prime-seq would be a good option for several reasons as laid out in the following.

A main difference among tag protocols is whether they tag the 5′ end, like Decode-seq, or tag the 3′ end like BRB-seq or prime-seq. 5′ tagging has some obvious advantages (see also [16]), including the possibility to read both ends of the cDNA as one cannot read through the poly(A) tail. Using the sequence information from the 5′ end is also important to distinguish alleles of B-cell receptors and T-cell receptors [69]. In scRNA-seq, both 5′ and 3′ tag protocols have been successfully used, but 3′ tagging is currently the standard. The reason for this is not obvious, but it might be that the incorporation of the barcode and the UMI is more difficult to optimize [10]. Additionally, the higher level of alternative splicing at the 5′ end could make gene-level quantification more difficult. More dedicated comparisons would be needed to further investigate these factors. Currently, 3′ tag protocols are more established and when using a suitable sequencing design, poly(A) priming does not compromise sequencing quality as validated by us and the widespread use of Chromium 10x v3 chemistry scRNA-seq libraries that have the same layout as prime-seq.

As shown above, prime-seq is among all protocols the most cost-efficient when starting from purified RNA. It is also currently the only protocol for which a direct lysis is validated, which further increases cost efficiency of library production. This is especially advantageous when processing many samples, shallow sequencing is sufficient, and/or as sequencing costs continue to drop.

Finally, we think that prime-seq is the easiest tag protocol to set up. While many such protocols have been published and all have argued that their method would be useful, few have actually become widely implemented. The reasons are in all likelihood complex, but we think that prime-seq has the lowest barriers to be set up by an individual lab or a core facility for three reasons: First, to our knowledge, it is the most validated non-commercial bulk RNA-seq protocol, based on the experiments presented here as well as our >5 years of experience in running various versions of the protocol with over 6000 samples across 17 species resulting in over 20 publications to date. It is the only protocol for which direct lysis and sensitivity are quantitatively validated. Also, it is well validated in combination with zUMIs, the computational pipeline that was developed and is maintained by our group [45]. Second, it is not only cost-efficient per sample, but it also has low setup costs. It requires no specialized equipment and only the barcoded primers as an initial investment of ~$2000 for 96 primers, which will be sufficient for processing more than 240,000 samples. Finally, prime-seq is well documented not only by this manuscript, but also by a step-by-step protocol, including all materials, expected results, and alternative versions depending on the type and amounts of input material (https://doi.org/10.17504/protocols.io.s9veh66). Hence, we think that prime-seq is not only a very useful protocol in principle, but also in practice.

Conclusion

The multi-dimensional phenotype of gene expression is highly informative for many biological and medical questions. As sequencing costs dropped, RNA-seq became a standard tool in investigating these questions. We argue that the decisive next step is to use the possibilities of lowered library costs by tag protocols to leverage even more of this potential. We show that prime-seq is currently the best option when establishing such a protocol as it performs as well as other established RNA-seq protocols with respect to its accuracy, power, and library complexity. Additionally, it is very sensitive, is well documented, and is the most cost-efficient bulk RNA-seq protocol currently available to set up and to run.

Methods

A step-by-step protocol of prime-seq, including all materials and expected results, is available on protocols.io (https://doi.org/10.17504/protocols.io.s9veh66). Below, we briefly outline the prime-seq protocol, as well as describe any experiment-specific methods and modifications that were made to prime-seq during testing and optimization.

Prime-seq

Cell lysates, generally containing around 1000–10,000 cells, were treated with 20 µg of Proteinase K (Thermo Fisher, #AM2546) and 1 µL 25 mM EDTA (Thermo Fisher, EN0525) at 50 °C for 15 min with a heat inactivation step at 75 °C for 10 min. The samples were then cleaned using cleanup beads, a custom-made mixture containing Speed-Beads (GE65152105050250, Sigma-Aldrich), at a 1:2 ratio of lysate to beads. DNA was digested on-beads using 1 unit of DNase I (Thermo Fisher, EN0525) at 20 °C for 10 min with a heat inactivation step at 65 °C for 5 min.

The samples were then cleaned and the RNA was eluted with the 10 µL reverse transcription mix, consisting of 30 units Maxima H- enzyme (Thermo Fisher, EP0753), 1× Maxima H- Buffer (Thermo Fisher), 1 mM each dNTPs (Thermo Fisher), 1 µM template-switching oligo (IDT), and 1 µM barcoded oligo (dT) primers (IDT). The reaction was incubated at 42 °C for 90 min.

Following cDNA synthesis, the samples were pooled, cleaned, and concentrated with cleanup beads at a 1:1 ratio and eluted in 17 µL of ddH_2O. Residual primers were digested using Exonuclease I (Thermo Fisher, EN0581) at 37 °C for 20 min followed by a heat inactivation step at 80 °C for 10 min. The samples were cleaned once more using cleanup beads at a 1:1 ratio, and eluted in 20 µL of ddH_2O.

Second-strand synthesis and pre-amplification were performed in a 50 µL reaction, consisting of 1× KAPA HiFi Ready Mix (Roche, 7958935001) and 0.6 µM SingV6 primer (IDT), with the following PCR setup: initial denaturation at 98 °C for 3 min, denaturation at 98 °C for 15 s, annealing at 65 °C for 30 s, elongation at 68 °C for 4 min, and a final elongation at 72 °C for 10 min. Denaturation, annealing, and elongation were repeated for 5–15 cycles depending on the initial input.

The DNA was cleaned using cleanup beads at a ratio of 1:0.8 of DNA to beads and eluted with 10 µL of ddH_2O. The quantity was assessed using a Quant-iT PicoGreen dsDNA assay kit (Thermo Fisher, P11496) and the quality was assessed using an Agilent 2100 Bioanalyzer with a High-Sensitivity DNA analysis kit (Agilent, 5067-4626).

Libraries were prepared with the NEBNext Ultra II FS Library Preparation Kit (NEB, E6177S) according to the manufacturer instructions in most steps, with the exception of adapter sequence and reaction volumes. Fragmentation was performed on 2.5 µL of cDNA (generally 2–20 ng) using Enzyme Mix and Reaction buffer in a 6 µL reaction. A custom prime-seq adapter (1.5 µM, IDT) was ligated using the Ligation Master Mix and Ligation Enhancer in a reaction volume of 12.7 µL. The samples were then double-size selected using SPRI-select Beads (Beckman Coulter, B23317), with a high cutoff of 0.5 and a low cutoff of 0.7. The samples were then amplified using Q5 Master Mix (NEB, M0544L), 1 µL i7 Index primer (Sigma-Aldrich), and 1 µL i5 Index primer (IDT) using the following setup: 98 °C for 30 s; 10–12 cycles of 98 °C for 10 s, 65 °C for 1 min 15 s, 65 °C for 5 min; and 65 °C for 4 min. Double-size selection was performed once more as before using SPRI-select Beads. The quantity and quality of the libraries were assessed as before.

Nextera XT Library Prep

Prior to using the NEBNext Ultra II FS Library Kit, libraries were prepared using the Nextera XT Kit (Illumina, FC-131-1096). This included the RNA extraction experiments (Fig. 4) as well as the AML experiment (Fig. 5B). These libraries were prepared as previously described [11].

Briefly, three replicates of 0.8 ng of DNA were tagmented in 20 μL reactions. Following tagmentation, the libraries were amplified using 0.1 μM P5NextPT5 primer (IDT) and 0.1 μM i7 index primer (IDT) in a reaction volume of 50 μL. The index PCR was incubated as follows: gap fill at 72 °C for 3 min, initial denaturation at 95 °C for 30 s, denaturation at 95 °C for 10 s, annealing at 62 °C for 30 s, elongation at 72 °C for 1 min, and a final elongation at 72 °C for 5 min. Denaturation, annealing, and elongation were repeated for 13 cycles.

Size selection was performed using gel electrophoresis. Libraries were loaded onto a 2% Agarose E-Gel EX (Invitrogen, G401002) and were excised between 300 and 900 bp and cleaned using the Monarch DNA Gel Extraction Kit (NEB, T1020). The libraries were quantified and qualified using an Agilent 2100 Bioanalyzer with a High-Sensitivity DNA analysis kit (Agilent, 5067-4626).

Barcoded oligo (dT) primer design

In order to enable more robust demultiplexing and to ensure full compatibility of our sequencing layout with the Chromium 10x v3 chemistry, oligo (dT) primers were designed to include a 12 nt cell barcode and 16 nt UMI. Candidate cell barcodes were created in R using the DNABarcodes package [70] to generate barcodes with a length of 12 nucleotides and a minimum Hamming distance (HD) of 4, with filtering for self-complementarity, homo-triplets, and GC-balance enabled. Candidate barcodes were filtered further, resulting in a barcode pool with a minimal HD of 5 and a minimal Sequence-Levenshtein distance of 4 within the set. In order to balance nucleotide compositions among cell barcodes at each position, BARCOSEL [71] was used to further reduce the candidate set down to the final 384 barcodes.

Sequencing

Sequencing was performed on an Illumina HiSeq 1500 instrument for all libraries except for the IPSC/NPC experiment where a NextSeq 550 instrument was used. The following setup was used: Read 1: 28 bp, Index 1: 8 bp; Read 2: 50-56 bp.

Pre-processing of RNA-seq data

The raw data was quality checked using fastqc (version 0.11.8 [72]) and then trimmed of poly(A) tails using Cutadapt (version 1.12, https://doi.org/10.14806/ej.17.1.200). Following trimming, the zUMIs pipeline (version 2.9.4 ,[45]) was used to filter the data, with a Phred quality score threshold of 20 for 2 BC bases and 3 UMI bases. The filtered data was mapped to the human genome (GRCh38) with the Gencode annotation (v35) or the mouse genome (GRCm38) with the Gencode annotation (vM25) using STAR (version 2.7.3a,[73]) and the reads counted using RSubread (version 1.32.4,[74]).

Sensitivity and differential gene expression analysis of RNA-seq data

The count matrix generated by zUMIs was loaded into RStudio (version 1.3.1093 [75]) using R (version 4.0.3 [76]). bioMart (version 2.46.0 [77]), dplyr (version 1.0.2 [78]), and tidyr (version 1.1.2 [79]) were used for data processing and calculating descriptive statistics (i.e., detected genes, reads, and UMIs). DESeq2 (version 1.30.0 [80]) was used for differential gene expression analysis. ggplot2 (version 3.3.3 [81]), cowplot (version 1.1.1 [82]), ggbeeswarm (0.6.0 [83]), ggsignif (version 0.6.0 [84]), ggsci (version 2.9 [85]), ggrepel (version 0.9.0 [86]), EnhancedVolcano (1.8.0 [87]), ggpointdensity (version 0.1.0 [88]), and pheatmap (version 1.0.12 [89]) were used for data visualization.

Power analysis of RNA-seq data

Power simulations were performed following the workflow of the powsimR package (version 1.2.3 [49]). Briefly, RNA-seq data per method was simulated based on parameters extracted from the UHRR comparison experiment. For each method and sample size setup (6 vs. 6, 12 vs. 12, 24 vs. 24, and 48 vs. 48), 20 simulations were performed with the following settings: normalization = “MR,” RNA-seq = “bulk,” Protocol = “Read/UMI,” Distribution = “NB,” ngenes = 30000, nsims = 20, p.DE = 0.10. We verified with the data generated from the AML and NPC differentiation data that the gamma distribution (shape = 1, scale = 0.5) would be an appropriate log fold change distribution in this case (Additional file 1: Fig. S7A).

To simulate contamination by cross-contamination, we assumed that contamination increases with expression as shown in Additional file 1: Fig. S11D and can thus be simulated by sampling from the overall counts per gene in a pool. Different levels of contamination (0.5%, 1%, 2.5%, 5%, 10%) were simulated and added to the original count matrix. Power simulations were run as described above.

Cell preparation

Human embryonic kidney 293T (HEK293T) cells were cultured in DMEM media (TH.Geyer, L0102) supplemented with 10% FBS (Thermo Fisher, 10500-064) and 100 U/ml Penicillin and 100 µg/ml Streptomycin (Thermo Fisher). Cells were grown to 80% confluency and harvested by trypsinization (Thermo Fisher, 25200072).

Peripheral blood mononuclear cells (PBMCs) were obtained from LGC Standards (PCS-800-011). Before use, the cells were thawed in a water bath at 37 °C and washed twice with PBS (Sigma-Aldrich, D8537).

Prior to lysis, cells were stained with 1 µg/ml Trypan Blue (Thermo Fisher Scientific, 15-250-061) and counted using a Neubauer counting chamber. Then, the desired number of cells (1000 or 10,000) was pelleted for 5 min at 200 rcf, resuspended in 50 µL of lysis buffer (RLT Plus (Qiagen, 1053393) and 1% ß-mercaptoethanol (Sigma-Aldrich,M3148) and transferred to a 96-well plate. Samples were then stored at − 80 °C until needed.

Tissue preparation

Striatal tissue from C57BL/6 mice between the ages of 6 and 12 months was harvested by first placing the mouse in a container with Isoflurane (Abbot, TU 061220) until the mouse was visibly still and exhibited labored breathing. The mice were then removed from the container, and a cervical dislocation was performed. The mice were briefly washed with 80% EtOH, the head decapitated, and the brain removed. The brain was transferred to a dish with ice-cold PBS and placed in a 1-mm slicing matrix.

Using steel blades (Wilkinson Sword, 19/03/2016DA), 5 coronal incisions were made. Biopsy punches (Kai Medical, BPP-20F) were then taken from the striatum and the tissue was transferred to a 1.5-mL tube with 50 µL of lysis buffer, RLT Plus, and 1% ß-mercaptoethanol. The tubes were snap frozen and stored at − 80 °C until needed.

RNA extraction experiments

To determine differences due to RNA extraction, we isolated RNA using columns from the Direct-zol RNA MicroPrep Kit (Zymo, R2062) (condition: "Column") and magnetic beads from the prime-seq protocol (conditions: "No Incubation," "Incubation," and "Magnetic Beads") (see above for details on prime-seq). For the "Column" condition, the manufacturer instructions were followed and both the Proteinase K and DNase digestion steps were performed as outlined in the protocol. For the magnetic bead isolation, the prime-seq protocol was used as outlined in the "Magnetic Beads" condition. For "No Incubation" condition, the Proteinase K digestion was skipped entirely. For the "Incubation" condition, the Proteinase K digestion was performed but with no enzyme; that is the heat cycling of 50 °C for 15 min and 75 °C for 10 min was carried out but no enzyme was added to the lysate.

gDNA priming experiment

For a graphical overview of the gDNA Priming experiment, see Fig. 2B. Frozen vials of mouse embryonic stem cells (mESC), which have been cultured as previously described (citation Bagnoli) (clone J1, frozen in Bambanker (NIPPON Genetics, BB01) on 04.2017), and HEK293T cells (frozen in Bambanker on 30.11.18, passage 25) were thawed. DNA was extracted from 1 million mESCs using DNeasy Blood & Tissue Kit (Qiagen, 69506) and RNA was extracted from 450,000 HEK293T cells using the Direct-zol RNA MicroPrep Kit (Zymo, R2062), according to the manufacturer instructions in both cases. The optional DNase treatment step during the RNA extraction was performed in order to remove any residual DNA.

After isolating DNA and RNA, the two were mixed to obtain the following conditions: 10 ng RNA/ 7 ng DNA, 7.5 ng RNA/ 1.75 ng DNA, and 10 ng RNA/ 0 ng DNA. The 10 ng RNA/ 7 ng DNA condition, which represents the highest contamination of DNA, was performed twice, once without DNase treatment and once with DNase treatment. Libraries were prepared from three replicates for each condition using prime-seq and were then sequenced (see above for detailed information).

MAQC-III comparison experiment

For a graphical overview of the experimental design, see Additional file 1: Fig. S5. As only Mix A from the original MAQC-III Study was compared, 122.2 µL of ddH$_2$O, 2.8 µL of UHRR (100 ng/µL) (Thermo Fisher, QS0639), and 2.5 µL of ERCC Mix 1 (1:1000) (Thermo Fisher, 4456740) were combined to generate a 1:500 dilution of Mix A. Eight

RNA-seq libraries were constructed using prime-seq (see above methods) with 5 μL of the 1:500 Mix A.

The samples were sequenced and the data processed and analyzed as outlined above. Of the comparison data from the original MAQC-III Study, Experiment SRX302130 to SRX302209 from Submission SRA090948 were used as this was the sequence data from one site (BGI) and was sequenced using an Illumina HiSeq 2000 [48]. The TruSeq data was first trimmed to be 50 bp long and then processed with zUMIs as outlined above, with the exception of using both cDNA reads and not providing UMIs as there were none. Paired-end data was used to not penalize TruSeq, as this is a feature of the method.

Barcode swapping experiments

In order to estimate cross-contamination levels in prime-seq introduced by barcode swapping, we isolated RNA from human-induced pluripotent stem cells (line 29B5, passage 34) [60] and mouse ES cells (line JM8, passage 27) [2] using the Direct-zol RNA MicroPrep Kit (Zymo, R2062). RNA concentrations were measured using the Quanti-Flour RNA Dye (Promega, E3310) and 8 ng of total RNA were added per well. For the experiment estimating the impact of amplification on contamination, different nanograms of RNA per well (0.5, 2, 8, 32, 128) were amplified with different numbers of cycles (17, 15, 13, 11, 9). Prime-seq was performed as described before with pooling of samples from the different species (Additional file 1: Fig. S11A). Contamination was assessed by mapping to a concatenated human and mouse genome and assigning reads to species based on which genome they mapped to best.

NPC differentiation experiment

To differentiate hiPSCs to NPCs, 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). A half-medium change was performed on days 2 and 4. On day 6, Neurospheres from 3 columns were pooled, dissociated using Accumax (Sigma-Aldrich) and seeded on Geltrex (Thermo Fisher) coated wells. After 2 days, cells were dissociated and counted and 2×10^4 were lysed in 100 μL of lysis buffer (RLT Plus (Qiagen, 1053393) and 1% ß-mercaptoethanol (Sigma-Aldrich,M3148).

AML-PDX sample collection

Acute myeloid leukemia (AML) cells were engrafted in NSG mice (The Jackson Laboratory, Bar Harbour, ME, USA) to establish patient-derived xenograft (PDX) cells [55]. AML-PDX cells were cryopreserved as 10 Mio cells in 1 mL of freezing medium (90% FBS, 10% DMSO) and stored at – 80 °C for biobanking purposes. To avoid thawing these samples and thus harming or even destroying them, the frozen cell stocks were first transferred to dry ice under a cell culture hood. Next a sterile 1-mm biopsy punch was used to punch the frozen cells in the vial and transfer the extracted cells to one well of a 96-well plate containing 100 μL RLTplus lysis buffer with 1% beta mercaptoethanol. To ensure complete lysis, the lysate was mixed and snap frozen on dry ice. One biopsy punch is estimated to contain 10 µL of cryopreserved cells corresponding to roughly 1 × 10^5 cells given an even distribution of cells within the original vial. All 96 samples were collected in this manner, biopsy punches were washed using RNAse Away (Thermo Fisher Scientific) and 80% Ethanol for reuse. These lysates were subjected to prime-seq, including RNA isolation using SPRI beads. In total, PDX samples from 11 different AML patients were analyzed in 6 to 16 biological replicates (engrafted mice) per sample.

Cost comparisons

Costs were determined by searching for general list prices from various vendors. When step by step protocols were available, each component was included in the cost calculation, such as for the SMARTer Stranded Total RNA Kit (Takara, 634862), SMART-Seq RNA Kit (v4) (Takara, 634891), TruSeq Library Prep (Illumina, RS-122-2001/2), TruSeq Stranded Library Prep (Illumina, 20020595), and Illumina Stranded mRNA Prep (Illumina, 20040534). In the case of BRB-seq, no publicly available step-by-step protocol was found, so the methods section was used to calculate costs [22]. Decode-seq has a publicly available protocol; however, the level of detail was insufficient to calculate exact costs; therefore, when specific vendors were not listed, we used the most affordable option that we have previously validated. In all cases, the prices included sales tax and were listed in euros and were therefore converted to USD using a conversion rate of 1.23 USD to EUR. The costs for all methods can be found in Table S4.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02660-8.

Additional file 1: Fig. S1. Molecular workflow of prime-seq. **Fig. S2**. prime-seq is a robust protocol and has been validated with numerous organisms. **Fig. S3**. Intronic reads are not derived from contaminating gDNA. **Fig. S4**. Intron counts are enriched at the 3' prime end and correlate with exon counts. **Fig. S5**. Experimental design comparing prime-seq to TruSeq data generated in the MAQC-III Study. **Fig. S6**. prime-seq and TruSeq have similar mapping, gene detection, and expression. **Fig. S7**. Power and FDR mostly depend on sample size and are similar between prime-seq and TruSeq. **Fig. S8**. Performance of isolation methods is similar independent of prefiltering or usage of only Exon data. **Fig. S9**. Most genes are detected independent of the extraction method used. **Fig. S10**. prime-seq performs equally well with high- and low-input samples. **Fig. S11**. Cross-contamination levels are low, increase with additional cycles but do not impact power simulations. **Fig. S12**. Power analysis shows prime-seq is able to reach 80% power earlier than less cost-efficient methods.

Additional file 2: Table S1. (Sensitivity) List of experiments performed with prime-seq including key characteristics of the experiments and data quality.

Additional file 3: Supplemental Text. Magnetic Beads used in prime-seq.

Additional file 4: Table S2. (Lysis Costs) Calculations for per sample costs of different commercial and non-commercial extraction methods.

Additional file 5: Table S3. (Lysis Time) Time needed for extraction of 24, 48 and 96 samples with SPRI beads or Silica Columns.

Additional file 6: Table S4. (Method Cost) Per sample cost calculations for popular commercial and non-commercial RNA-seq methods including all consumables and reagents.

Additional File 7: Table S5. (Method Time) Time needed for performing for popular commercial and non-commercial RNA-seq methods.

Additional file 8. Review history.

Acknowledgements

We would like to thank Karin Bauer and Ming Zhao for lab support, Ines Bliesener and Maike Fritschle for animal work, Sabrina Schenk, Irena Stähler, and the staff at the LMU Biology Faculty Animal Facility for mouse colony maintenance, Dr. Stefan Krebs and the staff of LAFUGA for sequencing services, and Dr. Boyan Bonev and his lab for suggesting the Ultra II FS Kit as an alternative to tagmentation. Some illustrations in Fig. 1, Fig. 3, and Additional file 1: Fig. S2 were created with BioRender.com.

Review history

The review history is available as Additional file 8.

Peer review information

Wenjing She 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

AJ, LEW, CZ, and WE conceived the study. JG, AJ, and PN prepared iPSC, HEK293T, and tissue samples. JG performed differentiation experiments. BVick and IJ generated AML-PDX samples. DR and JWB designed the barcoded primers. AJ, LEW, JWB, and PN conducted the RNA-seq experiments. AJ and LEW performed sensitivity and gene expression analysis. LEW performed power analysis. BVieth and IH provided computational and statistical support. AJ, LEW, JWB, and WE wrote the manuscript. All authors read and approved the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the LMU Excellence Initiative, SFB1243 (Subproject A05/A14/A15), DFG EN 1093/2-1 (project number 406901759), and the Cyliax foundation.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the ArrayExpress repository under the following accession numbers E-MTAB-10133, 10138-10142, 10175, 11455, 11456 [90–98]. The MAQC-III Study, Experiment SRX302130 to SRX302209 from Submission SRA090948 were retrieved from the short-read archive [99]. The code required to generate the figures can be found at https://github.com/Hellmann-Lab/prime-seq [100] (published under GPL-3 License). A stable version of the github repository is available through zenodo (https://doi.org/10.5281/zenodo.5932624) [101].

Declarations

Ethics approval and consent to participate

The human iPSC samples, which were differentiated into the NPCs, were ethically approved by the responsible committee on human experimentation (20-122, Ethikkommission LMU München) as previously published [60].

Bone marrow (BM) and peripheral blood (PB) samples from AML patients were obtained from the Department of Internal Medicine III, Ludwig-Maximilians-Universität, Munich, Germany. Specimens were collected for diagnostic purposes. Written informed consent was obtained from the patients. The study was performed in accordance with the ethical standards of the responsible committee on human experimentation (written approval by the Research Ethics Boards of the medical faculty of Ludwig-Maximilians-Universität, Munich, number 068-08 and 222-10) and with the Helsinki Declaration of 1975, as revised in 2013. All animal trials were performed in accordance with the current ethical standards of the official committee on animal experimentation (written approval by Regierung von Oberbayern, tierversuche@reg-ob.bayern.de; ROB-55.2Vet-2532.Vet_02-16-7 and ROB-55.2Vet-2532.Vet_03-16-56).

The mouse brain tissues were collected from mice that were bred and housed at the Biology Faculty Animal Facility at Ludwig Maximilian University in accordance with institutional ethical standards. The animal tissue was harvested according to the German Animal Welfare Act Paragraph 4 (organ removal for scientific reasons).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Anthropology & Human Genomics, Faculty of Biology, Ludwig-Maximilians University, Großhaderner Str. 2, 82152 Martinsried, Germany. [2] Graduate School of Systemic Neurosciences, Faculty of Biology, Ludwig-Maximilians University, Martinsried, Germany. [3] Research Unit Apoptosis in Hematopoietic Stem Cells, Helmholtz Zentrum München, German Research Center for Environmental Health (HMGU), Munich, Germany. [4] German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany. [5] Department of Pediatrics, Dr. von Hauner Children's Hospital, Ludwig-Maximilians University, Munich, Germany. [6] Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden.

Received: 27 September 2021 Accepted: 23 March 2022

Published online: 31 March 2022

References

1. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20:631–56.
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;65:631–43.e4.
3. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat Commun. 2019;10:4667.
4. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13:599–604.

5. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat Biotechnol. 2020;38:747–55.
6. Ziegenhain C, Vieth B, Parekh S, Hellmann I, Enard W. Quantitative single-cell transcriptomics. Brief Funct Genomics. 2018;17:220–32.
7. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. nature.com. 2011;9:72–4.
8. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012;2:666–73.
9. 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.
10. Hagemann-Jensen M, Ziegenhain C, Chen P, Ramsköld D, Hendriks G-J, Larsson AJM, et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nat Biotechnol. 2020;38:708–14.
11. Bagnoli JW, Ziegenhain C, Janjic A, Wange LE, Vieth B, Parekh S, et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. Nat Commun. 2018;9:2937.
12. 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.
13. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.
14. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201.
15. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commun. 2020;11:5650.
16. Li Y, Yang H, Zhang H, Liu Y, Shang H, Zhao H, et al. Decode-seq: a practical approach to improve differential gene expression analysis. Genome Biol. 2020;21:66.
17. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014;30:301–4.
18. Lazic SE, Clarke-Williams CJ, Munafò MR. What exactly is "N" in cell culture and animal experiments? PLoS Biol. 2018;16:e2005282.
19. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171:1437–52.e17.
20. Uzbas F, Opperer F, Sönmezer C, Shaposhnikov D, Sass S, Krendl C, et al. BART-Seq: cost-effective massively parallelized targeted sequencing for genomics, transcriptomics, and single-cell analysis. Genome Biol. 2019;20:155.
21. Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Zachery Cogan J, et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat Biotechnol. 2020;38:954–61 Nature Publishing Group.
22. Alpern D, Gardeux V, Russeil J, Mangeat B, Meireles-Filho ACA, Breysse R, et al. BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. Genome Biol. 2019;20:71.
23. Ebinger S, Özdemir EZ, Ziegenhain C, Tiedt S, Castro Alves C, Grunert M, et al. Characterization of rare, dormant, and therapy-resistant cells in acute lymphoblastic leukemia. Cancer Cell. 2016;30:849–62.
24. Schreck C, Istvánffy R, Ziegenhain C, Sippenauer T, Ruf F, Henkel L, et al. Niche WNT5A regulates the actin cytoskeleton during regeneration of hematopoietic stem cells. J Exp Med. 2017;214:165–81.
25. Gegenfurtner FA, Zisis T, Al Danaf N, Schrimpf W, Kliesmete Z, Ziegenhain C, et al. Transcriptional effects of actin-binding compounds: the cytoplasm sets the tone. Cell Mol Life Sci. 2018;75:4539–55.
26. Gegenfurtner FA, Jahn B, Wagner H, Ziegenhain C, Enard W, Geistlinger L, et al. Micropatterning as a tool to identify regulatory triggers and kinetics of actin-mediated endothelial mechanosensing. J Cell Sci. 2018;131. Available from:. https://doi.org/10.1242/jcs.212886.
27. Mueller S, Engleitner T, Maresch R, Zukowska M, Lange S, Kaltenbacher T, et al. Evolutionary routes and KRAS dosage define pancreatic cancer phenotypes. Nature. 2018;554:62–8.
28. Wang S, Crevenna AH, Ugur I, Marion A, Antes I, Kazmaier U, et al. Actin stabilizing compounds show specific biological effects due to their binding mode. Sci Rep. 2019;9:9731.
29. Wang S, Gegenfurtner FA, Crevenna AH, Ziegenhain C, Kliesmete Z, Enard W, et al. Chivosazole A modulates protein-protein interactions of actin. J Nat Prod. 2019;82:1961–70.
30. Ebinger S, Zeller C, Carlet M, Senft D, Bagnoli JW, Liu W-H, et al. Plasticity in growth behavior of patients' acute myeloid leukemia stem cells growing in mice. Haematologica. 2020;105:2855–60.
31. Garz A-K, Wolf S, Grath S, Gaidzik V, Habringer S, Vick B, et al. Azacitidine combined with the selective FLT3 kinase inhibitor crenolanib disrupts stromal protection and inhibits expansion of residual leukemia-initiating cells in FLT3-ITD AML with concurrent epigenetic mutations. Oncotarget. 2017;8:108738–59.
32. Mulholland CB, Nishiyama A, Ryan J, Nakamura R, Yiğit M, Glück IM, et al. Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals. Nat Commun. 2020;11:5972.
33. Redondo Monte E, Wilding A, Leubolt G, Kerbs P, Bagnoli JW, Hartmann L, et al. ZBTB7A prevents RUNX1-RUNX1T1-dependent clonal expansion of human hematopoietic stem and progenitor cells. Oncogene. 2020;39:3195–205.
34. Shami A, Atzler D, Bosmans LA, Winkels H, Meiler S, Lacy M, et al. Glucocorticoid-induced tumour necrosis factor receptor family-related protein (GITR) drives atherosclerosis in mice and is associated with an unstable plaque phenotype and cerebrovascular events in humans. Eur Heart J. 2020;41:2938–48.
35. LaClair KD, Zhou Q, Michaelsen M, Wefers B, Brill MS, Janjic A, et al. Congenic expression of poly-GA but not poly-PR in mice triggers selective neuron loss and interferon responses found in C9orf72 ALS. Acta Neuropathol. 2020;140:121–42.
36. Geuder J, Ohnuki M, Wange LE, Janjic A, Bagnoli JW, Müller S, et al. A non-invasive method to generate induced pluripotent stem cells from primate urine: Cold Spring Harbor Laboratory; 2020. p. 2020.08.12.247619. [cited 2021 Jan 21] Available from: https://www.biorxiv.org/content/10.1101/2020.08.12.247619v1

37. Alterauge D, Bagnoli JW, Dahlström F, Bradford BM, Mabbott NA, Buch T, et al. Continued Bcl6 expression prevents the transdifferentiation of established Tfh cells into Th1 cells during acute viral infection. Cell Rep. 2020;33:108232.
38. Kempf J, Knelles K, Hersbach BA, Petrik D, Riedemann T, Bednarova V, et al. Heterogeneity of neurons reprogrammed from spinal cord astrocytes by the proneural factors Ascl1 and Neurogenin2. Cell Rep. 2021;36:109409.
39. Porquier A, Tisserant C, Salinas F, Glassl C, Wange L, Enard W, et al. Retrotransposons as pathogenicity factors of the plant pathogenic fungus Botrytis cinerea. Genome Biol. 2021;22:1–19 BioMed Central.
40. Carlet M, Völse K, Vergalli J, Becker M, Herold T, Arner A, et al. In vivo inducible reverse genetics in patients' tumors to identify individual therapeutic targets. bioRxiv. 2020:2020.05.02.073577 [cited 2021 Sep 3]. Available from: https://www.biorxiv.org/content/10.1101/2020.05.02.073577v1.
41. Kempf JM, Weser S, Bartoschek MD, Metzeler KH, Vick B, Herold T, et al. Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance in AML. Sci Rep. 2021;11:5838.
42. Pekayvaz K, Leunig A, Kaiser R, Brambs S, Joppich M, Janjic A, et al. Protective immune trajectories in early viral containment of non-pneumonic SARS-CoV-2 infection: Cold Spring Harbor Laboratory; 2021. p. 2021.02.03.429351. [cited 2021 Feb 19]. Available from: https://www.biorxiv.org/content/10.1101/2021.02.03.429351v1
43. Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Hülsmann M, et al. TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals: Cold Spring Harbor Laboratory; 2021. p. 2021.02.05.429919. [cited 2021 Feb 19]. Available from: https://www.biorxiv.org/content/10.1101/2021.02.05.429919v2
44. Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. Characterization of directed differentiation by high-throughput single-cell RNA-Seq: Cold Spring Harbor Laboratory; 2014. p. 003236. [cited 2021 Jan 21]. Available from: http://biorxiv.org/content/early/2014/03/05/003236.abstract
45. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. Gigascience. 2018;7. Available from:. https://doi.org/10.1093/gigascience/giy059.
46. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560:494–8.
47. Lee S, Zhang AY, Su S, Ng AP, Holik AZ, Asselin-Labat M-L, et al. Covering all your bases: incorporating intron signal from RNA-seq data. NAR Genom Bioinform. 2020;2 [cited 2021 Jan 21]. Oxford Academic; Available from: https://academic.oup.com/nargab/article-pdf/2/3/lqaa073/34054975/lqaa073.pdf.
48. Xu J, Su Z, Hong H, Thierry-Mieg J, Thierry-Mieg D, Kreil DP, et al. Cross-platform ultradeep transcriptomic profiling of human reference RNA samples by RNA-Seq. Sci Data. 2014;1:140020.
49. Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: power analysis for bulk and single cell RNA-seq experiments. Bioinformatics. 2017;33:3486–8.
50. Oberacker P, Stepper P, Bond DM, Höhn S, Focken J, Meyer V, et al. Bio-On-Magnetic-Beads (BOMB): Open platform for high-throughput nucleic acid extraction and manipulation. PLoS Biol. 2019;17:e3000107.
51. Scholes AN, Lewis JA. Comparison of RNA isolation methods on RNA-Seq: implications for differential expression and meta-analyses. BMC Genomics. 2020;21:249.
52. 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 [cited 2020 Feb 17]. Available from: https://www.biorxiv.org/content/10.1101/791699v1.abstract.
53. Dixit A. Correcting chimeric crosstalk in single cell RNA-seq experiments. bioRxiv. 2021:093237 [cited 2021 Aug 26]. Available from: https://www.biorxiv.org/content/10.1101/093237v2.
54. Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol. 2016;34:942–9.
55. Vick B, Rothenberg M, Sandhöfer N, Carlet M, Finkenzeller C, Krupka C, et al. An advanced preclinical mouse model for acute myeloid leukemia using patients' cells of various genetic subgroups and in vivo bioluminescence imaging. PLoS One. 2015;10:e0120925.
56. Herold T, Jurinovic V, Batcha AMN, Bamopoulos SA, Rothenberg-Thurley M, Ksienzyk B, et al. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. Haematologica. 2018;103:456–65.
57. Chambers SM, Fasano CA, Papapetrou EP, Tomishima M, Sadelain M, Studer L. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat Biotechnol. 2009;27:275–80.
58. Liu Y, Yu C, Daley TP, Wang F, Cao WS, Bhate S, et al. CRISPR activation screens systematically identify factors that drive neuronal fate and reprogramming. Cell Stem Cell. 2018;23:758–71.e8.
59. Özdemir EZ, Ebinger S, Ziegenhain C, Enard W, Gires O, Schepers A, et al. Drug resistance and dormancy represent reversible characteristics in patients' ALL cells growing in mice. Blood. 2016;128:602 American Society of Hematology.
60. Geuder J, Wange LE, Janjic A, Radmer J, Janssen P, Bagnoli JW, et al. A non-invasive method to generate induced pluripotent stem cells from primate urine. Sci Rep. 2021;11:3516.
61. Sholder G, Lanz TA, Moccia R, Quan J, Aparicio-Prat E, Stanton R, et al. 3'Pool-seq: an optimized cost-efficient and scalable method of whole-transcriptome gene expression profiling. BMC Genomics. 2020;21:64.
62. Ye C, Ho DJ, Neri M, Yang C, Kulkarni T, Randhawa R, et al. DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. Nat Commun. 2018;9:4307.
63. Pandey S, Takahama M, Gruenbaum A, Zewde M, Cheronis K, Chevrier N. A whole-tissue RNA-seq toolkit for organism-wide studies of gene expression with PME-seq. Nat Protoc. 2020;15:1459–83.
64. Kamitani M, Kashima M, Tezuka A, Nagano AJ. Lasy-Seq: a high-throughput library preparation method for RNA-Seq and its application in the analysis of plant responses to fluctuating temperatures. Sci Rep. 2019;9:7091.
65. Giraldez MD, Spengler RM, Etheridge A, Godoy PM, Barczak AJ, Srinivasan S, et al. Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. Nat Biotechnol. 2018;36:746–57.
66. Xiong Y, Soumillon M, Wu J, Hansen J, Hu B, van Hasselt JGC, et al. A comparison of mRNA sequencing with random primed and 3'-directed libraries. Sci Rep. 2017;7:14626.

67. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096–8.
68. Westermann AJ, Vogel J. Cross-species RNA-seq for deciphering host-microbe interactions. Nat Rev Genet. 2021;22:361–78.
69. Trück J, Eugster A, Barennes P, Tipton CM, Luning Prak ET, Bagnara D, et al. Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling. Elife. 2021;10. Available from:. https://doi.org/10.7554/eLife.66274.
70. Buschmann T, Bystrykh LV. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. BMC Bioinformatics. 2013;14:272.
71. Somervuo P, Koskinen P, Mei P, Holm L, Auvinen P, Paulin L. BARCOSEL: a tool for selecting an optimal barcode set for high-throughput sequencing. BMC Bioinformatics. 2018;19:257.
72. Andrews S. FastQC: A quality control analysis tool for high throughput sequencing data. Github; [cited 2021 Sep 14]. Available from: https://github.com/s-andrews/FastQC
73. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
74. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 2019;47:e47.
75. Team R. RStudio: Integrated Development for R. Boston: RStudio, PBC; 2020. p. 2020.
76. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2016. Available from: https://www.r-project.org/
77. Steffen Durinck, Wolfgang Huber. biomaRt. Bioconductor; 2017. Available from: https://bioconductor.org/packages/biomaRt
78. Wickham H, Francois R, Henry L, Müller K. dplyr: a grammar of data manipulation. 2021. Available from: https://github.com/tidyverse/dplyr
79. Wickham H, Henry L. Tidyr: Tidy messy data. R package version, vol. 1; 2020. p. 397.
80. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
81. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer; 2010.
82. Wilke CO. cowplot: streamlined plot theme and plot annotations for "ggplot2."; 2019.
83. Clarke E, Sherrill-Mix S. ggbeeswarm: Categorical Scatter (Violin Point) Plots . 2017. Available from: https://CRAN.R-project.org/package=ggbeeswarm
84. Constantin A-E, Patil I. ggsignif: R Package for Displaying Significance Brackets for "ggplot2". PsyArxiv. 2021. Available from: https://psyarxiv.com/7awm6
85. Xiao N. ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for "ggplot2". 2018. Available from: https://CRAN.R-project.org/package=ggsci
86. Slowikowski K. ggrepel: Automatically position non-overlapping text labels with "ggplot2."; 2018.
87. Blighe K, Rana S, Lewis M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version; 2019.
88. Kremer LPM. ggpointdensity: a cross between a 2D density plot and a scatter plot. 2019. Available from: https://CRAN.R-project.org/package=ggpointdensity
89. Kolde R. Pheatmap: pretty heatmaps [Internet]. 2012. Available from: https://cran.r-project.org/web/packages/pheatmap/index.html
90. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. Impact of RNA isolation methods for RNA-seq on gene expression. (HEK293T). E-MTAB-10142: Array Express; https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10142/. Accessed 6 Mar 2022.
91. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. Impact of RNA isolation methods for RNA-seq on gene expression (mouse striatal tissue). E-MTAB-10140: Array Express; https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10140/. Accessed 6 Mar 2022.
92. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. Impact of RNA isolation methods for RNA-seq on gene expression. (PBMCs). E-MTAB-10138: Array Express; https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10138/. Accessed 6 Mar 2022.
93. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. RNA-seq of human RNA contaminated with different amounts of mouse gDNA to quantify the impact of gDNA contamination in prime-seq. E-MTAB-10141: Array Express; https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10141/. Accessed 6 Mar 2022.
94. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. Deep RNA-seq of Universal Human Reference RNA mixed with external spike in molecules ERCC mix 1 using prime-seq. E-MTAB-10139: Array Express; https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10139/. Accessed 6 Mar 2022.
95. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. Bulk RNA-seq of archived acute myeloid leukemia (AML) samples propagated in a mouse Xenograft model over several passages. E-MTAB-10175: Array Express; https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10175/. Accessed 6 Mar 2022.
96. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. Bulk RNA-seq of human induced pluripotent stem cells (hIPSC) and neural progenitor cells (NPC) differentiated using Dual SMAD inhibition using the prime-seq method. E-MTAB-10133: Array Express; https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10133/. Accessed 6 Mar 2022.
97. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. Human-Mouse Mixture experiment to estimate that contribution of Barcode swapping. E-MTAB-11455: Array Express; https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-11455/. Accessed 6 Mar 2022.
98. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. Human-Mouse Mixture experiment to estimate that contribution of Barcode swapping. E-MTAB-11456: Array Express; https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-11456/. Accessed 6 Mar 2022.

99. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. PRJNA208369. BioProject; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA208369. Accessed 18 Sept 2019.
100. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. prime-seq: prime-seq paper analysis: Github; 2022. https://github.com/Hellmann-Lab/prime-seq
101. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. prime-seq: prime-seq paper analysis (zenodo): Zenodo; 2022. https://zenodo.org/record/5932624

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

Supplementary Figures and Tables

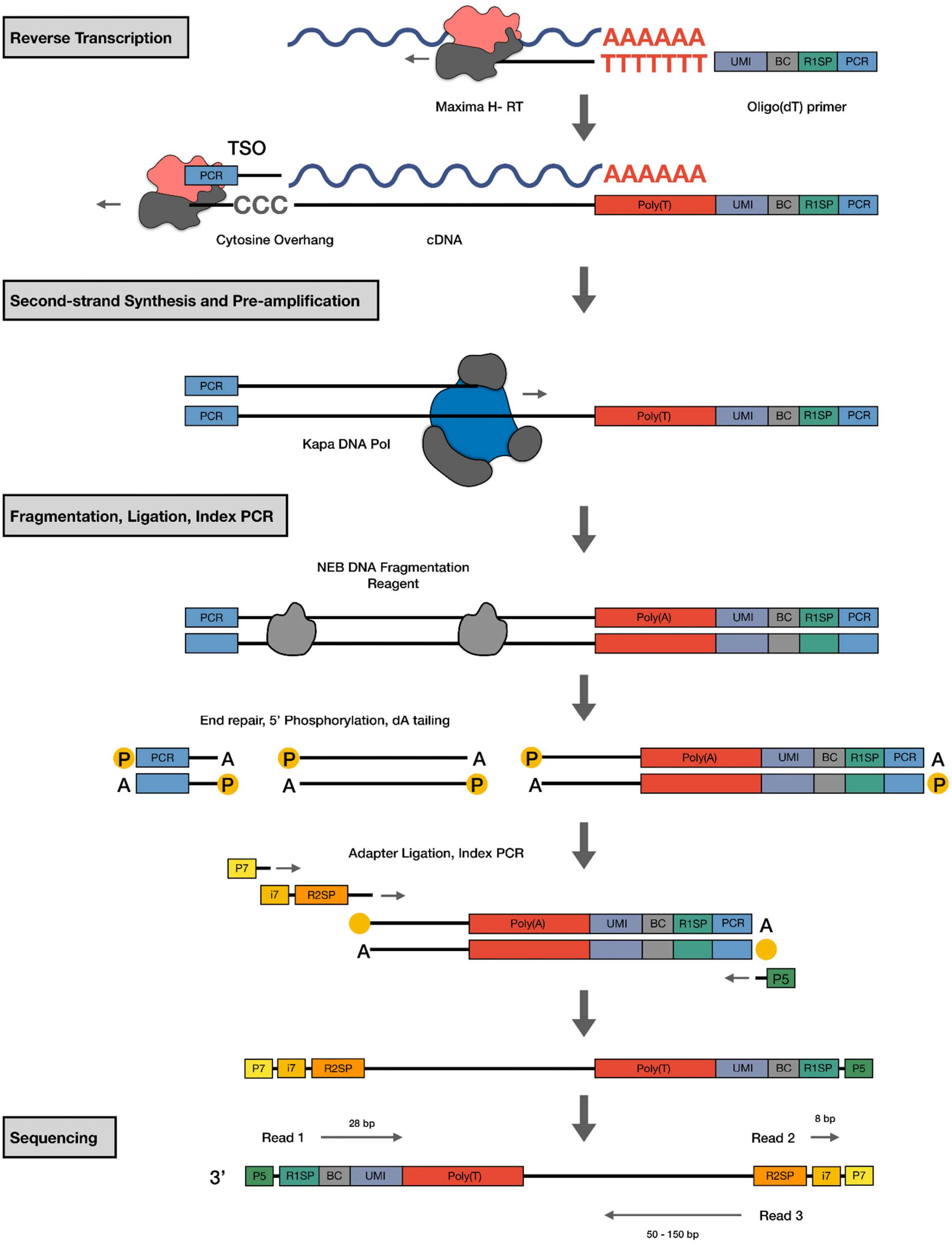

Reverse Transcription
AAAAAA
TTTTTTT
UMI
BC
R1SP
PCR
Maxima H- RT
Oligo(dT) primer
TSO
PCR
AAAAAA
CCC
Poly(T)
UMI
BC
R1SP
PCR
Cytosine Overhang
cDNA
Second-strand Synthesis and Pre-amplification
PCR
PCR
Poly(T)
UMI
BC
R1SP
PCR
Kapa DNA Pol
Fragmentation, Ligation, Index PCR
NEB DNA Fragmentation Reagent
PCR
Poly(A)
UMI
BC
R1SP
PCR
End repair, 5' Phosphorylation, dA tailing
P
PCR
A
A
P
P
A
A
P
P
Poly(A)
UMI
BC
R1SP
PCR
A
A
P
Adapter Ligation, Index PCR
P7
i7
R2SP
Poly(A)
UMI
BC
R1SP
PCR
A
A
P5
P7
i7
R2SP
Poly(T)
UMI
BC
R1SP
P5
Sequencing
Read 1
28 bp
Read 2
8 bp
3'
P5
R1SP
BC
UMI
Poly(T)
R2SP
i7
P7
Read 3
50 - 150 bp

Fig. S1. Molecular workflow of prime-seq. (Related to Figure 1) oligo(dT)-primers are used to enrich mRNA, which is then reverse transcribed using Maxima H-, a M-MLV reverse transcriptase. Full length first strand synthesis is performed using a template switching oligo. Second strand synthesis and cDNA pre-amplification is completed during the PCR using KAPA Hifi Polymerase, and this DNA is then used to generate libraries using the NEBNEXT Ultra II FS Kit. Finally the libraries are sequenced with the following setup: read 1: 28bp, read 2: 8bp, and read 3: 50-150bp.

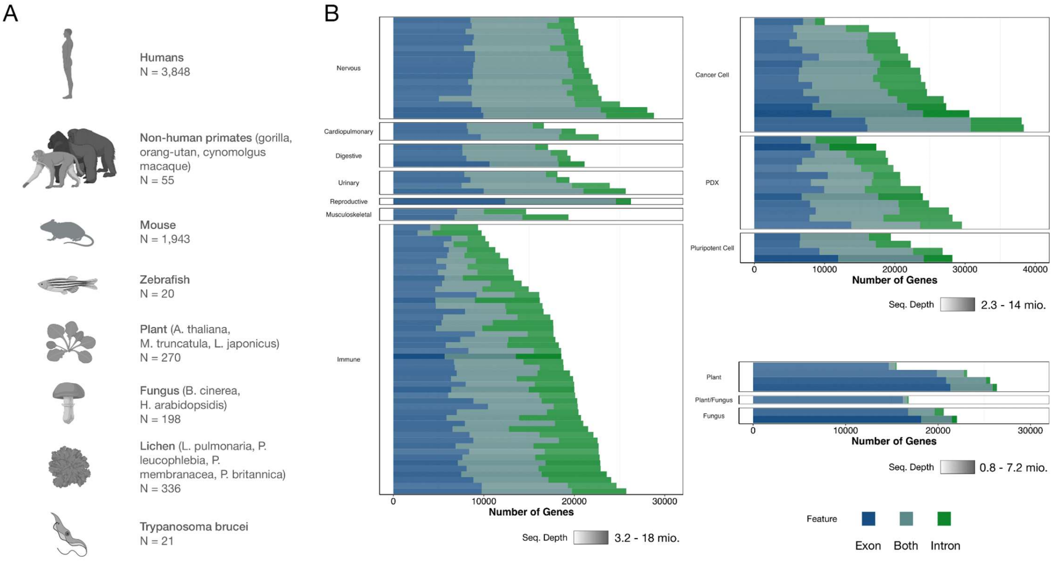

A
Humans
N = 3,848
Non-human primates (gorilla, orang-utan, cynomolgus macaque)
N = 55
Mouse
N = 1,943
Zebrafish
N = 20
Plant (A. thaliana, M. truncatula, L. japonicus)
N = 270
Fungus (B. cinerea, H. arabidopsidis)
N = 198
Lichen (L. pulmonaria, P. leucophlebia, P. membranacea, P. britannica)
N = 336
Trypanosoma brucei
N = 21
B
Nervous
Cardiopulmonary
Digestive
Urinary
Reproductive
Musculoskeletal
Immune
0
10000
20000
30000
Number of Genes
Seq. Depth
3.2 - 18 mio.
Cancer Cell
PDX
Pluripotent Cell
0
10000
20000
30000
40000
Number of Genes
Seq. Depth
2.3 - 14 mio.
Plant
Plant/Fungus
Fungus
0
10000
20000
30000
Number of Genes
Seq. Depth
0.8 - 7.2 mio.
Feature
Exon
Both
Intron

Fig. S2. prime-seq is a robust protocol and has been validated with numerous organisms. (Related to Figure 2A) (A) To date, 132 experiments consisting of 6,691 samples from 17 different organisms, ranging from arabidopsis to zebrafish, have been processed with prime-seq. (B) Data from experiments with well-annotated genomes suggests a substantial number of detected genes come from intronic reads.

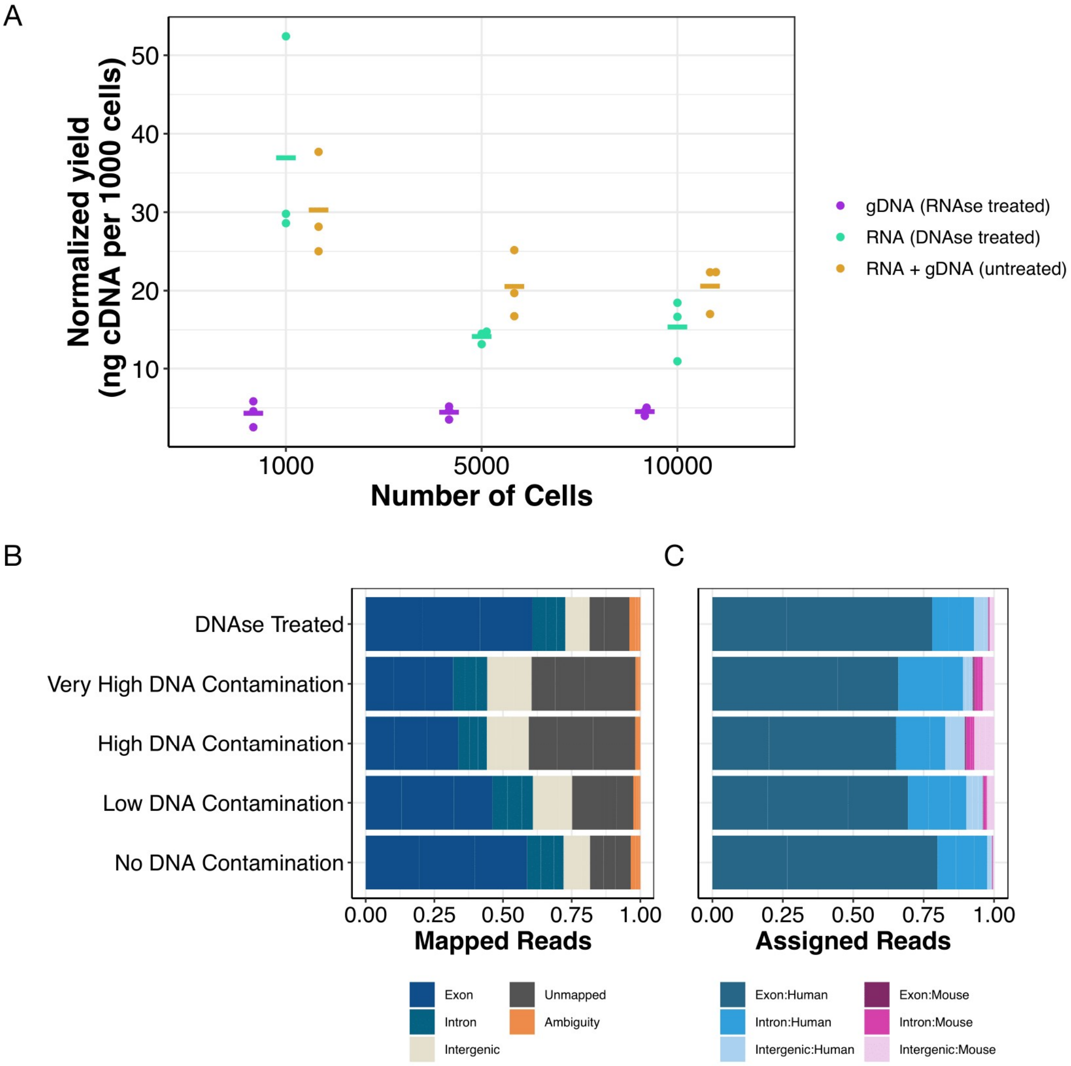

A
Normalized yield
(ng cDNA per 1000 cells)
50
40
30
20
10
1000
5000
10000
Number of Cells
gDNA (RNAse treated)
RNA (DNAse treated)
RNA + gDNA (untreated)
B
DNAse Treated
Very High DNA Contamination
High DNA Contamination
Low DNA Contamination
No DNA Contamination
0.00
0.25
0.50
0.75
1.00
Mapped Reads
Exon
Intron
Intergenic
Unmapped
Ambiguity
C
0.00
0.25
0.50
0.75
1.00
Assigned Reads
Exon:Human
Intron:Human
Intergenic:Human
Exon:Mouse
Intron:Mouse
Intergenic:Mouse

Fig. S3. Intronic reads are not derived from contaminating gDNA. (A) Samples containing total nucleic acids were either treated with RNase A or DNase I, or remained untreated. Untreated samples had the highest concentration, showing that genomic DNA is also used as a template when not removed, albeit less efficiently than mRNA. cDNA yields were normalized to the number of input cells. (Related to Figure 2B) (B) Mapped reads from different gDNA/RNA mixed conditions, showing that the DNase treated condition and the no DNA contamination condition had the lowest fraction of intergenic and unmapped reads. (C) Fraction of assigned mapped reads per genomic feature (exon, intron, intergenic) and species, showing an increase in mouse reads with higher gDNA contamination.

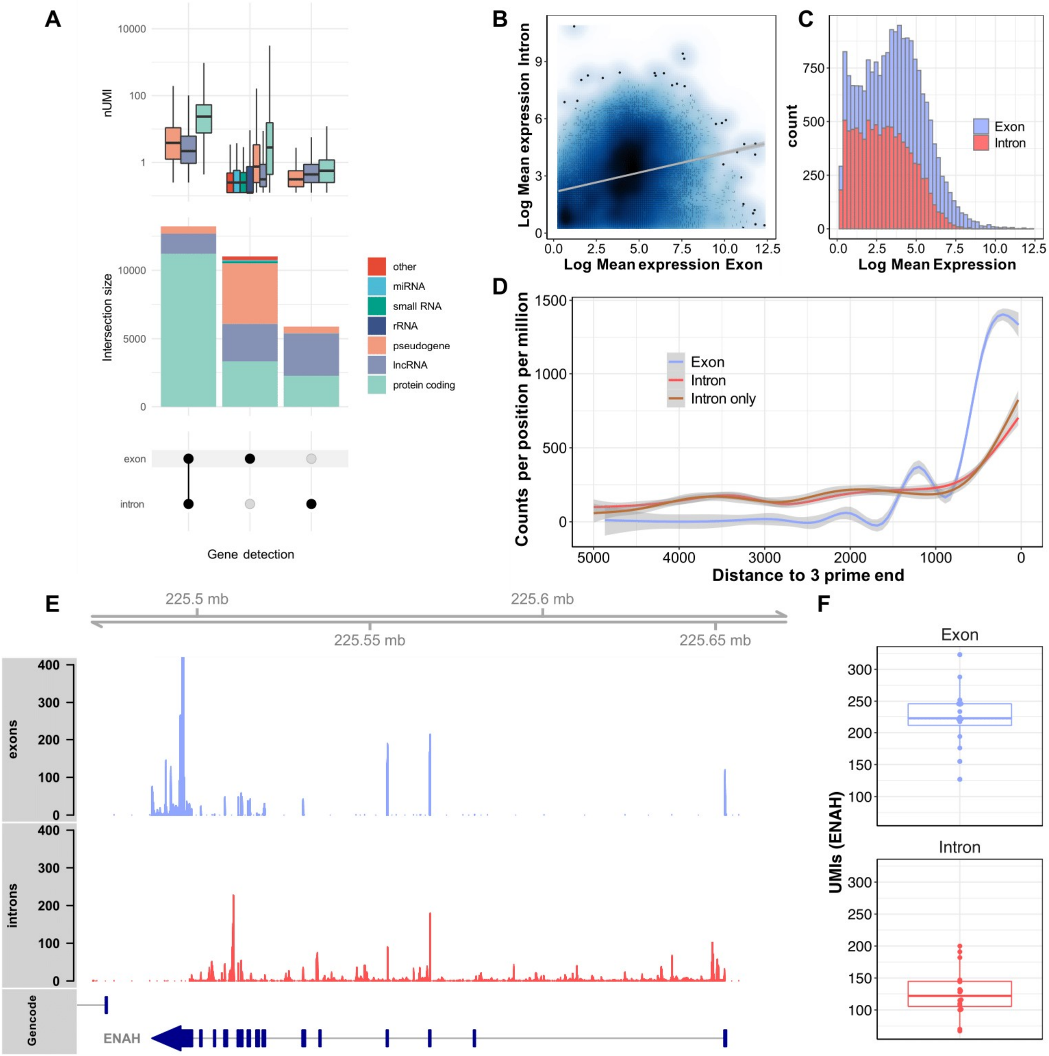

A
10000
100
1
nUMI
10000
5000
0
Intersection size
other
miRNA
small RNA
rRNA
pseudogene
lncRNA
protein coding
exon
intron
Gene detection
B
9
6
3
0
Log Mean expression Intron
0.0
2.5
5.0
7.5
10.0
12.5
Log Mean expression Exon
C
750
500
250
0
count
Exon
Intron
0.0
2.5
5.0
7.5
10.0
12.5
Log Mean Expression
D
1500
1000
500
0
Counts per position per million
Exon
Intron
Intron only
5000
4000
3000
2000
1000
0
Distance to 3 prime end
E
225.5 mb
225.6 mb
225.55 mb
225.65 mb
exons
400
300
200
100
0
introns
400
300
200
100
0
Gencode
ENAH
F
Exon
300
250
200
150
100
Intron
300
250
200
150
100
UMIs (ENAH)

Fig. S4. Intron counts are enriched at the 3' prime end and correlate with exon counts. (A) Upset plot showing the intersection of genes detected with reads mapped to exons, introns or both. Most genes are detected in both introns and exons, followed by exons and introns only. Color represents the biotypes of the detected genes. Genes detected in both introns and exons are enriched for protein coding genes. Boxplots above show the expression levels of the genes by biotype. Genes detected with both intron and exon mapped reads are most highly expressed, intron only detected genes are lowly expressed. (B) Mean expression based on exon counts shows weak correlation to intron counts. (C) Histograms of expression levels of exon counts and intron counts normalized to total counts (intron plus exon) show higher average expression for exon counts. (D) 3’ prime enrichment of exon counts, intron counts and intron only counts. Counts per position relative to the 3’ prime per million averaged over 2000 genes with highest overall expression. Exon and intron counts are enriched at the 3’ prime end of the genbody. Intron only counts follow the same pattern as intron counts in genes with exon counts. (E) Exemplary exon and intron coverage for the gene ENAH show mapping of the intron counts coincides with mapping of exon counts along the gene body. (F) Corresponding UMI counts of ENAH based on intron and exon counting.

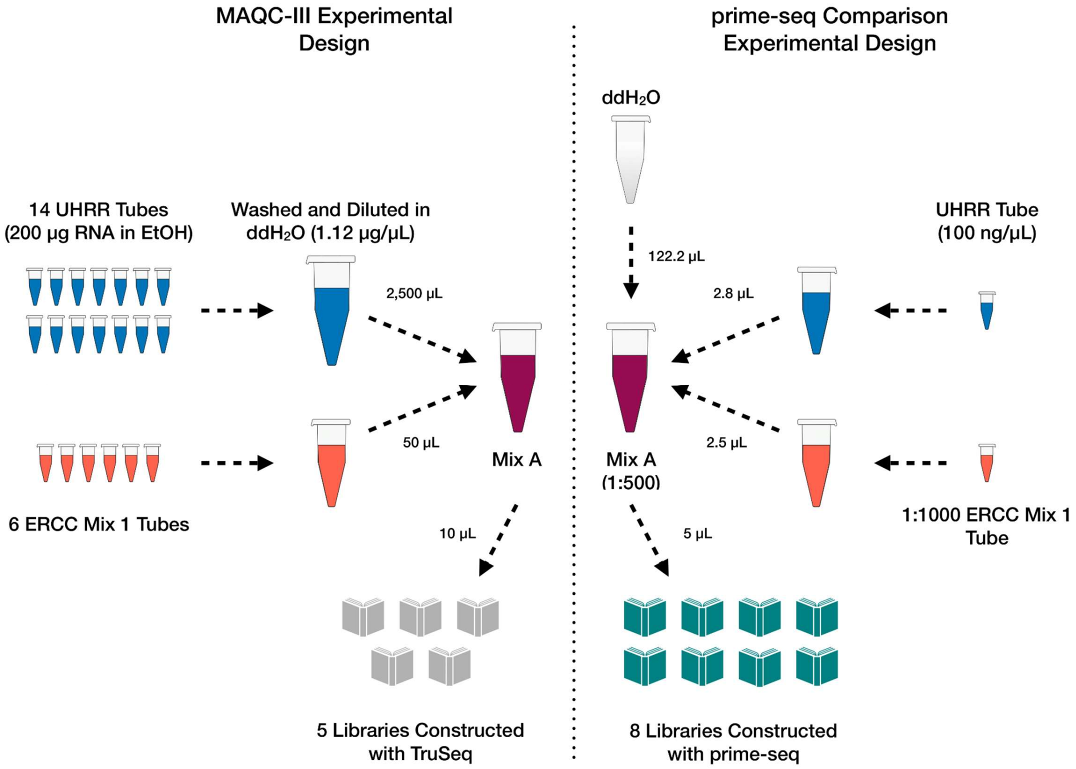

MAQC-III Experimental Design
prime-seq Comparison Experimental Design
ddH₂O
14 UHRR Tubes (200 µg RNA in EtOH)
Washed and Diluted in ddH₂O (1.12 µg/µL)
UHRR Tube (100 ng/µL)
122.2 µL
2,500 µL
2.8 µL
50 µL
2.5 µL
Mix A
Mix A (1:500)
6 ERCC Mix 1 Tubes
1:1000 ERCC Mix 1 Tube
10 µL
5 µL
5 Libraries Constructed with TruSeq
8 Libraries Constructed with prime-seq

Fig. S5. Experimental design comparing prime-seq to TruSeq data generated in the MAQC-III Study. (Related to Figure 3) A 1:1000 concentration of Mix A, from the MAQC-III Study, was generated by mixing UHRR and ERCC Mix 1. From this, eight libraries were generated using prime-seq and compared to five TruSeq generated libraries.

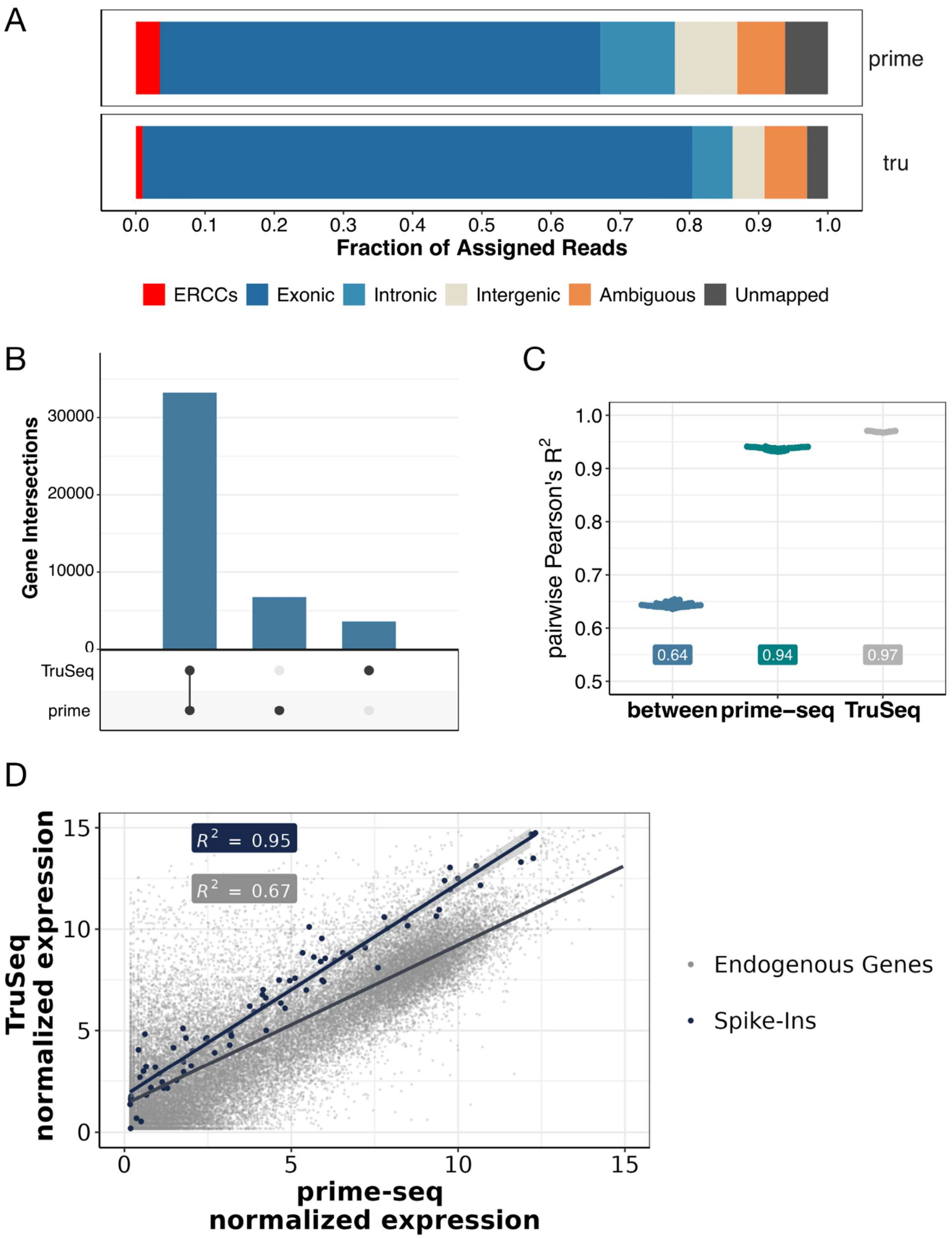

A
prime
tru
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Fraction of Assigned Reads
ERCCs
Exonic
Intronic
Intergenic
Ambiguous
Unmapped
B
Gene Intersections
30000
20000
10000
0
TruSeq
prime
C
pairwise Pearson's R²
1.0
0.9
0.8
0.7
0.6
0.5
0.64
0.94
0.97
between
prime−seq
TruSeq
D
TruSeq
normalized expression
15
10
5
0
R² = 0.95
R² = 0.67
prime-seq
normalized expression
Endogenous Genes
Spike-Ins

Fig. S6. prime-seq and TruSeq have similar mapping, gene detection, and expression. (Related to Figure 3) (A) Feature distribution from prime-seq and TruSeq shows 78% and 85% of reads are exonic, intronic, and ERCCs, respectively. (B) TruSeq and prime-seq exhibit a strong overlap of detected genes (33,230), with 3,589 and 6,766 genes expressed only in TruSeq and prime-seq, respectively. (C) Coefficient of determination of two samples, either between ($R^2 = 0.64$) or within methods ($R^2 = 0.94$ for prime-seq and 0.97 for TruSeq). (D) Gene-wise scatterplot of prime-seq and TruSeq mean normalized expression showing decent correlation of endogenous genes ($R^2 = 0.67$) and strong correlation of ERCC spike-in molecules ($R^2 = 0.95$).

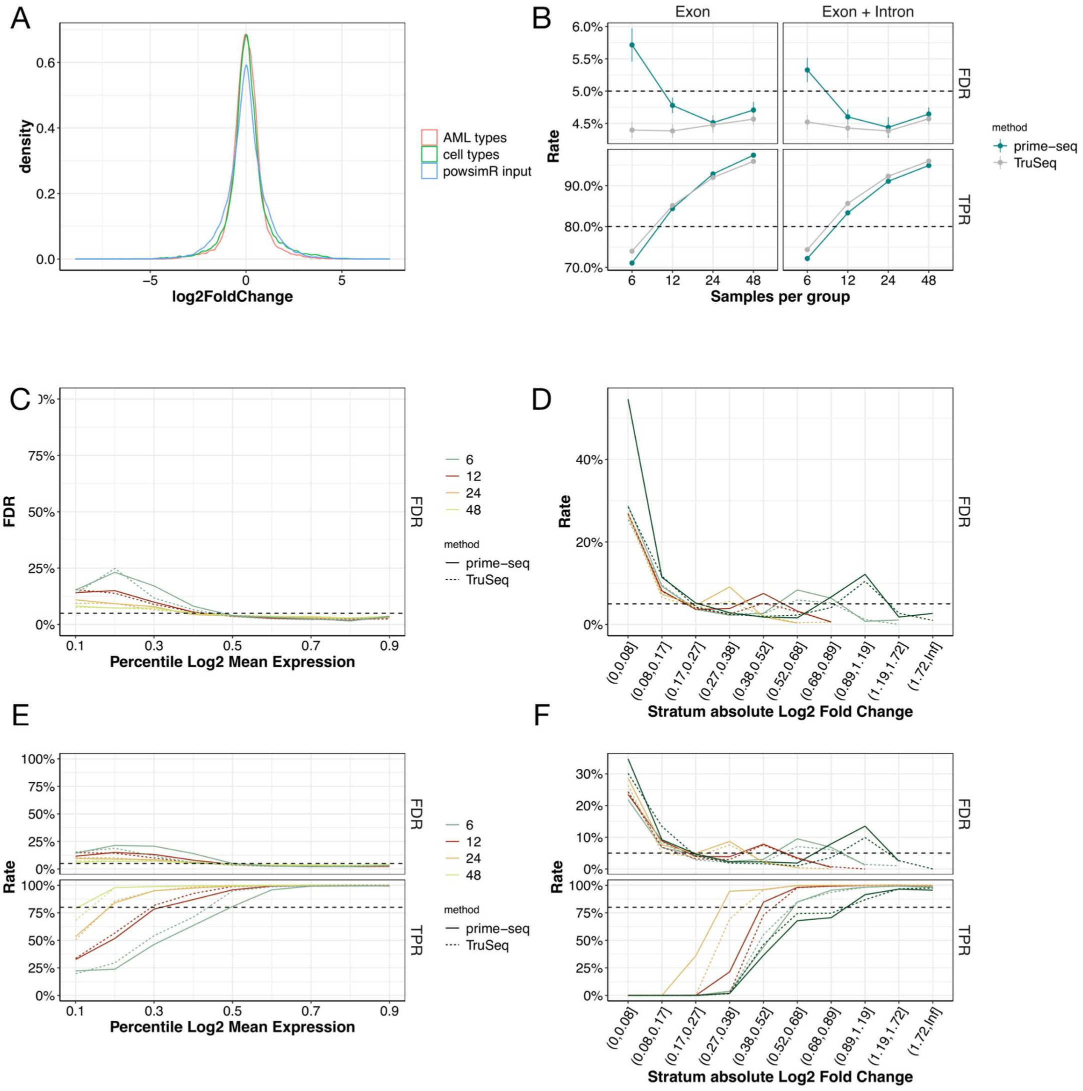

A
density
0.6
0.4
0.2
0.0
−5
0
5
log2FoldChange
AML types
cell types
powsimR input
B
Exon
Exon + Intron
6.0%
5.5%
5.0%
4.5%
FDR
Rate
90.0%
80.0%
70.0%
TPR
6
12
24
48
Samples per group
method
prime−seq
TruSeq
C
75%
50%
25%
0%
FDR
0.1
0.3
0.5
0.7
0.9
Percentile Log2 Mean Expression
6
12
24
48
method
prime−seq
TruSeq
D
40%
20%
0%
Rate
FDR
(0,0.08]
(0.08,0.17]
(0.17,0.27]
(0.27,0.38]
(0.38,0.52]
(0.52,0.68]
(0.68,0.89]
(0.89,1.19]
(1.19,1.72]
(1.72,Inf]
Stratum absolute Log2 Fold Change
E
100%
75%
50%
25%
0%
Rate
FDR
TPR
0.1
0.3
0.5
0.7
0.9
Percentile Log2 Mean Expression
6
12
24
48
method
prime−seq
TruSeq
F
30%
20%
10%
0%
100%
75%
50%
25%
0%
Rate
FDR
TPR
(0,0.08]
(0.08,0.17]
(0.17,0.27]
(0.27,0.38]
(0.38,0.52]
(0.52,0.68]
(0.68,0.89]
(0.89,1.19]
(1.19,1.72]
(1.72,Inf]
Stratum absolute Log2 Fold Change

Fig. S7. Power and FDR mostly depend on sample size and are similar between prime-seq and TruSeq. (Related to Figure 3) (A) Log2 fold change distribution from the AML and NPC differentiation experiment (Figure 4) compared to the log2 fold change distribution used in powsimR for power analysis confirms that simulation settings match expected distributions. (B) Marginal power of prime-seq and TruSeq at differing samples per condition shows both methods perform similarly well, crossing the 80% threshold with roughly 12 samples both for exon plus intron and only exon counts. (C and D) FDR over different mean expression and log2 fold change strata (Related to 3F and 3G). (E and F) analogous to Figure 3F and 3G but including only Exonic counts; prime-seq and TruSeq exhibit similar TPR and FDR over different mean expression and log2 fold change strata. Filtering parameters: detected UMI ≥ 1, detected gene present in at least 25%.

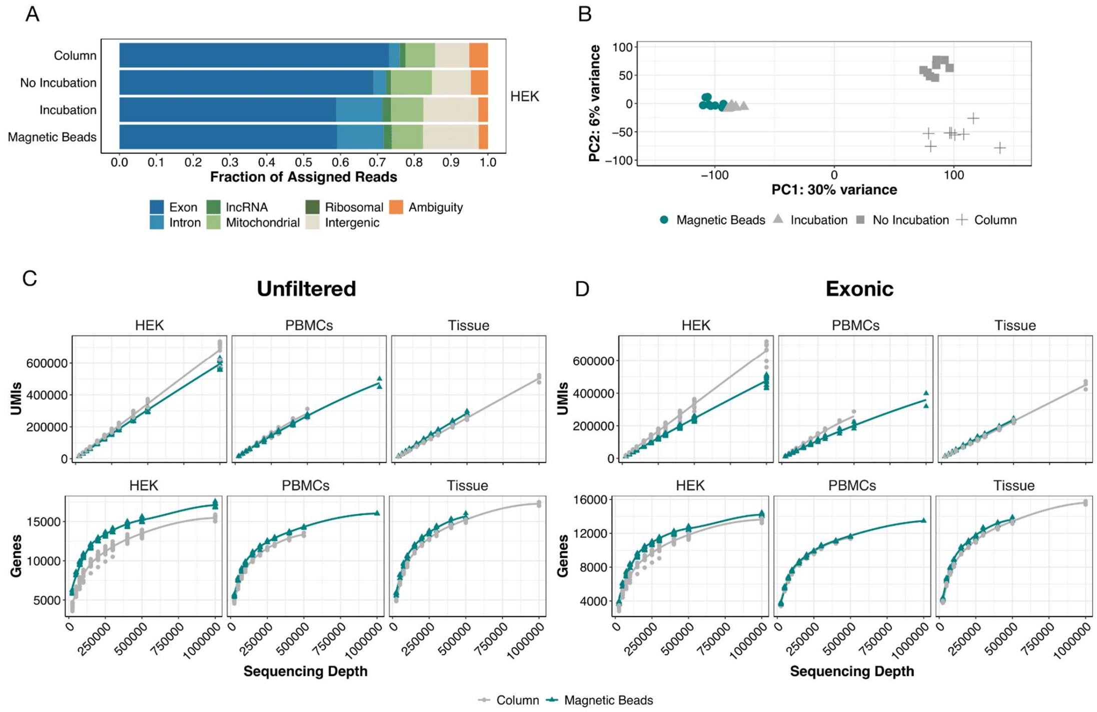


Fig. S8. Performance of isolation methods is similar independent of prefiltering or usage of only Exon data. (Related to Figure 4) (A) HEK293T cell samples were extracted using columns and magnetic beads, employing the

standard prime-seq protocol (“Magnetic Beads”), as well as variant protocols without proteinase K digestion (“No Incubation”) and a proteinase K digestion control without enzyme (“Incubation”). All conditions had similar fractions of usable reads (all but intergenic and ambiguity), with an increase in intronic reads in “Incubation” and “Magnetic Beads” suggesting this increase is due to heat incubation. (B) Principal component analysis (PCA) of the 500 most variable genes shows the largest variable is heat incubation. (C and D) Analysis of detected UMIs and detected genes for unfiltered data and exonic only data shows that prime-seq using magnetic bead isolation is more sensitive in HEK cells and similarly sensitive in PBMCs and tissue compared to prime-seq using column isolation. Filtering parameters: detected UMI ≥ 1, detected gene present in at least 25% of samples and is protein coding.

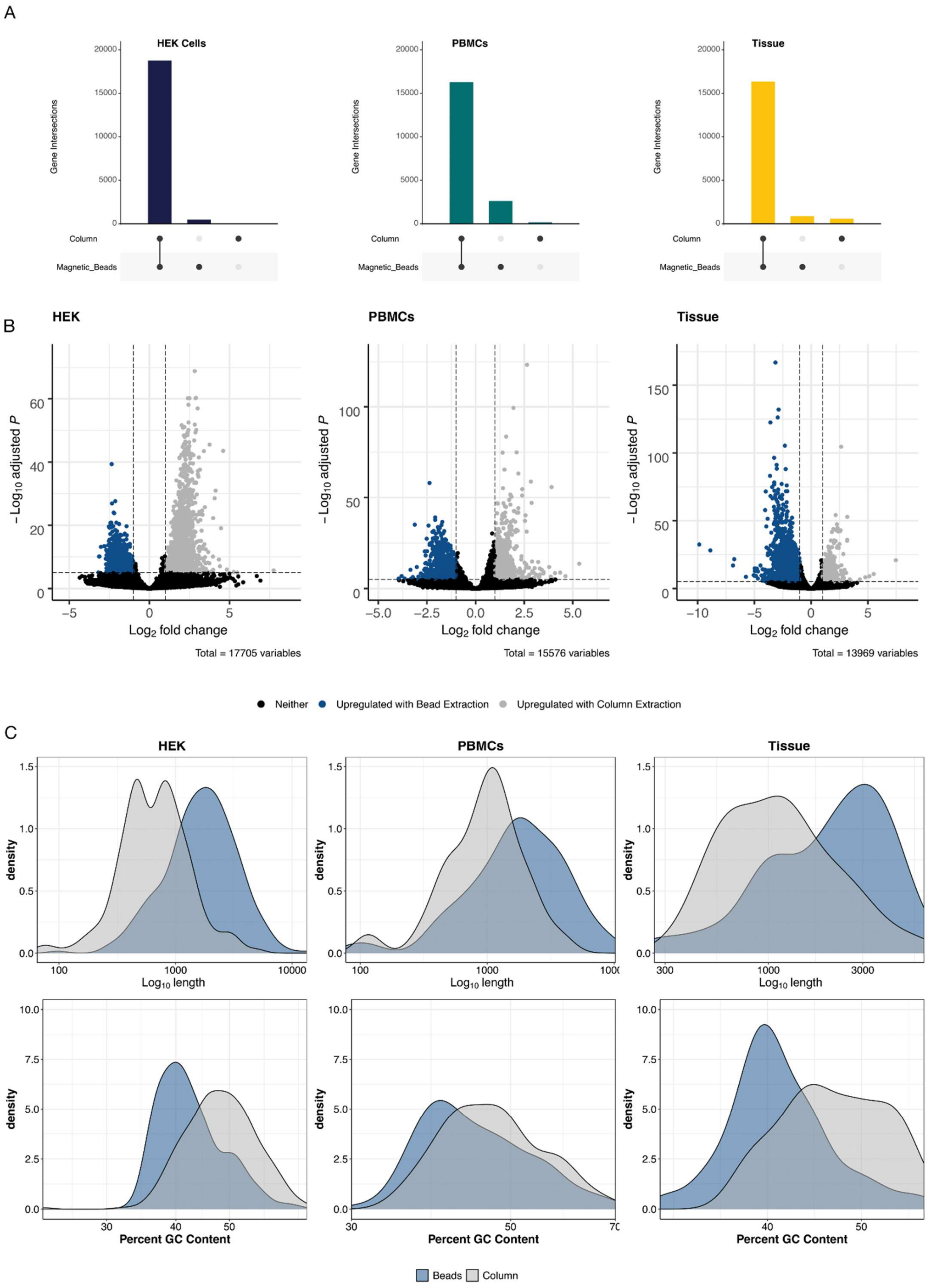

A
HEK Cells
PBMCs
Tissue
Gene Intersections
20000
15000
10000
5000
0
Column
Magnetic_Beads
B
HEK
PBMCs
Tissue
−Log10 adjusted P
Log2 fold change
Total = 17705 variables
Total = 15576 variables
Total = 13969 variables
● Neither ● Upregulated with Bead Extraction ● Upregulated with Column Extraction
C
HEK
PBMCs
Tissue
density
Log10 length
Percent GC Content
Beads
Column

Fig. S9. Most genes are detected independent of the extraction method used. (Related to Figure 4) (A) Upset plots showing a strong overlap of detected genes between columns and magnetic beads. (B) Up- and down-regulated genes between column and bead-based RNA extractions ($p>0.05$, $\log_2$ FC > 2). (C) Density plots of the differentially expressed genes relative to length and GC content. Genes upregulated in columns tend to be longer with lower GC content.

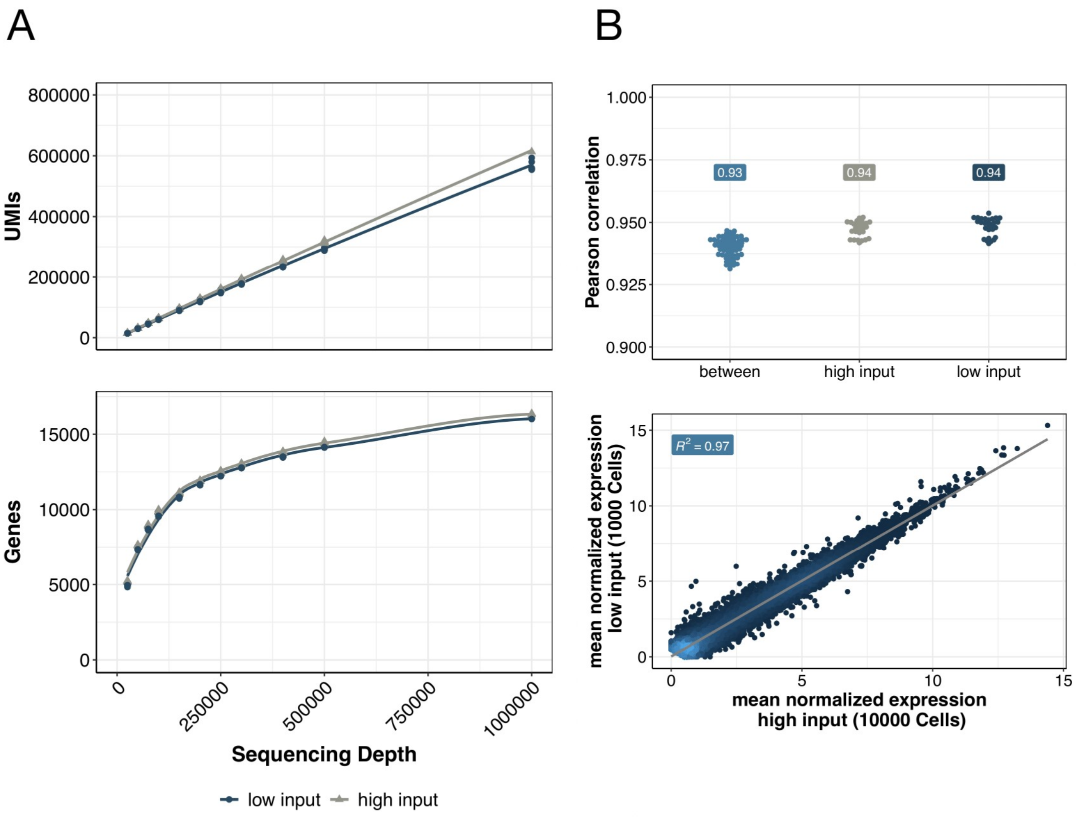

A
B
UMIs
800000
600000
400000
200000
0
Genes
15000
10000
5000
0
0
250000
500000
750000
1000000
Sequencing Depth
low input
high input
Pearson correlation
1.000
0.975
0.950
0.925
0.900
0.93
0.94
0.94
between
high input
low input
$R^2 = 0.97$
mean normalized expression
low input (1000 Cells)
15
10
5
0
mean normalized expression
high input (10000 Cells)

Fig. S10. prime-seq performs equally well with high- and low-input samples. (Related to Figure 5) (A) Sensitivity, measured in detected UMIs and genes, is similar between high input (10,000 HEK293T cells) and low input (1,000 HEK293T cells) conditions at various sequencing depths (filtering parameters: detected UMI ≥ 1, detected gene present in at least 25% of samples and is protein coding). (B) Additionally, Pearson's correlations between the high- and low-input conditions were high (pairwise comparison between: $r = 0.93$, pairwise comparison within: $r = 0.94$, and average normalized mean expression, $R^2 = 0.97$).

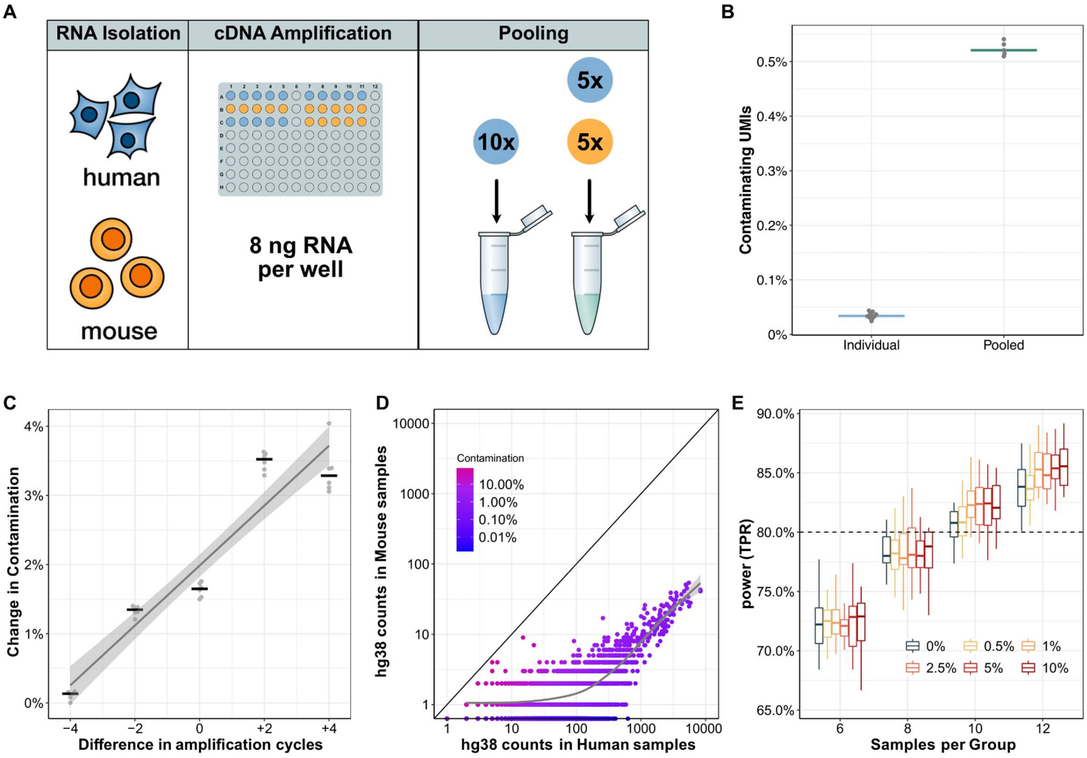

A
RNA Isolation
cDNA Amplification
Pooling
human
mouse
8 ng RNA per well
10x
5x
5x
B
Contaminating UMIs
0%
0.1%
0.2%
0.3%
0.4%
0.5%
Individual
Pooled
C
Change in Contamination
0%
1%
2%
3%
4%
−4
−2
0
+2
+4
Difference in amplification cycles
D
hg38 counts in Mouse samples
hg38 counts in Human samples
1
10
100
1000
10000
Contamination
10.00%
1.00%
0.10%
0.01%
E
power (TPR)
65.0%
70.0%
75.0%
80.0%
85.0%
90.0%
6
8
10
12
Samples per Group
0%
0.5%
1%
2.5%
5%
10%

Fig. S11. Cross-contamination levels are low, increase with additional cycles but do not impact power simulations. (A) Experimental overview to detect cross-contamination. 1.RNA was isolated from hiPSCs and mESC; 2. cDNA amplification of 8ng RNA per well; 3. pooling of only human samples or mouse and human samples. (B) The percentage of contaminating UMIs (mapping best to the mouse genome) increases with pooling but is generally low median early pooling: 0.52%. (C) Impact of amplification cycles on cross-contamination. 0 corresponds to the condition shown in panel B, 13 cycles of pre-amplification for 96 ng of input RNA (8 ng per well). (D) Genewise contamination ranges from 0% to up to 10 % for lowly expressed genes. Contamination decreases with increasing expression levels. (E) Power simulation with different levels of computationally added contamination shows little impact on marginal TPR. An increase in the number of replicates leads to a small increase in power for highly contaminated conditions relative to no contamination.

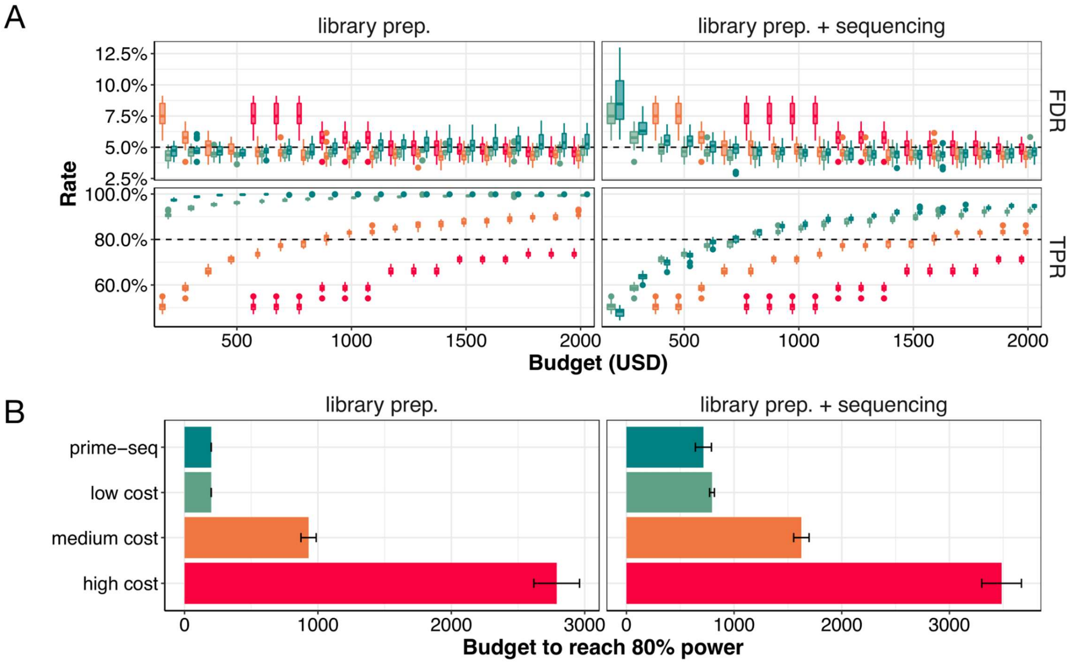

A
library prep.
library prep. + sequencing
Rate
12.5%
10.0%
7.5%
5.0%
2.5%
100.0%
80.0%
60.0%
FDR
TPR
500
1000
1500
2000
Budget (USD)
B
library prep.
library prep. + sequencing
prime−seq
low cost
medium cost
high cost
0
1000
2000
3000
Budget to reach 80% power

Fig. S12. Power analysis shows prime-seq is able to reach 80% power earlier than less cost-efficient methods. (Related to Figure 6) (A) True positive rate (TPR) and false discovery rates (FDR) corresponding to Figure 6B, but with more incremental values. (B) prime-seq crosses an 80% power threshold with $715 when sequencing costs are included compared to $795, $1,625, and $3,485 for low, middle, and high cost methods respectively (10 million reads used for analysis at a cost of $3.40 per 1 mio. reads).

Date	Species	Sample_Type	Sample_Subtype	Sample_Number	Depth	Average_reads_p	Genes_Exon	Genes_Intron	Genes_Inex	Genes_Exon_Onl	Genes_Intron_Or	Genes_Both	Frac_Genes_Exo	Frac_Genes_Intr	Frac_Genes_Bot	UMI_Exonic	UMI_Intronic	UMI_Inex	frac_inex_umi
20170216	human	Cell Line	iPSC	14															
20170613	primate (h3, c7)	Cell Line	iPSC	10															
20170713	human	Cell Line	AML	90	667264509	7414050	30780	22284	38057	15773	7277	15007	0,4	0,2	0,39	2655186	365921	2997550	0,11
20170728	mouse	Nervous	Striatum	31	168160761	5424541	19968	17953	23006	5053	3038	14915	0,2	0,1	0,65	1484727	1303746	2808176	0,47
20180502	mouse	Musculoskeletal	HSC	20	154988721	7749436	14256	12579	19353	6774	5097	7482	0,4	0,3	0,39	278785	185268	465648	0,40
20180514	mouse	Nervous	Hippocampus/Co	90	607862036	6754023	19765	13938	22596	8658	2831	11107	0,4	0,1	0,49	2125493	260766	2378718	0,11
20180614	mouse	Immune	CD4 T Cell	11	89435464	8130497	12583	12488	18286	5798	5703	6785	0,3	0,3	0,37	181556	132982	312245	0,42
20180619	human	Cell Line	AML	90	667264509	7410000	30780	22284	38357	16074	7577	14707	0,4	0,2	0,38	2655186	365921	2997550	0,11
20180619	human	Cell Line	AML	99	1192954561	12100000	21743	19047	27306	8259	5563	13484	0,3	0,2	0,49	3304665	1191187	4431165	0,25
20180619	human	Immune	NA	83	471649567	5680000	13724	11847	17347	5500	3623	8224	0,3	0,2	0,47	344655	155923	500578	0,31
20181004	mouse	Cell Line	Breast/Lung Tum	24	159697008	6654042	16109	13942	20782	6840	4673	9269	0,3	0,2	0,45	2131298	444135	2565736	0,17
20181004	primates (h14, o6	Renal/Urinary	UDSC/iPSC	24	84623082	3525962	19734	16367	23918	7551	4184	12183	0,3	0,2	0,51	1665334	388900	2110512	0,21
20190103	mouse	Immune	CD4 T Cell	81	855843229	10565966	16819	15731	22874	7143	6055	9676	0,3	0,3	0,42	850164	428217	1293394	0,34
20190205	primates (g12, h1	Cell Line	iPSC/NPC	53	340762162	6429475	17290	15821	22280	6459	4990	10831	0,3	0,2	0,49	429644	292376	696806	0,38
20190508	mouse	Immune	Macrophage	24	200076477	8336520	15914	12495	19512	7017	3598	8897	0,4	0,2	0,46	3875014	268950	4147086	0,07
20190508	mouse	Immune	T Effector/Monoc	39	281646648	7221709	14372	12068	18776	6708	4404	7664	0,4	0,2	0,41	415141	155510	566621	0,27
20190612	human	Immune	BMMNC	47	178517479	3798244	16136	14029	20092	6063	3956	10073	0,3	0,2	0,50	844310	209392	1009098	0,16
20190819	mouse	Nervous	Hippocampus/Co	120	497208266	4143402	16986	11396	20046	8650	3060	8336	0,4	0,2	0,42	686832	113622	793238	0,13
20190819	mouse	Nervous	Microglia	70	424723001	6067471	19174	14074	22704	8630	3530	10544	0,4	0,2	0,46	1308820	216483	1492922	0,12
20191118	mouse	Immune	B Cell	12	134183422	11181952	14932	13595	20002	6407	5070	8525	0,3	0,3	0,43	400606	216112	610824	0,34
20191118	L. pulmonaria	Lichen	Lichen	96															
20191118	zebrafish	Immune	Macrophage	20	152308641	7615432	16972	13442	20462	7020	3490	9952	0,3	0,2	0,49	1672986	161080	1848200	0,09
20191205	mouse	Immune	CD4 T Cell	9	72386086	8042898	17858	13079	20989	7910	3131	9948	0,4	0,1	0,47	2704451	240463	2912509	0,07
20191206	L. pulmonaria	Lichen	Lichen	96															
20191227	human	PDX	ALL	34	145693500	4290000	18361	20345	28214	7869	9853	10492	0,3	0,3	0,37	383630	406278	813672	0,53
20191227	human	Immune	AML	20	178814900	8940000	9816	12977	17695	4718	7879	5098	0,3	0,4	0,29	99198	84539	197190	0,50
20191227	human	Immune	AML	32	289590651	9050000	13900	14859	20750	5891	6850	8009	0,3	0,3	0,39	210185	153093	358296	0,41
20191227	human	PDX	AML	36	368006698	10200000	17674	17263	23989	6726	6315	10949	0,3	0,3	0,46	656509	406156	1035629	0,37
20191227	human	Immune	FL	27	177007598	6560000	9113	7951	12712	4761	3599	4352	0,4	0,3	0,34	67388	42333	108740	0,38
20191227	human	Immune	FL	31	237700974	7670000	4310	7085	9753	2668	5443	1642	0,3	0,6	0,17	48140	40191	91913	0,48
20191227	Lotus Japonicus	Plant	NA	64	527421623	8240000	26004	NA	NA	NA	NA	NA	NA	NA	NA	1402552	NA	NA	NA
20191227	Arabidopsis/Botry	Plant/Fungus	Root	24	20589676	858000	16658	616	16831	16215	173	443	1,0	0,0	0,03	378177	3060	381727	0,01
20191227	Medicago Trunca	Plant	Root	36	43527940	1210000	22775	3281	23150	19869	376	2906	0,9	0,0	0,13	646296	17926	659930	0,02
20191227	Medicago Trunca	Plant	Root	60	251655144	4190000	25198	4776	25647	20872	449	4327	0,8	0,0	0,17	1014887	42333	1052161	0,04
20200115	Medicago Trunca	Plant	Root	36	187608414	5210000	25883	5030	26373	21344	491	4539	0,8	0,0	0,17	717637	72114	773685	0,07
20200615	human	PDX	ALL	643	1534379380	2390000	15202	13757	19792	6035	4590	9167	0,3	0,2	0,46	542453	220402	780316	0,30
20200615	mouse	Cell Line	AML	16	53078855	3320000	13004	10760	16299	5539	3295	7465	0,3	0,2	0,46	564427	138393	716526	0,21
20200615	human	PDX	AML	96	380580493	3960000	13060	10090	18703	8613	5644	4447	0,5	0,3	0,24	72438	38432	108412	0,33
20200615	human	PDX	AML	65	321067720	4940000	20668	18948	27714	8766	7046	11902	0,3	0,3	0,43	1243211	314865	1535515	0,19
20200615	human	Cell Line	AML	72	367524312	5100000	16915	17421	23727	6306	6812	10609	0,3	0,3	0,45	536213	226825	765653	0,30
20200615	human	PDX	AML	61	364845827	5980000	14234	12203	19029	6826	4795	7408	0,4	0,3	0,39	384612	95537	458911	0,16
20200615	human	PDX	AML	24	147072360	6130000	20503	16952	24871	7920	4368	12584	0,3	0,2	0,51	2432474	706012	3131424	0,22
20200615	human	Cell Line	AML	30	193649969	6450000	15951	15417	20396	4980	4446	10971	0,2	0,2	0,54	2460357	485167	2931711	0,16
20200615	human	Cell Line	AML	48	311114082	6480000	17495	17237	23616	6379	6121	11116	0,3	0,3	0,47	1552399	387650	1922947	0,19
20200615	human	Cell Line	AML	24	175231123	7300000	19017	17602	24573	6972	5556	12046	0,3	0,2	0,49	2425468	522129	2904592	0,16
20200615	human	Cell Line	AML	37	302458592	8170000	17899	15438	22229	6791	4330	11108	0,3	0,2	0,50	2275883	392365	2664432	0,15
20200615	human	Immune	Macrophage	189	834993304	4420000	9312	8286	13209	4923	3897	4389	0,4	0,3	0,33	79047	43686	122081	0,35
20200615	human	Immune	Mast Cell Tumor	47	178517479	3800000	16136	14029	20092	6063	3956	10073	0,3	0,2	0,50	844310	209392	1009098	0,16
20200615	Arabidopsis	Plant	NA	50	41558415	831000	15385	844	15531	14687	146	698	0,9	0,0	0,04	468026	6672	474046	0,01
20200615	Hyaloperonospor	Fungus	NA	96	288531004	3010000	19621	3857	20620	16763	999	2859	0,8	0,0	0,14	599511	28319	628338	0,05
20200615	mouse	NA	NA	18	100324027	5570000	9983	7162	13629	6468	3647	3515	0,5	0,3	0,26	209765	46570	255041	0,18
20200615	Hyaloperonospor	Fungus	NA	72	519351123	7210000	21510	3860	22046	18186	536	3324	0,8	0,0	0,15	1334577	27015	1364579	0,02
20200615	mouse	Immune	T Cell	25	121011147	4840000	12351	11785	16486	4701	4135	7650	0,3	0,3	0,46	656450	358475	1011342	0,35
20200615	mouse	Immune	T Cell	16	87810387	5490000	10735	8780	14182	5402	3447	5333	0,4	0,2	0,38	390166	154191	548080	0,29
20200615	mouse	Immune	T Cell	10	181413918	18100000	13541	12886	18566	5681	5026	7860	0,3	0,3	0,42	1167876	413326	1560973	0,25
20200626	human	Immune	FL	30	173289413	5780000	9839	10321	14953	4633	5114	5207	0,3	0,3	0,35	72974	78604	154274	0,53
20200626	P. britannica	Lichen	Lichen	48															
20200713	human	Immune	FL	21	153761335	7320000	19711	15980	25747	9767	6036	9944	0,4	0,2	0,39	2775491	685586	3456828	0,20
20200718	P. leucophlebia	Lichen	Lichen	48															
20200718	P. membranacea	Lichen	Lichen	48															
20200728	human	Cell Line	AML	36	188949807	5250000	17902	16131	24355	8224	6454	9678	0,3	0,3	0,40	1705263	478377	2213546	0,23
20200805	primates (h4, g3	Cell Line	Germ Layer Cells	13	69469374	5343798	22583	17555	26803	9248	4220	13335	0,3	0,2	0,50	3250664	410911	3707673	0,12
20200914	human	Cell Line	AML	11	151550105	13800000	23991	19659	30615	10956	6624	13035	0,4	0,2	0,43	1977816	516988	2507980	0,21
20200914	human	Immune	B Cell	54	315568870	5840000	16154	13694	22112	8419	5959	7735	0,4	0,3	0,35	836132	118243	936227	0,11
20200915	mouse	Immune	CD4 T Cell	15	72791966	4852798	13415	6905	16127	9222	2712	4193	0,6	0,2	0,26	1344806	72338	1409602	0,05
20200915	mouse	Digestive/Excreto	Liver	12	134321077	11193423	18256	10487	21138	10651	2882	7605	0,5	0,1	0,36	3298310	191962	3491292	0,06
20200915	human	Immune	Neutrophil	30	333866108	11128870	9026	11545	16194	4649	7168	4377	0,3	0,4	0,27	119156	131017	247988	0,52
20200917	mouse	Cell Line	AML	24	94184242	3920000	16958	12681	21886	9205	4928	7753	0,4	0,2	0,35	730935	87212	813445	0,10
20201013	mouse	Nervous	Hippocampus/Co	84	826769283	9842491	23505	18874	28804	9930	5299	13575	0,3	0,2	0,47	1566200	681048	2254111	0,31
20201013	mouse	Nervous	Hippocampus/Co	73	738341739	10114270	22910	18336	28069	9733	5159	13177	0,3	0,2	0,47	1347825	647520	2002625	0,33
20201123	human	Cell Line	AML	48	343687027	7160000	20329	17710	26950	9240	6621	11089	0,3	0,2	0,41	2413964	635285	3056800	0,21
20201127	human	Renal/Urinary	UDSC	19	165756189	8724010	21006	15687	25680	9993	4674	11013	0,4	0,2	0,43	2172970	454127	2633742	0,17
20201130	mouse	Immune	CD4 T Cell	26	110175505	4237519	15556	10252	18548	8296	2992	7260	0,4	0,2	0,39	661148	91030	760280	0,13
20201130	mouse	Immune	CD4 T Cell	23	128115947	5570259	15627	11976	19986	8010	4359	7617	0,4	0,2	0,38	901509	167042	1066079	0,15
20201201	human	PDX	AML	75	306621078	4090000	15701	13716	23674	9958	7973	5743	0,4	0,3	0,24	189404	48716	237956	0,20
20201211	mouse	Cell Line	ESC	48	362302997	7550000	16279	12883	19429	6546	3150	9733	0,3	0,2	0,50	1385787	281726	1609035	0,14
20201211	human	NA	NA	44	174750108	3970000	16932	13992	21514	7522	4582	9410	0,3	0,2	0,44	1948246	204655	2139705	0,09
20201217	mouse	NA	NA	20	187679331	9380000	24087	17731	29023	11292	4936	12795	0,4	0,2	0,44	3577873	677420	4257410	0,16

20210103	human	Immune	CD4 T Cell	61	453475416	7434023	13615	13511	19876	6365	6261	7250	0,3	0,3	0,36	179999	141165	324770	0,45
20210103	human	Immune	CD8 T Cell	61	526290640	8627715	15502	15781	22689	6908	7187	8594	0,3	0,3	0,38	173058	186307	353092	0,51
20210103	human	Immune	Monocyte	61	630918344	10342924	16394	16083	23586	7503	7192	8891	0,3	0,3	0,38	336456	248703	562609	0,40
20210103	human	Immune	NK Cell	61	513908493	8424729	10998	12319	17659	5340	6661	5658	0,3	0,4	0,32	72194	73267	144028	0,50
20210118	mouse	Immune	B Cell	40	196257893	4906447	14848	8684	17700	9016	2852	5832	0,5	0,2	0,33	311156	65039	381987	0,19
20210118	mouse	Immune	Macrophage	8	63964151	7995519	16406	11447	20321	8874	3915	7532	0,4	0,2	0,37	680922	130476	806924	0,16
20210118	human	Immune	Monocyte	14	118480281	8462877	12692	10516	17915	7399	5223	5293	0,4	0,3	0,30	219944	76638	296322	0,26
20210118	mouse	Immune	NKT Cell	8	60576522	7572065	12476	8666	16605	7939	4129	4537	0,5	0,2	0,27	156631	60440	216728	0,28
20210118	human	Immune	PBMC	10	32493987	3249399	16716	14304	22702	8398	5986	8318	0,4	0,3	0,37	334012	139677	476715	0,30
20210119	human	Cell Line	AML	36	242069701	6720000	8628	3083	10026	6943	1398	1685	0,7	0,1	0,17	57428	6895	64502	0,11
20210125	human	PDX	AML	36	215428861	5980000	16931	10291	20749	10458	3818	6473	0,5	0,2	0,31	552710	78789	630861	0,12
20210209	mouse	Nervous	Amygdala	24	160583036	6690960	18824	11588	20472	8884	1648	9940	0,4	0,1	0,49	2051065	198900	2239209	0,08
20210209	mouse	Renal/Urinary	Bladder	24	147605071	6150211	16839	10288	18125	7837	1286	9002	0,4	0,1	0,50	1121691	127950	1248498	0,10
20210209	mouse	Nervous	Cerebellum	24	152361734	6348406	18542	12451	20432	7981	1890	10561	0,4	0,1	0,52	1624876	363134	1993624	0,18
20210209	mouse	Digestive/Excreto	Colon	24	149444378	6226849	18119	11518	19584	8066	1465	10053	0,4	0,1	0,51	2324318	186208	2487607	0,07
20210209	mouse	Nervous	Cortex	24	184164613	7673526	19182	11926	20972	9046	1790	10136	0,4	0,1	0,48	2744274	205742	2967972	0,08
20210209	mouse	Cardiopulmonary	Heart	24	165797552	6908231	15392	8513	16648	8135	1256	7257	0,5	0,1	0,44	2991790	72944	3067478	0,02
20210209	mouse	Renal/Urinary	Kidney	24	178812341	7450514	17998	10928	19466	8538	1468	9460	0,4	0,1	0,49	3429545	136196	3565195	0,04
20210209	mouse	Digestive/Excreto	Liver	24	171165193	7131883	15672	9478	17104	7626	1432	8046	0,4	0,1	0,47	3059222	127490	3176702	0,04
20210209	mouse	Cardiopulmonary	Lung	24	177049706	7377071	18580	11888	20140	8252	1560	10328	0,4	0,1	0,51	2297220	235524	2525928	0,09
20210209	mouse	Nervous	Medulla	24	155876551	6494856	19348	11985	20996	9011	1648	10337	0,4	0,1	0,49	2372248	166350	2559347	0,07
20210209	mouse	Nervous	Midbrain	24	201188900	8382871	19928	13068	21853	8785	1925	11143	0,4	0,1	0,51	2524544	362878	2780702	0,09
20210209	mouse	Nervous	Pons	24	162613566	6775565	18304	11439	19970	8531	1666	9773	0,4	0,1	0,49	1854192	149854	2003800	0,07
20210209	mouse	Digestive/Excreto	Small Intestines	24	142122775	5921782	17382	11602	19235	7633	1853	9749	0,4	0,1	0,51	1968784	258936	2204667	0,11
20210209	mouse	Immune	Spleen	24	167265038	6969377	16178	12038	18835	6797	2657	9381	0,4	0,1	0,50	1123878	326388	1424022	0,21
20210209	mouse	Nervous	Striatum	24	165797552	6908231	19266	12258	21116	8858	1850	10408	0,4	0,1	0,49	2471752	255206	2738774	0,10
20210209	mouse	Nervous	Striatum	24	166467850	6936160	18937	11908	20662	8754	1725	10183	0,4	0,1	0,49	2216540	219890	2415078	0,08
20210209	mouse	Reproductive	Testis	13	156423616	12032586	24578	13907	26253	12346	1675	12232	0,5	0,1	0,47	4405060	186188	4581557	0,04
20210209	mouse	Nervous	Thalamus	24	177246717	7385280	19734	12842	21618	8776	1884	10958	0,4	0,1	0,51	2434781	327080	2751315	0,12
20210211	human	PDX	ALL	384	995409360	2590000	23578	15729	29555	13826	5977	9752	0,5	0,2	0,33	660745	217554	858678	0,23
20210211	human	PDX	ALL	38	536560853	14100000	10646	9374	17365	7992	6719	2655	0,5	0,4	0,15	61348	34768	95886	0,36
20210217	human	PDX	ALL	40	327596793	8190000	8732	7886	14537	6652	5806	2080	0,5	0,4	0,14	41534	25245	66250	0,37
20210218	human	PDX	AML	429	1233893899	2880000	16705	12757	20813	8056	4108	8649	0,4	0,2	0,42	915369	181311	1114425	0,18
20210326	mouse	Cell Line	ESC	12	157174744	13100000	24111	16245	28187	11942	4076	12169	0,4	0,1	0,43	5933541	522365	6472862	0,08
20210412	Botrytis Cinarea	Fungus	NA	30	185621636	6190000	10743	NA	NA	NA	NA	NA	NA	NA	NA	2126547	NA	NA	NA
20210413	mouse	Immune	CD8 T Cell	10	53627688	5362769	17716	9904	20374	10470	2658	7246	0,5	0,1	0,36	2753182	110626	2860232	0,04
20210414	human	Immune	DC	31	196588639	6340000	18297	13699	22559	8860	4263	9437	0,4	0,2	0,42	1420437	271109	1695214	0,16
20210414	human	Immune	FL	34	120113718	3530000	8176	6878	12738	5860	4562	2316	0,5	0,4	0,18	32528	17497	49986	0,35
20210415	human	Immune	FL	40	218417875	5460000	6658	5015	10560	5545	3902	1113	0,5	0,4	0,11	21860	9676	34946	0,37
20210415	mouse	Immune	Macrophage	50	304634373	6092687	8205	3746	10166	6420	1961	1785	0,6	0,2	0,18	73713	10132	83726	0,12
20210427	human	Immune	B Cell	32	126302924	3950000	19258	14886	24673	9787	5415	9471	0,4	0,2	0,38	1319719	150754	1480357	0,11
20210430	mouse	Immune	Monocyte	9	44186913	4909657	5172	5274	9362	4088	4190	1084	0,4	0,4	0,12	33789	12888	46603	0,27
20210430	mouse	Immune	Pericyte	20	147775553	7388778	8146	5146	11276	6130	3130	2016	0,5	0,3	0,18	58888	14306	73148	0,19
20210430	mouse	Musculoskeletal	Smooth Muscle C	9	48832166	5425796	10028	7588	14665	7077	4637	2951	0,5	0,3	0,20	94434	24284	118628	0,20
20210502	human	Immune	Classical Monocy	10	77154105	7715411	16038	14792	22904	8112	6866	7926	0,4	0,3	0,35	491084	221964	711314	0,31
20210502	human	Immune	Classical Monocy	10	85455003	8545500	17137	15231	24062	8831	6925	8306	0,4	0,3	0,35	619562	191606	809719	0,23
20210502	mouse	Cardiopulmonary	Endothelial Cell	11	61997694	5636154	18331	12997	22682	9685	4351	8646	0,4	0,2	0,38	452803	108094	559417	0,19
20210502	mouse	Immune	Kuppfer Cell	7	61042871	8720410	7666	7689	13348	5659	5682	2007	0,4	0,4	0,15	92278	22298	114536	0,19
20210502	mouse	Immune	Megakaryocyte	14	73753980	5268141	7851	5820	11804	5984	3953	1867	0,5	0,3	0,16	116326	26898	142025	0,18
20210502	human	Immune	Nonclassical Mon	10	72959389	7295939	15676	15160	22805	7645	7129	8031	0,3	0,3	0,35	516104	243798	758112	0,32
20210502	human	Immune	Nonclassical Mon	10	84192630	8419263	12781	14969	21519	6550	8738	6231	0,3	0,4	0,29	649354	363742	1006100	0,35
20210602	mouse	Nervous	Striatum	94	372405310	3961759	17297	13136	20942	7806	3645	9491	0,4	0,2	0,45	715505	232070	942622	0,24
20210602	mouse	Nervous	Striatum	94	441023073	4691735	18618	13747	22021	8274	3403	10344	0,4	0,2	0,47	1041476	421844	1477710	0,30
20210602	mouse	Nervous	Striatum	95	451305556	4750585	20092	16341	25044	8703	4952	11389	0,3	0,2	0,45	911100	389184	1303004	0,30

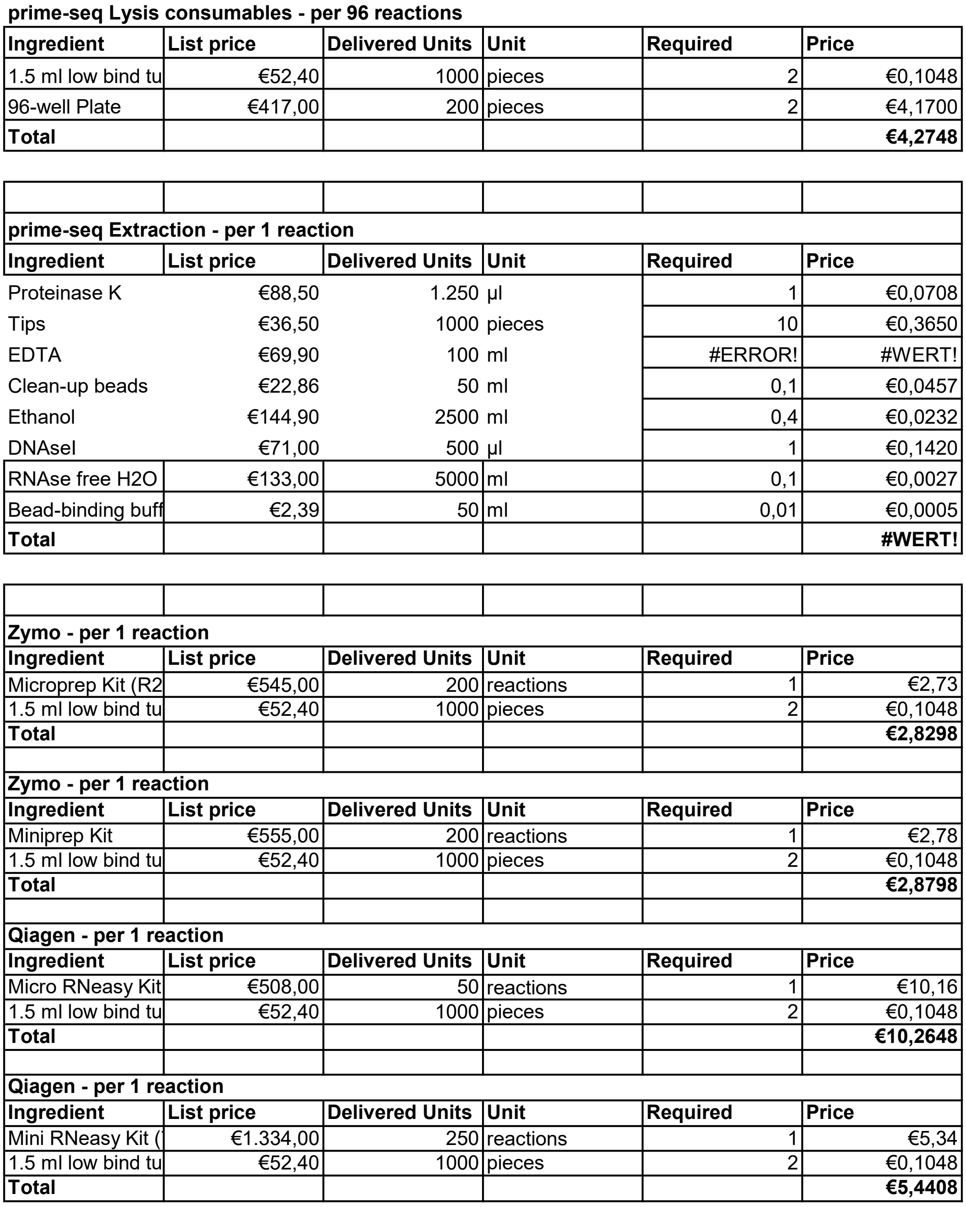
Supplemental Text 1. Magnetic beads used in prime-seq

Magnetic beads for nucleic acid isolation are a suitable alternative to column-based extraction, especially as they are more easily scalable and generally more cost efficient. These magnetic nanoparticles are coated to prevent oxidation and clumping, frequently with silica or a carboxyl coating as this provides an inert or negatively charged surface on the beads, respectively [50]. This allows for solid phase reversible immobilization (SPRI), that is the negatively charged nucleic acids can be precipitated out of solution, bind to the magnetic beads, be immobilized through the use of a magnet, washed, and then eluted without risk of irreversible binding. Additionally, carboxylated beads have carboxyl groups on the surface of the beads which form covalent bonds with nucleic acids in the presence of a crowding agent (i.e. polyethylene glycol) and high salt conditions [50]. These advantages have made carboxylated magnetic beads especially useful in isolating nucleic acids for high-throughput NGS applications.

For prime-seq, we use two different sets of beads, Carboxylated Sera-Mag SpeedBeads (GE Healthcare, now Cytiva) and SPRIselect (Beckman Coulter). The former are used in the RNA extraction step and the subsequent cleanup steps, whereas the latter are used only in the size selection steps during library preparation. Therefore, in the case of prime-seq, the Sera-Mag SpeedBeads are used to replace RNA extraction columns as well as nucleic acid concentrator columns, and the SPRISelect beads replace the gel excision and cleanup.

Numerous carboxylated magnetic beads exist for nucleic acid cleanups, with Ampure XP (Beckman Coulter) consistently used across many RNA-seq and NGS protocols. Additionally, the Ampure XP beads have worked well in our hands when used for standard nucleic acid cleanups. Within the prime-seq protocol, however, we specifically do not use the Ampure XP beads due to the cost factor involved. For example, a 24 sample prime-seq experiment would require 2.46 mL of Ampure XP beads and 2.7 mL of diluted Sera-Mag SpeedBeads prior to library preparation, which amounts to $203 and $2.25, respectively. The almost 90-fold lower cost makes it apparent that the Sera-Mag SpeedBeads are a better choice for many researchers. It could be possible that the Ampure XP beads yield a better nucleic acid recovery or provide more consistent performance, however, when tested we recover 80 % of input with the Sera-Mag SpeedBeads. Thus, even if there were slight improvements in performance, this would not outweigh the substantial increase in cost.

During library preparation, we are not only cleaning the solution to remove residual primers and salts, but are also specifically selecting for a range of fragment sizes (e.g. 300-800 bp). And, although both the Sera-Mag SpeedBeads and SPRIselect beads are carboxylated magnetic beads, the SPRIselect beads are validated for size selection properties ensuring consistent performance between lots. From our experience it is unclear if there are added advantages of using the SPRISelect beads for the library size selection, but as our Sera-Mag SpeedBead lot-to-lot verifications are not as rigorous as those performed by Beckman Coulter, and one only needs 70 µL of SPRISelect per prime-seq library, the increased cost of using SPRISelect for this portion of the protocol does not substantially alter the overall cost.



prime-seq Lysis consumables - per 96 reactions

Ingredient	List price	Delivered Units	Unit	Required	Price
1.5 ml low bind tu	€52,40	1000	pieces	2	€0,1048
96-well Plate	€417,00	200	pieces	2	€4,1700
Total					**€4,2748**

prime-seq Extraction - per 1 reaction

Ingredient	List price	Delivered Units	Unit	Required	Price
Proteinase K	€88,50	1.250	µl	1	€0,0708
Tips	€36,50	1000	pieces	10	€0,3650
EDTA	€69,90	100	ml	#ERROR!	#WERT!
Clean-up beads	€22,86	50	ml	0,1	€0,0457
Ethanol	€144,90	2500	ml	0,4	€0,0232
DNAseI	€71,00	500	µl	1	€0,1420
RNAse free H2O	€133,00	5000	ml	0,1	€0,0027
Bead-binding buff	€2,39	50	ml	0,01	€0,0005
Total					**#WERT!**

Zymo - per 1 reaction

Ingredient	List price	Delivered Units	Unit	Required	Price
Microprep Kit (R2	€545,00	200	reactions	1	€2,73
1.5 ml low bind tu	€52,40	1000	pieces	2	€0,1048
Total					**€2,8298**

Zymo - per 1 reaction

Ingredient	List price	Delivered Units	Unit	Required	Price
Miniprep Kit	€555,00	200	reactions	1	€2,78
1.5 ml low bind tu	€52,40	1000	pieces	2	€0,1048
Total					**€2,8798**

Qiagen - per 1 reaction

Ingredient	List price	Delivered Units	Unit	Required	Price
Micro RNeasy Kit	€508,00	50	reactions	1	€10,16
1.5 ml low bind tu	€52,40	1000	pieces	2	€0,1048
Total					**€10,2648**

Qiagen - per 1 reaction

Ingredient	List price	Delivered Units	Unit	Required	Price
Mini RNeasy Kit (	€1.334,00	250	reactions	1	€5,34
1.5 ml low bind tu	€52,40	1000	pieces	2	€0,1048
Total					**€5,4408**

Hands-On Requi	Task	Beads (24)	Beads (48)	Beads (96)
Hands-on	Workspace Prepa	2	2	2
Hands-on	Sample Preparati	1	2	3
Hands-on	Add ProtK	1	1	1
Hands-off	Incubate	25	25	25
Hands-on	Add Beads	1	1	1
Hands-off	Incubate	5	5	5
Hands-off	Magnet	3	3	3
Hands-on	Wash	4	6	8
Hands-off	Dry	3	3	3
Hands-on	Add Water and R	2	2	2
Hands-on	Add DNase	1	1	1
Hands-off	Incubate	16	16	16
Hands-off	Magnet	3	3	3
Hands-on	Wash	4	6	8
Hands-off	Dry	3	3	3
Hands-off	Elute	5	5	5
	Total	**79**	**84**	**89**
	Hands On	16	21	26
	Hands Off	63	63	63

Hands-On Requi	Task	Column (24)	Column (48)	Column (96)
Hands-on	Add ProtK	1	2	4
Hands-off	Incubate	30	30	30
Hands-on	Workspace Prepa	2	2	2
Hands-on	Sample Preparati	1	2	3
Hands-on	Add EtOH	1	2	4
Hands-on	Transfer to Colum	3	6	12
Hands-on	Add to Centrifuge	1	2	4
Hands-off	Spin	0,5	0,5	0,5
Hands-on	Remove from Cer	2	4	8
Hands-on	Discard Flow Thro	2	4	8
Hands-on	Add RNA Wash	1	2	4
Hands-on	Add to Centrifuge	1	2	4
Hands-off	Spin	0,5	0,5	0,5
Hands-on	Remove from Cer	2	4	8
Hands-on	Discard Flow Thro	2	4	8
Hands-on	Add DNAse	2	4	8
Hands-off	Incubate	15	15	15
Hands-on	Add to Centrifuge	1	2	4
Hands-off	Spin	0,5	0,5	0,5
Hands-on	Remove from Cer	2	4	8
Hands-on	Discard Flow Thro	2	4	8
Hands-on	Add Pre Wash	1	2	4
Hands-on	Add to Centrifuge	1	2	4
Hands-off	Spin	0,5	0,5	0,5
Hands-on	Remove from Cer	2	4	8
Hands-on	Discard Flow Thro	2	4	8
Hands-on	Add RNA Wash	1	2	4
Hands-on	Add to Centrifuge	1	2	4
Hands-off	Spin	1	1	1
Hands-on	Remove from Cer	2	4	8
Hands-on	Transfer Column	2	4	8
Hands-on	Add H2O	1	2	4
Hands-off	Incubate	1	1	1
Hands-on	Add to Centrifuge	1	2	4
Hands-off	Spin	1	1	1
Hands-on	Remove from Cer	2	4	8
	Total	**92**	**132**	**211**
	Hands On	42	82	161
	Hands Off	50	50	50

	Catalog Number	Cost	Amount	Units	Required	Cost	Required	Cost	Required	Cost	Required	Cost	Required	Cost	Required	Cost	Required	Cost	Required	Cost	Required	Cost	Required	Cost
Method					Ligation		Prep		Kit v2, Set A or B		RNA Kit for Sequencing		Sample Prep Kit - HT		prime-seq		brb-seq (Nextera)		brb-seq (Custom Transposase)		NEBNext		DECODE-seq	
Kits																								
Stranded mRNA	20040534	€3.964,00	1	kit (96 rxn)	1	€3.964,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00		€0,00	0	€0,00
mRNA Library	20020595	€4.261,00	1	kit (96 rxn)	0	€0,00	1	€4.261,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00		€0,00	0	€0,00
Library	RS-122-2001/2	€3.750,00	1	kit (48 rxn)	0	€0,00	0	€0,00	2	€7.500,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00		€0,00	0	€0,00
Ultra Low Input	634891	€6.360,00	1	kit (96 rxn)	0	€0,00	0	€0,00	0	€0,00	1	€6.360,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00		€0,00	0	€0,00
Stranded Total	634862	€5.004,00	1	kit (96 rxn)	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€5.004,00	0	€0,00	0	€0,00	0	€0,00		€0,00	0	€0,00
Nextera XT	FC-131-1096	€3.179,37	1	kit (96 rxn)	0	€0,00	0	€0,00	0	€0,00	1	€3.179,37	0	€0,00	0	€0,00	0,010416667	€33,12	0	€0,00		€0,00	0,010416667	€33,12
FS	E6177S	€6,08	1	reaction	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€6,08	0	€0,00	0	€0,00		€0,00	0	€0,00
RiboGone	634847	€1.860,00	1	kit (24 rxn)	0	€0,00	0	€0,00	0	€0,00	0	€0,00	4	€7.440,00	0	€0,00	0	€0,00	0	€0,00		€0,00	0	€0,00
Transposase	C01070010-20	€4,60	1	reaction	0	€0,00	0	€0,00	0	€0,00	1	€4,60	0	€0,00	0	€0,00	0	€0,00	1	€4,60		€0,00	0	€0,00
Poly(A) mRNA	E7490	€247,00	1	kit (96 rxn)	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€247,00	0	€0,00
RNA Library	E7775L	€3.685,00	1	kit (96 rxn)	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€3.685,00	0	€0,00
Library Indicies																								
RNA UD Indexes	20040553	€157,75	96	sample	1	€157,75	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00
Index Plate	20019792	€128,25	96	sample	0	€0,00	1	€128,25	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00
prime-seq TSO	Sigma	€90,68	1	μmol	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0,001104	€0,10	0	€0,00	0	€0,00	0	€0,00	0	€0,00
prime-seq barcod	Sigma	€1.607,04	1	nmol	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	#ERROR!	#WERT!	0	€0,00	0	€0,00	0	€0,00	0	€0,00
prime-seq pre-am	Sigma	€0,02	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€0,02	0	€0,00	0	€0,00	0	€0,00	0,25	€0,00
Indicies	IDT	€0,18	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€0,18	0	€0,00	0	€0,00	0	€0,00	5	€0,89
prime-seq p5	Sigma	€0,04	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€0,04	0	€0,00	0	€0,00	0	€0,00	0	€0,00
BRB-seq barcoded oligo-dT	Microsynth	€1,99	96	samples	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€1,99	1	€1,99	0	€0,00	0	€0,00
BRB-seq i7	IDT	€0,18	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	2,5	€0,45	2,5	€0,45	0	€0,00	0	€0,00
BRB-seq p5	Microsynth	#ERROR!	1	μl	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	2,5	#WERT!	2,5	#WERT!	0	#WERT!	0	#WERT!
NEBNext Multiple	E7600S	€535,00	1	kit (96 rxn)	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€535,00	0	€0,00
5bc-TSO (10 uM)	IDT	#ERROR!	1	μl	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	96	#WERT!
5bc-RT (10 uM)	Sigma	#ERROR!	1	μl	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	#ERROR!	#WERT!
Single-P primer	Sigma	€0,02	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	#ERROR!	#WERT!
P5Read1 (5 uM)	Sigma	€0,04	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	2,5	€0,09
Reagents																								
UltraPure Water	10977-049	€0,03	1	ml	21	€0,56	31	€0,82	38,4	€1,02	16	€0,43	23,04	€0,61	0,554	€0,01	0,554	€0,01	0,554	€0,01	0	€0,00	#ERROR!	#WERT!
Ethanol	32221-2.5L-M	#ERROR!	1	ml	80	#WERT!	123	#WERT!	153,6	#WERT!	62	#WERT!	92,16	#WERT!	10,2	#WERT!	5	#WERT!	5	#WERT!	115,2	#WERT!	1	#WERT!
Maxima H- Rever	EP0753	€3,18	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	#ERROR!	#WERT!	0	€0,00	0	€0,00	0	€0,00	0	€0,00
SuperScript II Reverse Transcriptase	18064-014	€6,50	1	ul	0	€0,00	96	€624,00	96	€624,00	0	€0,00	0	€0,00	0	€0,00	#ERROR!	#WERT!	#ERROR!	#WERT!	0	€0,00	#ERROR!	#WERT!
Q5 Master Mix 2x	M0544L	€0,79	1	reaction	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€0,79	0	€0,00	0	€0,00	0	€0,00	0	€0,00
NEBNext High-Fidelity 2X PCR MM	M0541L	€1,44	1	reaction	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€1,44	1	€1,44	0	€0,00	0	€0,00
KAPA HiFI ready	7958935001	€0,06	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	25	€1,54	0	€0,00	0	€0,00	0	€0,00	#ERROR!	#WERT!
Exonuclease I	M0293L	€0,38	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€0,38	1	€0,38	1	€0,38	0	€0,00	0	€0,00
100mM Tris-HCl (pH6.9)	A3452	€0,23	1	g	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0,00016	€0,00	0,00016	€0,00	0	€0,00	0	€0,00
25mM MgCl2	M2670	€0,35	1	g	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0,00005	€0,00	0,00005	€0,00	0	€0,00	0	€0,00
450mM KCl	A2939	€0,48	1	g	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0,00059	€0,00	0,00059	€0,00	0	€0,00	0	€0,00
β-NAD	N0632	€54,00	1	g	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0,00001	€0,00	0,00001	€0,00	0	€0,00	0	€0,00
(NH4)2SO4	AC20587	€0,18	1	g	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0,00008	€0,00	0,00008	€0,00	0	€0,00	0	€0,00
RNAse H	M0297S	€1,42	1	μL	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	2	€2,84	2	€2,84	0	€0,00	0	€0,00
E. coli DNA ligase	M0205 L	€2,52	1	μL	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€2,52	1	€2,52	0	€0,00	0	€0,00
E. coli DNA Polymerase	M0209 L	€1,10	1	μL	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	5	€5,52	5	€5,52	0	€0,00	0	€0,00
dNTPs (25 mM)	R0186	€0,02	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	#ERROR!	#WERT!	96	€1,60	96	€1,60	0	€0,00	#ERROR!	#WERT!
SPRI Select Rea	B23317	€0,06	1	μl	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	72	€4,36	0	€0,00	0	€0,00	0	€0,00	#ERROR!	#WERT!
Clean up beads	GE65152105050	€0,46	1	ml	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€0,46	0	€0,00	0	€0,00	0	€0,00	0	€0,00
Agencourt AMPureXP	A63881	€0,02	1	μL	16704	€345,22	#ERROR!	#WERT!	37632	€777,73	7680	€158,72	8640	€178,56	0	€0,00	90	€1,86	90	€1,86	0	€0,00	0	€0,00
5M Betaine	B0300	#ERROR!	1	μL	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	192	#WERT!
100 mM DTT	43816	#ERROR!	1	μL	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	#ERROR!	#WERT!
1 M MgCl2	M1028	#ERROR!	1	μL	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	#ERROR!	#WERT!
Recombinant RNase Inhibitor	2313A	€0,92	1	μL	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	#ERROR!	#WERT!
Consumables																								
Quick PCR purification Kit	GO-PCRF-100	€1,00	1	reaction	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	2	€2,00
Gel Extraction Beads Kit	GO-GELU-100	€1,00	1	reaction	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€1,00
DNA Clean & Concentrator-5 kit	D4014	€1,36	1	reaction	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	0	€0,00	1	€1,36	1	€1,36	0	€0,00	0	€0,00
Filter Tips (10,20, and 200)	737257	#ERROR!	1	tip	2714	#WERT!	2815	#WERT!	2825	#WERT!	1276	#WERT!	985	#WERT!	255	#WERT!	255	#WERT!	255	#WERT!	2325	#WERT!	136	#WERT!
1.5 ml low bind tubes	0030108051	#ERROR!	1	tube	2	#WERT!	1	#WERT!	0	#WERT!	4	#WERT!	3	#WERT!	5	#WERT!	5	#WERT!	5	#WERT!	14	#WERT!	2	#WERT!
96-well plate	0030129504	€5,60	1	plate	5	€28,00	8	€44,80	7	€39,20	4	€22,40	4	€22,40	2	€11,20	2	€11,20	2	€11,20	4	€22,40	2	€11,20
PCR Seal	04-097-0100	€1,11	1	seal	16	€17,76	22	€24,42	24	€26,64	8	€8,88	10	€11,10	5	€5,55	5	€5,55	5	€5,55	5	€5,55	4	€4,44
Stepper Tip		#ERROR!	1	tip	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	0	#WERT!	3	#WERT!	3	#WERT!	3	#WERT!	18	#WERT!	3	#WERT!
QC																								
Agilent DNA 1000 Kit	5067-1504	€53,80	1	chip	2	€107,60	2	€107,60	2	€107,60	2	€107,60	2	€107,60	2	€107,60	2	€107,60	2	€107,60	2	€107,60	2	€107,60

		Stranded mRNA	Stranded mRNA	Library	SMARTer	(homemade)	SMART-seq v4	NEBNext	DECODE-seq	(with pooling)	BRB-seq	prime-seq	
	Hands-on	166	191	#WERT!	53	52	52	#WERT!	54	30,0	#WERT!	#WERT!	
	Hands-off	245	245	551	168	197	197	122	298	148,0	342	370	
	Total	411	436	#WERT!	221	249	249	#WERT!	352	#WERT!	#WERT!	#WERT!	
	cDNA synthesis	86	86	#WERT!	26,5	26	26	#WERT!	27	15,0	#WERT!	#WERT!	
Hands-on	Library preparatio	80	105	#WERT!	26,5	26	26	#WERT!	27	15,0	#WERT!	#WERT!	
	cDNA synthesis	145	145	246	84	157	157	122	298	148,0	280	249	
Hands-off	Library preparatio	100	100	305	84	40	40	115	0	#ERROR!	62	121	
	cDNA synthesis	231	231	#WERT!	110,5	183	183	#WERT!	325	163,0	#WERT!	#WERT!	
Total	Library preparatio	180	205	#WERT!	110,5	66	66	#WERT!	27	#WERT!	#WERT!	#WERT!	
	bead clean up 96	#ERROR!	#ERROR!	#ERROR!	3	2	2	5	4	0	0	0	
	clean up pool (co	0	0	0	0	0	0	#ERROR!	0	4	#ERROR!	#ERROR!	
	add master mix 9	8	9	9	3	7	7	10	4	3	1	1	
	add individually p	4	3	3	4	5	5	1	2	1	1	1	
	add master mix to	0	0	0	0	0	0	0	0	2	5	4	
													time
25	bead clean up 96	#WERT!	#WERT!	#WERT!	30	20	20	50	40	0	0	0	10
15	bead clean up po	0	0	0	0	0	0	#WERT!	0	20	#WERT!	#WERT!	5
5	add master mix 9	8	9	9	3	7	7	10	4	3	1	1	1
10	add per well	20	15	15	20	25	25	5	10	5	5	5	5
1	add master mix p	0	0	0	0	0	0	0	0	2	5	4	1
	Total Hands on ti	#WERT!	#WERT!	#WERT!	53	52	52	#WERT!	54	30	#WERT!	#WERT!	22

2.3 Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq

Bagnoli, Johannes W. and Ziegenhain, Christoph and Janjic, Aleksandar, Wange, Lucas E., Vieth, Beate, Parekh, Swati, **Geuder, Johanna**, Hellmann, Ines, Enard, Wolfgang

"Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq" (2018)

Nature Communications 9, 2937 (2018).

doi: https://doi.org/10.1038/s41467-018-05347-6

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

`https://www.nature.com/articles/s41467-018-05347-6#Sec33`

Abstract

Single-cell RNA sequencing (scRNA-seq) has emerged as a central genome-wide method to characterize cellular identities and processes. Consequently, improving its sensitivity, flexibility, and cost-efficiency can advance many research questions. Among the flexible plate-based methods, single-cell RNA barcoding and sequencing (SCRB-seq) is highly sensitive and efficient. Here, we systematically evaluate experimental conditions of this protocol and find that adding polyethylene glycol considerably increases sensitivity by enhancing cDNA synthesis. Furthermore, using Terra polymerase increases efficiency due to a more even cDNA amplification that requires less sequencing of libraries. We combined these and other improvements to develop a scRNA-seq library protocol we call molecular crowding SCRB-seq (mcSCRB-seq), which we show to be one of the most sensitive, efficient, and flexible scRNA-seq methods to date.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-05347-6 OPEN

Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq

Johannes W. Bagnoli[1], Christoph Ziegenhain[1,2], Aleksandar Janjic[1], Lucas E. Wange[1], Beate Vieth[1], Swati Parekh[1,3], Johanna Geuder[1], Ines Hellmann[1] & Wolfgang Enard[1]

Single-cell RNA sequencing (scRNA-seq) has emerged as a central genome-wide method to characterize cellular identities and processes. Consequently, improving its sensitivity, flexibility, and cost-efficiency can advance many research questions. Among the flexible plate-based methods, single-cell RNA barcoding and sequencing (SCRB-seq) is highly sensitive and efficient. Here, we systematically evaluate experimental conditions of this protocol and find that adding polyethylene glycol considerably increases sensitivity by enhancing cDNA synthesis. Furthermore, using Terra polymerase increases efficiency due to a more even cDNA amplification that requires less sequencing of libraries. We combined these and other improvements to develop a scRNA-seq library protocol we call molecular crowding SCRB-seq (mcSCRB-seq), which we show to be one of the most sensitive, efficient, and flexible scRNA-seq methods to date.

---

[1] Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Großhaderner Straße 2, 82152 Martinsried, Germany. [2] Present address: Department of Cell & Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden. [3] Present address: Max Planck Institute for Biology of Ageing, 50931 Cologne, Germany. These authors contributed equally: Johannes W. Bagnoli, Christoph Ziegenhain, Aleksandar Janjic. Correspondence and requests for materials should be addressed to W.E. (email: enard@bio.lmu.de)

Whole transcriptome single-cell RNA sequencing (scRNA-seq) is a transformative tool with wide applicability to biological and biomedical questions[1,2]. Recently, many scRNA-seq protocols have been developed to overcome the challenge of isolating, reverse transcribing, and amplifying the small amounts of mRNA in single cells to generate high-throughput sequencing libraries[3,4]. However, as there is no optimal, one-size-fits all protocol, various inherent strengths and trade-offs exist[5–7]. Among flexible, plate-based methods, single-cell RNA barcoding and sequencing (SCRB-seq)[8] is one of the most powerful and cost-efficient[6], as it combines good sensitivity, the use of unique molecular identifiers (UMIs) to remove amplification bias and early cell barcodes to reduce costs. Here, we systematically optimize the sensitivity and efficiency of SCRB-seq and generate molecular crowding SCRB-seq (mcSCRB-seq), one of the most powerful and cost-efficient plate-based methods to date (Fig. 1a).

Results

Systematic optimization of SCRB-seq. We started to test improvements to SCRB-seq by optimizing the cDNA yield and quality generated from universal human reference RNA (UHRR)[9] in a standardized SCRB-seq assay (see Supplementary Fig. 1a and Methods). By including the barcoded oligo-dT primers in the lysis buffer, we increased cDNA yield by 10% and avoid a time-consuming pipetting step during the critical phase of the protocol (Supplementary Fig. 1b). Next, we compared the performance of nine Moloney murine leukemia virus (MMLV) reverse transcriptase (RT) enzymes that have the necessary template-switching properties. Especially at input amounts below 100 pg, Maxima H- (Thermo Fisher) performed best closely followed by SmartScribe (Clontech) (Supplementary Fig. 1c). In order to reduce the costs of the reaction, we showed that cDNA yield and quality is not measurably affected when we reduced the enzyme (Maxima H-) by 20%, reduced the oligo-dT primer by 80%, or used the cheaper unblocked template-switching oligo (Supplementary Fig. 2). Next, we evaluated the effect of $MgCl_2$, betaine and trehalose, as these led to the increased sensitivity of the Smart-seq2 protocol[10]. Since both Smart-seq2 and SCRB-seq generate cDNA by oligo-dT priming, template switching, and PCR amplification, we were surprised that these additives decreased cDNA yield for SCRB-seq (Supplementary Fig. 3a). Apparently, the interactions between enzymes and buffer conditions are complex and optimizations cannot be easily transferred from one protocol to another.

Molecular crowding significantly increases sensitivity. An additive that has not yet been explored for scRNA-seq protocols is polyethylene glycol (PEG 8000). It makes ligation reactions more efficient[11] and is thought to increase enzymatic reaction rates by mimicking (macro)molecular crowding, i.e., by reducing the effective reaction volume[12]. As small reaction volumes can increase the sensitivity of scRNA-seq protocols[5,13], we tested whether PEG 8000 can also increase the cDNA yield of SCRB-seq. Indeed, we observed that PEG 8000 increased cDNA yield in a concentration-dependent manner up to tenfold (Supplementary Fig. 3b). However, at higher PEG concentrations, unspecific DNA fragments accumulated in reactions without RNA (Supplementary Fig. 3d) and therefore we chose 7.5% PEG 8000 as an optimal concentration balancing yield and specificity (Supplementary

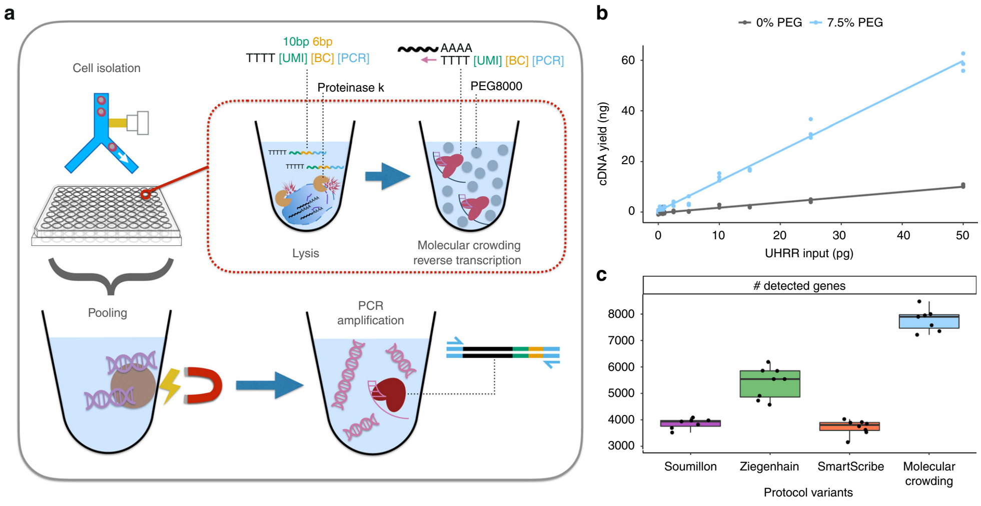


Fig. 1 mcSCRB-seq workflow and the effect of molecular crowding. **a** Overview of the mcSCRB-seq protocol workflow. Single cells are isolated via FACS in multiwell plates containing lysis buffer, barcoded oligo-dT primers, and Proteinase K. Reverse transcription and template switching are carried out in the presence of 7.5% PEG 8000 to induce molecular crowding conditions. After pooling the barcoded cDNA with magnetic SPRI beads, PCR amplification using Terra polymerase is performed. **b** cDNA yield dependent on the absence (gray) or presence (blue) of 7.5% PEG 8000 during reverse transcription and template switching. Shown are three independent reactions for each input concentration of total standardized RNA (UHRR) and the resulting linear model fit. **c** Number of genes detected (>=1 exonic read) per replicate in RNA-seq libraries, generated from 10 pg of UHRR using four protocol variants (see Supplementary Table 1) at a sequencing depth of one million raw reads. Each dot represents a replicate ($n = 8$) and each box represents the median and first and third quartiles per method with the whiskers indicating the most extreme data point, which is no more than 1.5 times the length of the box away from the box

Fig. 3c). With the addition of PEG 8000, yield increased substantially, making it possible to detect RNA inputs under 1 pg (Fig. 1b).

To test whether these increases in cDNA yield indeed correspond to increases in sensitivity, we generated and sequenced 32 RNA-seq libraries from 10 pg of total RNA (UHRR) using eight replicates for each of the following four SCRB-seq protocol variants (Supplementary Tables 1, 2): the original SCRB-seq protocol[8] ("Soumillon"; with Maxima H- as RT and Advantage2 as PCR enzyme), the slightly adapted protocol benchmarked in Ziegenhain et al.[6] ("Ziegenhain"; with Maxima H- and KAPA), the same protocol with SmartScribe as the RT enzyme ("SmartScribe") and our optimized protocol ("molecular crowding"; with Maxima H-, KAPA, 7.5% PEG, 80% less oligo-dT, and 20% less Maxima H-). As expected, the molecular crowding protocol yielded the most cDNA, while variant "Soumillon" yielded the least, confirming our systematic optimization (Supplementary Fig. 4a). After sequencing, we processed data using *zUMIs*[14] and downsampled each of the 32 libraries to one million reads per sample, which has been suggested to correspond to reasonable saturation for single-cell RNA-seq experiments[5,6]. Of the 32 libraries, 31 passed quality control with a median of 71% of the reads mapping to exons (range: 50–77%), 12% to introns (9–15%), 13% to intergenic regions (10–31%), and 4% (3–7%) to no region in the human genome (Supplementary Fig. 4b). Of note, we observe that a higher proportion of reads are mapping to intergenic regions for the "molecular crowding" condition (Supplementary Fig. 4b). As UHRR is provided as DNAse-digested RNA, these reads are likely derived from endogenous transcripts, but why their proportion is increased in the molecular crowding protocol is unclear. In any case, we assessed the sensitivity of the protocols by the number of detected genes per cell (>=1 exonic read), representing a conservative estimate for the molecular crowding protocol with its higher fraction of intergenic reads (Fig. 1c). This sensitivity measure correlates fairly well with cDNA yield (Supplementary Fig. 4a). Hence, it shows that Maxima H- is indeed more sensitive than SmartScribe (5542 detected genes per sample in "Ziegenhain" vs. 3805 in "SmartScribe", $p = 3 \times 10^{-5}$, Welch two sample *t*-test) and that the molecular crowding protocol is the most sensitive one (7898 vs. 5542 detected genes, $p = 7 \times 10^{-7}$, Welch two sample *t*-test). In summary, we can show that our optimized SCRB-seq protocol, in particular due to the addition of PEG 8000, increases the sensitivity compared to previous protocol variants at reduced costs.

Terra retains more complexity during cDNA amplification. Next, we aimed to increase the efficiency of this protocol by optimizing the cDNA amplification step. Depending on the number of cycles, reaction conditions, and polymerases, substantial noise and bias is introduced when the small amounts of cDNA molecules are amplified by PCR[15,16]. While UMIs allow for the correction of these effects computationally, scRNA-seq methods that have less amplification bias require fewer reads to obtain the same number of UMIs and hence are more efficient[6,17]. As a first step, we evaluated 12 polymerases for cDNA yield and found KAPA, SeqAmp, and Terra to perform best (Supplementary Fig. 5a). We disregarded SeqAmp because of a decreased median length of the amplified cDNA molecules (Supplementary Fig. 5b) as well as the higher cost of the enzyme and continued to compare the amplification bias of KAPA and Terra polymerases. To this end, we sorted 64 single mouse embryonic stem cells (mESCs) and generated cDNA using our optimized molecular crowding protocol. Two pools of cDNA from 32 cells were amplified with KAPA or Terra polymerase (18 cycles) and used to generate libraries. After sequencing and downsampling each transcriptome to one million raw reads[14], we found that amplification using Terra yielded twice as much library complexity (UMIs) than when using KAPA (Supplementary Fig. 5c). This is in agreement with a recent study that optimized the scRNA-seq protocol Quartz-seq2, which also found Terra to retain a higher library complexity[17]. In addition to choosing Terra for cDNA amplification, we also reduced the number of cycles from 19 in the original SCRB-seq protocol to 14, as fewer cycles are expected to decrease amplification bias further[15] and 14 cycles still generated sufficient amounts of cDNA (~1.6–2.4 ng/μl) from mouse ESCs to prepare libraries with Nextera XT (~0.8 ng needed). Depending on the investigated cells, which may have a lower or higher RNA content than ESCs, the cycle number might need to be adapted to generate enough cDNA while avoiding overcycling.

With the final improved version of the molecular crowding protocol (mcSCRB-seq), we tested to what extent cross-contamination occurs. For example, chimeric PCR products may occur following the pooling of cDNA[18] and we assessed whether this might potentially be influenced by PEG that is present during cDNA synthesis before pooling. To this end, we sorted 96 cells of a mixture of mESCs and human-induced pluripotent stem cells, synthesized cDNA according to the mcSCRB-seq protocol with and without the addition of PEG and generated libraries for each of the two conditions. After mapping the sequenced reads to the joint human and mouse reference genomes, each barcode/well could be clearly classified into human or mouse cells, indicating that no doublets were sorted into wells, as may be expected for a fluorescence-activated cell sorting (FACS)-based cell isolation (Supplementary Fig. 6a). Importantly, the median number of reads mapping best to the wrong species is less than 2000 per cell (<0.4% of all reads or <1.5% of uniquely mapped reads). This is not influenced by the addition of PEG, as may be expected, since PEG is only present during cDNA generation (Supplementary Fig. 6b; two-sided *t*-test, p value = 0.81). In summary, we developed an optimized protocol, mcSCRB-seq, that has higher sensitivity, a less biased amplification and little crosstalk of reads across cells.

mcSCRB-seq increases sensitivity 2.5-fold more than SCRB-seq. To directly compare the entire mcSCRB-seq protocol to the previously benchmarked SCRB-seq protocol used in Ziegenhain et al.[6] (Supplementary Table 2), we sorted for each method 48 and 96 single mESCs from one culture into plates, and added ERCC spike-ins[19]. Following sequencing, we filtered cells to discard doublets/dividing cells, broken cells, and failed libraries (see Methods). The remaining 249 high-quality libraries all show a similar mapping distribution with ~50% of reads falling into exonic regions (Supplementary Fig. 7). When plotting the number of detected endogenous mRNAs (UMIs) against sequencing depth, mcSCRB-seq clearly outperforms SCRB-seq and detects 2.5 times as many UMIs per cell at depths above 200,000 reads (Fig. 2a and Supplementary Fig. 8a). At two million reads, mcSCRB-seq detected a median of 102,282 UMIs per cell and a median of 34,760 ERCC molecules, representing 48.9% of all spiked in ERCC molecules (Supplementary Fig. 8b). Assuming that the efficiency of detecting ERCC molecules is representative of the efficiency to detect endogenous mRNAs, the median content per mESC is 227,467 molecules (Supplementary Fig. 8c and 8d), which is very similar to previous estimates using mESCs and STRT-seq, a 5′ tagged UMI-based scRNA-seq protocol[20]. As expected, the higher number of UMIs in mcSCRB-seq also results in a higher number of detected genes. For instance, at 500,000 reads, mcSCRB-seq detected 50,969 UMIs that corresponded to

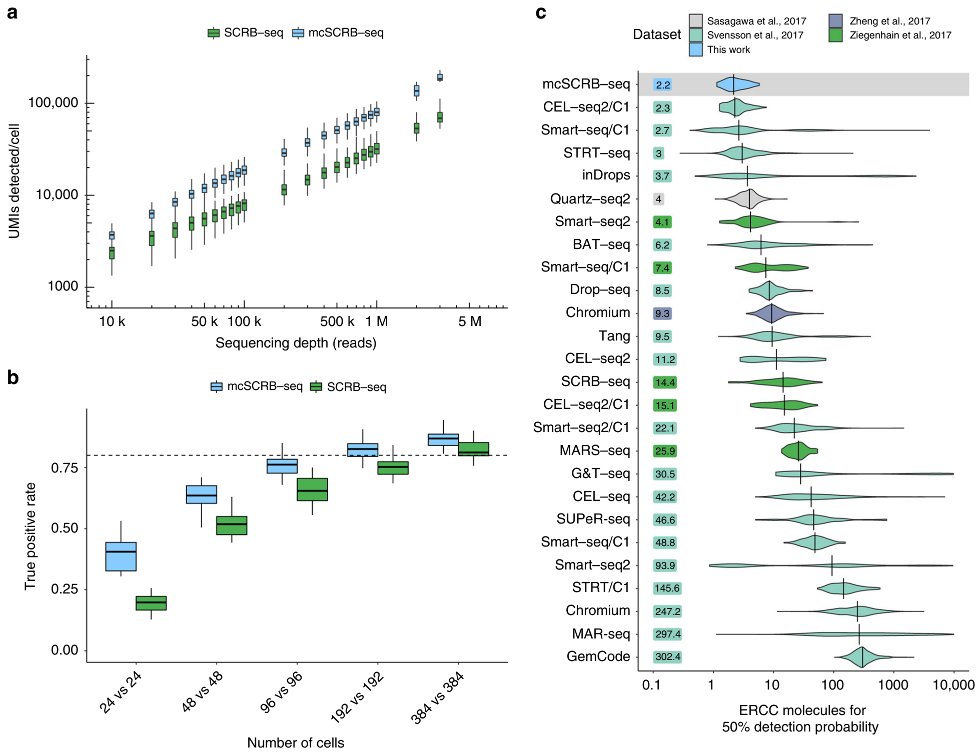


Fig. 2 Comparison of mcSCRB-seq to SCRB-seq and other protocols. **a** Number of UMIs detected in libraries generated from 249 single mESCs using SCRB-seq or mcSCRB-seq when downsampled to different numbers of raw sequence reads. Each box represents the median and first and third quartiles per cell, sequencing depth and method. Whiskers indicate the most extreme data point that is no more than 1.5 times the length of the box away from the box. **b** The true positive rate of mcSCRB-seq and SCRB-seq estimated by power simulations using the powsimR package[22]. The empirical mean–variance distribution of the 10,904 genes that were detected in at least 10 cells in either mcSCRB-seq or SCRB-seq (500,000 reads) was used to simulate read counts when 10% of the genes are differentially expressed. Boxplots represent the median and first and third quartiles of 25 simulations with whiskers indicating the most extreme data point hat is no more than 1.5 times the length of the box away from the box. The dashed line indicates a true positive rate of 0.8. The matching plot for the false discovery rate is shown in Supplementary Fig. 11d. **c** Sensitivity of mcSCRB-seq and other protocols, calculated as the number of ERCC molecules needed to reach a 50% detection probability as calculated in Svensson et al[5]. Per-cell distributions are shown using violin plots with vertical lines and numbers indicating the median per protocol

5866 different genes, 1000 more than SCRB-seq (Supplementary Fig. 9). Congruent with the above comparison of Terra and KAPA polymerase, mcSCRB-seq showed a less noisy and less-biased amplification (Supplementary Fig. 10). Furthermore, expression levels differed much less between the two batches of mcSCRB-seq libraries, indicating that it could be more robust than SCRB-seq (Supplementary Fig. 11a). In contrast to findings for other protocols[21], neither mcSCRB-seq nor SCRB-seq showed GC content or transcript length-dependent expression levels (Supplementary Fig. 11b, c).

Decisively, we find by using power simulations[6,22] that mcSCRB-seq requires approximately half as many cells as SCRB-seq to detect differentially expressed genes between two groups of cells (Fig. 2b and Supplementary Fig. 11d). Hence, the higher sensitivity and lower noise of mcSCRB-seq compared to SCRB-seq, as measured in parallelly processed cells, indeed matters for quantifying gene expression levels and can be quantified as a doubling of cost-efficiency. Furthermore, we have reduced the reagent costs from about 1.70 € per cell for SCRB-seq[6] to less than 0.54 € for mcSCRB-seq (Supplementary Fig. 12a and Supplementary Table 3). Together, this makes mcSCRB-seq sixfold more cost-efficient than SCRB-seq. Moreover, owing to an optimized workflow, we could reduce the library preparation time to one working day with minimal hands-on time (Supplementary Fig. 12b and Supplementary Table 4). As SCRB-seq was already one of the most cost-efficient protocols in our recent benchmarking study[6], this likely makes mcSCRB-seq the most cost-efficient plate-based method available.

Benchmarking by ERCCs. The widespread use of ERCC spike-ins also allows us to estimate and compare the absolute sensitivity across many scRNA-seq protocols using published data[5]. As in Svensson et al.[5], we used a binomial logistic regression to estimate the number of ERCC transcripts that are needed on average to reach a 50% detection probability (Supplementary Fig. 13a).

mcSCRB-seq reached this threshold with 2.2 molecules, when ERCCs are sequenced to saturation (Supplementary Fig. 13b). When comparing this to a total of 26 estimates for 20 different protocols obtained from two major protocol comparisons[5,6] as well as additional relevant protocols[17,23], mcSCRB-seq has the highest sensitivity among all protocols compared to date (Fig. 2c). It should be noted that the data show large amounts of variation within protocols, even for well-established, sensitive methods like Smart-seq2. This is the case, especially in Svensson et al.[5], because the data were generated from many varying cell types sequenced in numerous labs. Similarly, mcSCRB-seq sensitivity estimates could be variable across labs and conditions. Nevertheless, the average ERCC detection efficiency is the most representative measure to compare sensitivities across many protocols.

mcSCRB-seq detects biological differences in complex tissues. Finally, we applied mcSCRB-seq to peripheral blood mononuclear cells (PBMCs), a complex cell population with low mRNA amounts, to test whether it is efficient in recapitulating biological differences. We obtained PBMCs from one healthy donor, FACS-sorted cells in four 96-well plates and prepared libraries using mcSCRB-seq with a more stringent lysis condition (see Methods; Fig. 3a). We sequenced ~203 million reads for the resulting pool, of which ~189 million passed filtering criteria in the *zUMIs* pipeline (see Methods). Next, we filtered low-quality cells (<50,000 raw reads or mapping rates <75%; Supplementary Fig. 14a), leaving 349 high-quality cells for further analysis (Supplementary Fig. 14b). Using the Seurat package[24], we clustered the expression data and obtained five clusters that could be easily attributed to expected cell types: B cells, Monocytes, NK cells, and T cells (Fig. 3b). Rare cell types, such as dendritic cells or megakaryocytes that are known to occur in PBMCs at frequencies of ~0.5–1%, could not be detected, as expected from the low power to cluster 2–3 cells. For the detected cell types, known marker gene expression fits closely to previously described results[23] (Fig. 3c, d). Overall, we show that mcSCRB-seq is a powerful tool to highlight biological differences, already when a low number of cells are sequenced.

Discussion

In this work, we developed mcSCRB-seq, a scRNA-seq protocol utilizing molecular crowding. Based on benchmarking data generated from mouse ES cells, we show that mcSCRB-seq considerably increases sensitivity and decreases amplification bias due to the addition of PEG 8000 and the use of Terra polymerase, respectively. Furthermore, it shows no indication of bias for GC content and transcript lengths, and has low levels of crosstalk between cell barcodes, which has been seen especially in droplet-based RNA-seq approaches[23,25]. Compared to the previous SCRB-seq protocol, mcSCRB-seq increases the power to quantify gene expression twofold. Additionally, optimized reagents and workflows reduce costs by a factor of three. Qualitatively, we validate our protocol by sequencing PBMCs, a complex mixture of different cell types. We show that mcSCRB-seq can identify the different subpopulations and marker gene expression correctly and distinctively detect the major cell types present in the population.

In this context, we found that it was necessary to use different lysis conditions for the PBMCs than for mESCs. In our experience, some cell types may require a more stringent lysis buffer to stabilize mRNA, which might be a result of internal RNAses and/or lower RNA content. Therefore, we also provide an alternative lysis strategy for mcSCRB-seq to deal with more difficult cell types or samples.

Taken together, mcSCRB-seq is—to the best of our knowledge—not only the most sensitive protocol when benchmarked using ERCCs, it is also the most cost-efficient and flexible plate-based protocol currently available, and could be a valuable methodological addition to many laboratories, in particular as it requires no specialized equipment and reagents.

Methods

cDNA yield assay. For all optimization experiments, universal human reference RNA (UHRR; Agilent) was utilized to exclude biological variability. Unless otherwise noted, 1 ng of UHRR was used as input per replicate. Additionally, Proteinase K digestion and desiccation were not necessary prior to reverse transcription. In order to accommodate all the reagents, the total volume for reverse transcription was increased to 10 µl. All concentrations were kept the same, with the exception that we added the same total amount of reverse transcriptase (25 U), thus lowering the concentration from 12.5 to 2.5 U/µl. After reverse transcription, no pooling was performed, rather preamplification was done per replicate. For each sample, we measured the cDNA concentration using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher).

Comparison of reverse transcriptases. Nine reverse transcriptases, Maxima H- (Thermo Fisher), SMARTScribe (Clontech), Revert Aid (Thermo Fisher), EnzScript (Biozym), ProtoScript II (New England Biolabs), Superscript II (Thermo Fisher), GoScript (Promega), Revert UP II (Biozym), and M-MLV Point Mutant (Promega), were compared to determine which enzyme yielded the most cDNA. Several dilutions ranging from 1 to 1000 pg of universal human reference RNA (UHRR; Agilent) were used as input for the RT reactions.

RT reactions contained final concentrations of 1 × M-MuLV reaction buffer (NEB), 1 mM dNTPs (Thermo Fisher), 1 µM E3V6NEXT barcoded oligo-dT primer (IDT), and 1 µM E5V6NEXT template-switching oligo (IDT). For reverse transcriptases with unknown buffer conditions, the provided proprietary buffers were used. Reverse transcriptases were added for a final amount of 25 U per reaction.

All reactions were amplified using 25 PCR cycles to be able to detect low inputs.

Comparison of template-switching oligos (TSO). Unblocked (IDT) and blocked (Eurogentec) template-switching oligonucleotides were compared to determine yield when reverse transcribing 10 pg UHRR and primer-dimer formation without UHRR input. Reaction conditions for RT and PCR were as described above.

Effect of reaction enhancers. In order to improve the efficiency of the RT, we tested the addition of reaction enhancers, including $MgCl_2$, betaine, trehalose, and polyethylene glycol (PEG 8000). The final reaction volume of 10 µl was maintained by adjusting the volume of H_2O.

For this, we added increasing concentrations of $MgCl_2$ (3, 6, 9, and 12 mM; Sigma-Aldrich) in the RT buffer in the presence or absence of 1 M betaine (Sigma-Aldrich). Furthermore, the addition of 1 M betaine and 0.6 M trehalose (Sigma-Aldrich) was compared to the standard RT protocol. Lastly, increasing concentrations of PEG 8000 (0, 3, 6, 9, 12, and 15% W/V) were also tested.

Comparison of PCR DNA polymerases. The following 12 DNA polymerases were evaluated in preamplification: KAPA HiFi HotStart (KAPA Biosystems), SeqAmp (Clontech), Terra direct (Clontech), Platinum SuperFi (Thermo Fisher), Precisor (Biocat), Advantage2 (Clontech), AccuPrime Taq (Invitrogen), Phusion Flash (Thermo Fisher), AccuStart (QuantaBio), PicoMaxx (Agilent), FideliTaq (Affymetrix), and Q5 (New England Biolabs). For each enzyme, at least three replicates of 1 ng UHRR were reverse transcribed using the optimized molecular crowding reverse transcription in 10 µl reactions. Optimal concentrations for dNTPs, reaction buffer, stabilizers, and enzyme were determined using the manufacturer's recommendations. For all amplification reactions, we used the original SCRB-seq PCR cycling conditions[8].

Cell culture of mouse embryonic stem cells. J1[26] and JM8[27] mouse embryonic stem cells (mESCs) were provided by the Leonhardt lab (LMU Munich) and originally provided by Kerry Tucker (Ruprecht-Karls-University,Heidelberg) and by the European Mouse Mutant Cell repository (JM8A3; www.eummcr.org), respectively. They were used for the comparison of KAPA vs. Terra PCR amplification (Supplementary Fig. 5c) and the comparison of SCRB-seq and mcSCRB-seq, respectively. Both were cultured under feeder-free conditions on gelatine-coated dishes in high-glucose Dulbecco's modified Eagle's medium (Thermo Fisher) supplemented with 15% fetal bovine serum (FBS, Thermo Fisher), 100 U/ml penicillin, 100 µg/ml streptomycin (Thermo Fisher), 2 mM L-glutamine (Thermo Fisher), 1 × MEM non-essential amino acids (NEAA, Thermo Fisher), 0.1 mM β-mercaptoethanol (Thermo Fisher), 1000 U/ml recombinant mouse LIF (Merck Millipore) and 2i (1 µM PD032591 and 3 µM CHIR99021 (Sigma-Aldrich)). mESCs were routinely passaged using 0.25% trypsin (Thermo Fisher).

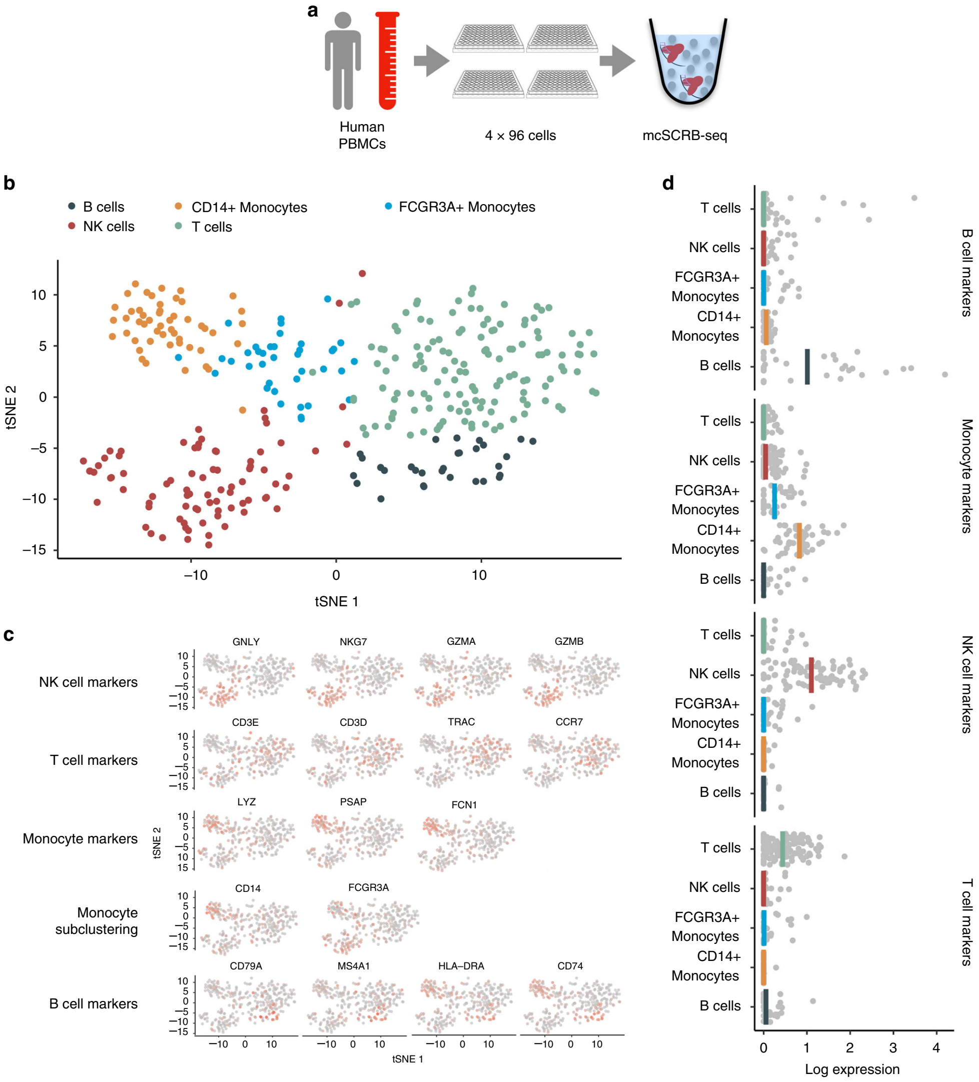


Fig. 3 mcSCRB-seq distinguishes cell types of peripheral blood mononuclear cells. **a** PBMCs were obtained from a healthy male donor and FACS sorted into four 96-well plates. Using the mcSCRB-seq protocol, sequencing libraries were generated. **b** tSNE projection of PBMC cells ($n = 349$) that were grouped into five clusters using the Seurat package[24]. Colors denote cluster identity. **c** tSNE projection of PBMC cells ($n = 349$) where each cell is colored according to its expression level of various marker genes for the indicated cell types. Expression levels were log-normalized using the Seurat package. **d** Marker gene expression from **c** was summarized as the mean log-normalized expression level per cell. B-cell markers: *CD79A*, *CD74*, *MS4A1*, *HLA-DRA*; Monocyte markers: *LYZ*, *PSAP*, *FCN1*, *CD14*, *FCGR3A*; NK-cell markers: *GNLY*, *NKG7*, *GZMA*, *GZMB*; T-cell markers: *CD3E*, *CD3D*, *TRAC*, *CCR7*

mESC cultures were confirmed to be free of mycoplasma contamination by a PCR-based test[28].

Cell culture of human-induced pluripotent stem cells. Human-induced pluripotent stem cells were generated using standard techniques from renal epithelial cells obtained from a healthy donor with written informed consent in accordance with the ethical standards of the responsible committee on human experimentation (216–08, Ethikkommission LMU München) and with the current (2013) version of the Declaration of Helsinki. hiPSCs 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) and 100 U/ml penicillin, 100 µg/ml streptomycin (Thermo Fisher). Cells were routinely passaged using 0.5 mM EDTA. Whenever cells were dissociated into single cells using 0.5 × TrypLE Select (Thermo Fisher), the culture medium was supplemented with 10 µM Rho-associated kinase (ROCK) inhibitor Y27632 (BIOZOL) to prevent apoptosis.

hiPSC cultures were confirmed to be free of mycoplasma contamination by a PCR-based test[28].

SCRB-seq cDNA synthesis. Cells were dissociated using trypsin and resuspended in 100 µl of RNAprotect Cell Reagent (Qiagen) per 100,000 cells. Directly prior to FACS sorting, the cell suspension was diluted with PBS (Gibco). Single cells were sorted into 96-well DNA LoBind plates (Eppendorf) containing lysis buffer using a Sony SH800 sorter (Sony Biotechnology; 100 µm chip) in "Single Cell (3 Drops)" purity. Lysis buffer consisted of a 1:500 dilution of Phusion HF buffer (New England Biolabs). After sorting, plates were spun down and frozen at −80 °C. Libraries were prepared as previously described[6,8]. Briefly, proteins were digested with Proteinase K (Ambion) followed by desiccation to inactivate Proteinase K and reduce the reaction volume. RNA was then reverse transcribed in a 2 µl reaction at 42 °C for 90 min. Unincorporated barcode primers were digested using Exonuclease I (Thermo Fisher). cDNA was pooled using the Clean & Concentrator-5 kit (Zymo Research) and PCR amplified with the KAPA HiFi HotStart polymerase (KAPA Biosystems) in 50 µl reaction volumes.

mcSCRB-seq cDNA synthesis. A full step-by-step protocol for mcSCRB-seq has been deposited in the protocols.io repository[29]. Briefly, cells were dissociated using trypsin and resuspended in PBS. Single cells ("3 drops" purity mode) were sorted into 96-well DNA LoBind plates (Eppendorf) containing 5 µl lysis buffer using a Sony SH800 sorter (Sony Biotechnology; 100 µm chip). Lysis buffer consisted of a 1:500 dilution of Phusion HF buffer (New England Biolabs), 1.25 µg/µl Proteinase K (Clontech), and 0.4 µM barcoded oligo-dT primer (E3V6NEXT, IDT). After sorting, plates were immediately spun down and frozen at −80 °C. For libraries containing ERCCs, 0.1 µl of 1:80,000 dilution of ERCC spike-in Mix 1 was used.

Before library preparation, proteins were digested by incubation at 50 °C for 10 min. Proteinase K was then heat inactivated for 10 min at 80 °C. Next, 5 µl reverse transcription master mix consisting of 20 units Maxima H- enzyme (Thermo Fisher), 2 × Maxima H- Buffer (Thermo Fisher), 2 mM each dNTPs (Thermo Fisher), 4 µM template-switching oligo (IDT), and 15% PEG 8000 (Sigma-Aldrich) was dispensed per well. cDNA synthesis and template switching was performed for 90 min at 42 °C. Barcoded cDNA was then pooled in 2 ml DNA LoBind tubes (Eppendorf) and cleaned up using SPRI beads. Purified cDNA was eluted in 17 µl and residual primers digested with Exonuclease I (Thermo Fisher) for 20 min at 37 °C. After heat inactivation for 10 min at 80 °C, 30 µl PCR master mix consisting of 1.25 U Terra direct polymerase (Clontech) 1.66 × Terra direct buffer and 0.33 µM SINGV6 primer (IDT) was added. PCR was cycled as given: 3 min at 98 °C for initial denaturation followed by 15 cycles of 15 s at 98 °C, 30 s at 65 °C, 4 min at 68 °C. Final elongation was performed for 10 min at 72 °C.

Library preparation. Following preamplification, all samples were purified using SPRI beads at a ratio of 1:0.8 with a final elution in 10 µl of H_2O (Invitrogen). The cDNA was then quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher). Size distributions were checked on high-sensitivity DNA chips (Agilent Bioanalyzer). Samples passing the quantity and quality controls were used to construct Nextera XT libraries from 0.8 ng of preamplified cDNA.

During library PCR, 3′ ends were enriched with a custom P5 primer (P5NEXTPT5, IDT). Libraries were pooled and size-selected using 2% E-Gel Agarose EX Gels (Life Technologies), cut out in the range of 300–800 bp, and extracted using the MinElute Kit (Qiagen) according to manufacturer's recommendations.

Sequencing. Libraries were paired-end sequenced on high output flow cells of an Illumina HiSeq 1500 instrument. Sixteen 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. When several libraries were multiplexed on sequencing lanes, an additional 8 base i7 barcode read was done.

Primary data processing. All raw fastq data were processed using zUMIs together with STAR to efficiently generate expression profiles for barcoded UMI data[14,30]. For UHRR experiments, we mapped to the human reference genome (hg38) while mouse cells were mapped to the mouse genome (mm10) concatenated with the ERCC reference. Gene annotations were obtained from Ensembl (GRCh38.84 or GRCm38.75). Downsampling to fixed numbers of raw sequencing reads per cell were performed using the "-d" option in zUMIs.

Filtering of scRNA-seq libraries. After initial data processing, we filtered cells by excluding doublets and identifying failed libraries. For doublet identification, we plotted distributions of total numbers of detected UMIs per cell, where doublets were readily identifiable as multiples of the major peak.

In order to discard broken cells and failed libraries, spearman rank correlations of expression values were constructed in an all-to-all matrix. We then plotted the distribution of "nearest-neighbor" correlations, i.e., the highest observed correlation value per cell. Here, low-quality libraries had visibly lower correlations than average cells.

Species-mixing experiment. Mouse ES cells (JM8) and human iPS cells were mixed and sorted into a 96-well plate containing lysis buffer as described for mcSCRB-seq using a Sony SH800 sorter (Sony Biotechnology; 100 µm chip). cDNA was synthesized according to the mcSCRB-seq protocol (see above), but without addition of PEG 8000 for half of the plate. Wells containing or lacking PEG were pooled and amplified separately. Sequencing and primary data analysis was performed as described above with the following changes: cDNA reads were mapped against a combined reference genome (hg38 and mm10) and only reads with unique alignments were considered for expression profiling.

Complex tissue analysis. PBMCs were obtained from a healthy male donor with written informed consent in accordance with the ethical standards of the responsible committee on human experimentation (216–08, Ethikkommission LMU München) and with the current (2013) version of the Declaration of Helsinki. Cells were sorted into 96-well plates containing 5 µl lysis buffer using a Sony SH800 sorter (Sony Biotechnology; 100 µm chip). Lysis buffer consisted of 5 M Guanidine hydrochloride (Sigma-Aldrich), 1% 2-mercaptoethanol (Sigma-Aldrich) and a 1:500 dilution of Phusion HF buffer (New England Biolabs). Before library preparation, each well was cleaned up using SPRI beads and resuspended in a mix of 5 µl reverse transcription master mix (see above) and 4 µl ddH_2O. After the addition of 1 µl 2 µM barcoded oligo-dT primer (E3V6NEXT, IDT), cDNA was synthesized according to the mcSCRB-seq protocol (see above). Pooling was performed by adding SPRI bead buffer. Sequencing and primary data analysis was performed as described above using the human reference genome (hg38). We retained only high-quality cells with at least 50,000 reads and a mapping rate above 75%. Furthermore, we discarded potential doublets that contained more than 40,000 UMIs and 5000 genes. Next, we used Seurat[24] to perform normalization (LogNormalize) and scaling. We selected the most variable genes using the "FindVariableGenes" command (1108 genes). Next, we performed dimensionality reduction with PCA and selected components with significant variance using the "JackStraw" algorithm. Statistically significant components were used for shared nearest-neighbor clustering (FindClusters) and tSNE visualization (RunTSNE). Log-normalized expression values were used to plot marker genes.

Estimation of cellular mRNA content. For the estimation of cellular mRNA content in mESCs, we utilized the known total amount of ERCC spike-in molecules added per cell. First, we calculated a detection efficiency as the fraction of detected ERCC molecules by dividing UMI counts to total spiked ERCC molecule counts. Next, dividing the total number of detected cellular UMI counts by the detection efficiency yields the number of estimated total mRNA molecules per cell.

ERCC analysis. In order to estimate sensitivity from ERCC spike-in data, we modeled the probability of detection in relation to the number of spiked molecules. An ERCC transcript was considered detected from 1 UMI. For each cell, we fitted a binomial logistic regression model to the detection of ERCC genes given their input molecule numbers. Using the MASS R-package, we determined the molecule number necessary for 50% detection probability.

For public data from Svensson et al.[5], we used their published molecular abundances calculated using the same logistic regression model obtained from Supplementary Table 2 (https://www.nature.com/nmeth/journal/v14/n4/extref/nmeth.4220-S3.csv). For Quartz-seq2[17], we obtained expression values for ERCCs from Gene Expression Omnibus (GEO; GSE99866), sample GSM2656466; for Chromium[23] we obtained expression tables from the 10 × Genomics webpage (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/ercc) and for SCRB-seq, Smart-seq2, CEL-seq2/C1, MARS-seq and Smart-seq/C1[6], we obtained count tables from GEO (GSE75790). For these methods, we calculated molecular detection limits given their published ERCC dilution factors.

Power simulations. For power simulation studies, we used the powsimR package[22]. Parameter estimation of the negative binomial distribution was done using scran normalized counts at 500,000 raw reads per cell[31]. Next, we simulated two-group comparisons with 10% differentially expressed genes. Log2 fold-changes were drawn from a normal distribution with a mean of 0 and a standard deviation of 1.5. In each of the 25 simulation iterations, we draw equal sample sizes of 24, 48, 96, 192 and 384 cells per group and test for differential expression using ROTS[32] and scran normalization[31].

Batch effect analysis. In order to detect genes differing between batches of one scRNA-seq protocol, data were normalized using scran[31]. Next, we tested for differentially expressed genes using limma-voom[33,34]. Genes were labeled as significantly differentially expressed between batches with Benjamini–Hochberg adjusted p values <0.01.

Code availability. Analysis code to reproduce major analyses can be found at https://github.com/cziegenhain/Bagnoli_2017.

Data availability. RNA-seq data generated here are available at GEO under accession GSE103568.

Further data including cDNA yield of optimization experiments is available on GitHub (https://github.com/cziegenhain/Bagnoli_2017). A detailed step-by-step protocol for mcSCRB-seq has been submitted to the protocols.io repository (mcSCRB-seq protocol 2018). All other data available from the authors upon reasonable request.

Received: 22 December 2017 Accepted: 26 June 2018

Published online: 26 July 2018

References

1. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
2. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
3. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
4. Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I. & Enard, W. Quantitative single-cell transcriptomics. *Brief. Funct. Genomics* https://doi.org/10.1093/bfgp/ely009 (2018).
5. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
6. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 (2017).
7. Menon, V. Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Brief. Funct. Genomics* https://doi.org/10.1093/bfgp/ely001 (2018).
8. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. Preprint at https://doi.org/10.1101/003236 (2014).
9. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).
10. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
11. Zimmerman, S. B. & Pheiffer, B. H. Macromolecular crowding allows blunt-end ligation by DNA ligases from rat liver or *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **80**, 5852–5856 (1983).
12. Rivas, G. & Minton, A. P. Macromolecular crowding in vitro, in vivo, and in between. *Trends Biochem. Sci.* **41**, 970–981 (2016).
13. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
14. 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).
15. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 25533 (2016).
16. Quail, M. A. et al. Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* **9**, 10–11 (2012).
17. Sasagawa, Y. et al. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, 29 (2018).
18. Dixit, A. Correcting chimeric crosstalk in single cell RNA-seq experiments. Preprint at https://doi.org/10.1101/093237 (2016).
19. Baker, S. C. et al. The external RNA controls consortium: a progress report. *Nat. Methods* **2**, 731–734 (2005).
20. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
21. Phipson, B., Zappia, L. & Oshlack, A. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res.* **6**, 595 (2017).
22. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33**, 3486–3488 (2017).
23. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
24. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
25. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. Preprint at https://doi.org/10.1101/303727 (2018).
26. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).
27. Pettitt, S. J. et al. Agouti C57BL/6N embryonic stem cells for mouse genetic resources. *Nat. Methods* **6**, 493–495 (2009).
28. Young, L., Sung, J., Stacey, G. & Masters, J. R. Detection of mycoplasma in cell cultures. *Nat. Protoc.* **5**, 929–934 (2010).
29. Bagnoli, J., Ziegenhain, C., Janjic, A., Wange, L. E. & Vieth, B. mcSCRB-seq protocol. *protocols.io* https://doi.org/10.17504/protocols.io.nrkdd4w (2018).
30. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
31. 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**, 75 (2016).
32. Seyednasrollah, F., Rantanen, K., Jaakkola, P. & Elo, L. L. ROTS: reproducible RNA-seq biomarker detector—prognostic markers for clear cell renal cell cancer. *Nucleic Acids Res.* **44**, e1 (2015).
33. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
34. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

Acknowledgements

We thank Ines Bliesener for expert technical assistance. We are grateful to Magali Soumillon and Tarjei Mikkelsen for providing the original SCRB-seq protocol and to Stefan Krebs and Helmut Blum for sequencing. We would like to thank Elena Winheim for the PBMC sample. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through LMUexcellent and the SFB1243 (Subproject A14/A15).

Author contributions

C.Z. and W.E. conceived the study. J.W.B., C.Z., A.J. and L.E.W. performed experiments and prepared sequencing libraries. J.G. and J.W.B. cultured mouse ES and human iPS cells. Sequencing data were processed by S.P. and C.Z. J.W.B., C.Z., A.J. and B.V. analyzed the data. J.W.B., C.Z., A.J., I.H. and W.E. wrote the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05347-6.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018

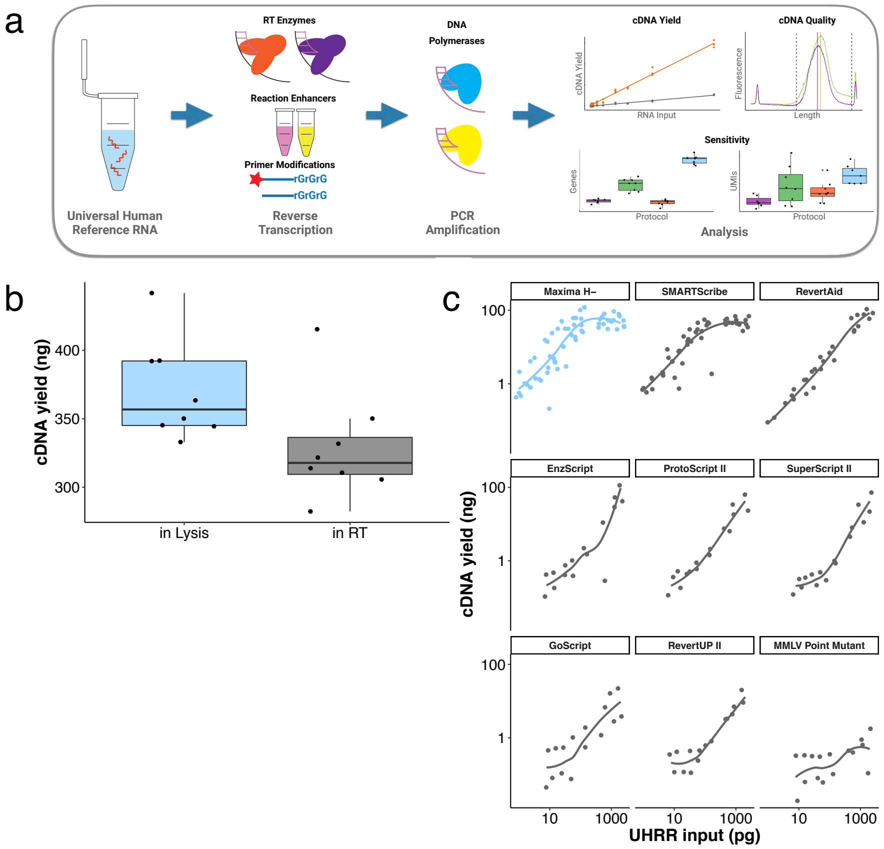
Supplementary Figures and Tables

Supplementary Information

Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq

Bagnoli et al.

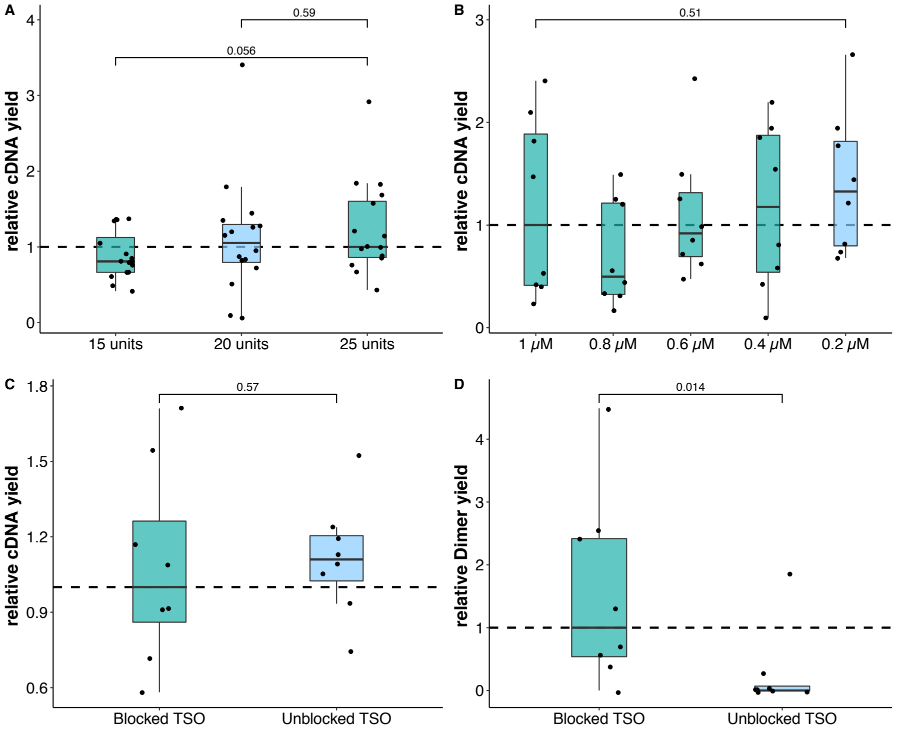
Supplementary Figure 1



Supplementary Figure 1: Schematic overview and optimization of reverse transcription

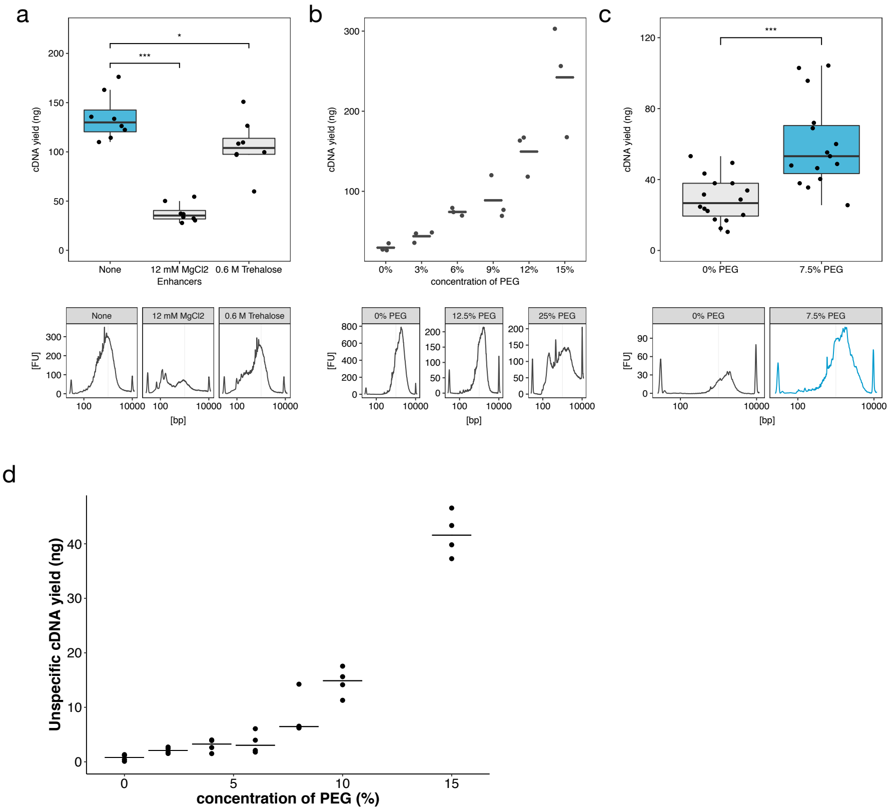
a) Low amounts (1-1000pg) of universal human reference RNA (UHRR) were used in optimization experiments. We assessed components affecting reverse transcription and PCR amplification with respect to cDNA yield and cDNA quality and verified effects on gene and transcript sensitivity by sequencing scRNA-seq libraries to develop the mcSCRB-seq protocol.
b) cDNA yield (ng) after reverse transcription with oligo-dT primers already in the lysis buffer (“in Lysis”) or separately added before reverse transcription (“in RT”). Each dot represents a replicate and each box represents the median and first and third quartiles. The condition selected for the final mcSCRB-seq protocol is highlighted in blue.
c) cDNA yield (ng) dependent on varying UHRR input using 9 different RT enzymes. Each dot represents a replicate. Lines were fitted using local regression. The condition selected for the final mcSCRB-seq protocol is highlighted in blue.

Supplementary Figure 2



Supplementary Figure 2: Optimization of reverse transcription conditions.
Shown are relative cDNA yields after reverse transcription and PCR amplification of UHRR using:
a) varying amounts of reverse transcriptase enzyme (15-25 units, Maxima H-; 1 ng UHRR input per replicate)
b) varying amounts of oligo-dT primer (E3V6; 1 ng UHRR input per replicate)
c) blocked or unblocked Template switching oligo (TSO, E5V6; 10 pg UHRR per replicate)
d) relative primer dimer yield using blocked or unblocked Template switching oligo (TSO, E5V6) estimated using no-input controls (see Methods).
All values are relative to the median of the condition used in the original SCRB-seq protocol[1], which is indicated by a dashed horizontal line. Each dot represents a replicate and each box represents the median and first and third quartiles method. Numbers above boxes indicate p-values (Welch Two Sample t-test).
Optimized conditions selected for the mcSCRB-seq protocol are marked in blue.

Supplementary Figure 3



Supplementary Figure 3: Reverse transcription yield is increased by molecular crowding.

cDNA yield as well as representative length distributions (Bioanalyzer traces, bottom) using various additives in the reverse transcription and template switching reaction.

Each dot represents a replicate, lines represent the median and boxes the first and third quartile. Stars above boxes indicate p-values < 0.05 (Welch Two Sample t-test)

a) Influence of MgCl2 and Trehalose on cDNA synthesis (1 ng UHRR input per replicate; 21 PCR cycles).

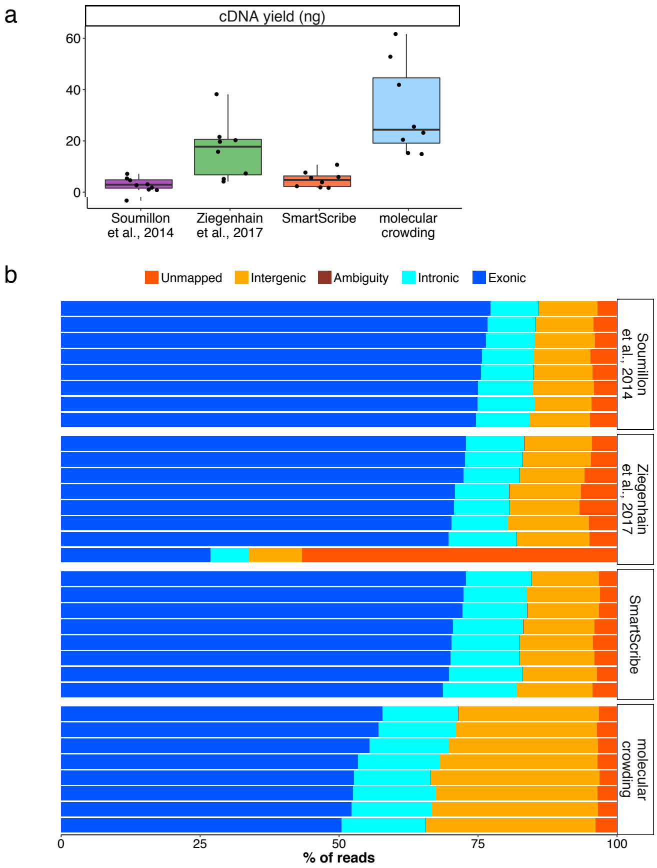
b) Concentration-dependent influence of PEG 8000 on cDNA yield (100 pg UHRR input per replicate; 23 PCR cycles).

c) Effect of 7.5% PEG 8000 (100 pg UHRR input per replicate; 23 PCR cycles).

d) Concentration-dependent generation of unspecific reverse transcription products (0 pg UHRR input per replicate; 23 PCR cycles).

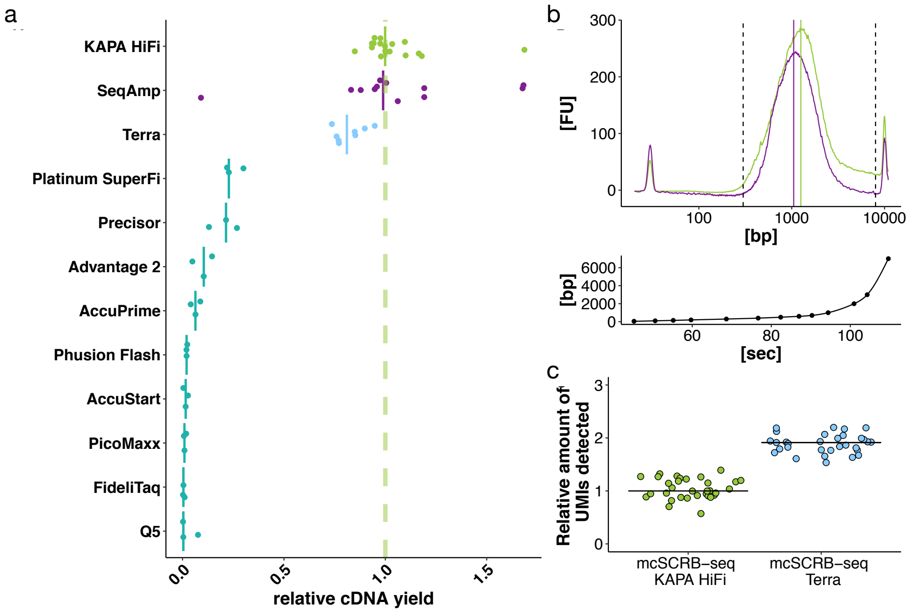
The conditions selected for the final mcSCRB-seq protocol are highlighted in blue.

Supplementary Figure 4



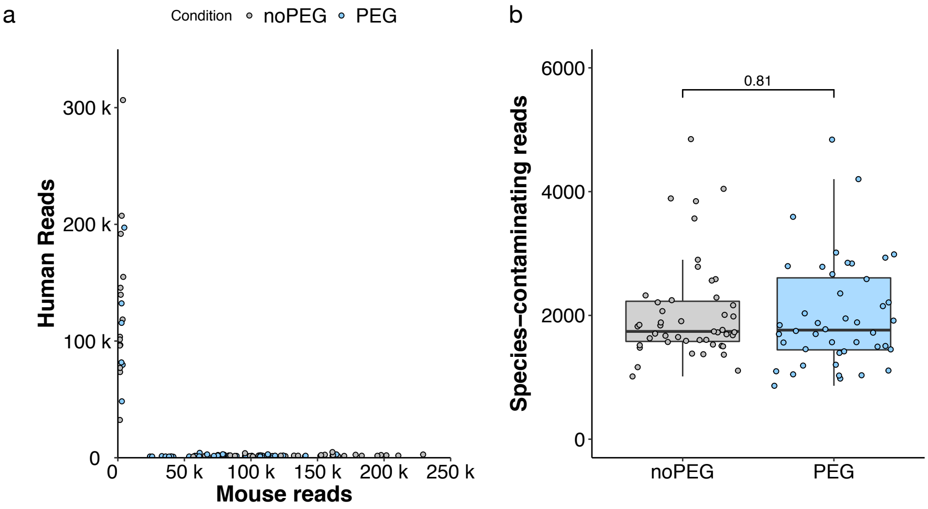
Supplementary Figure 4: Sequencing of UHRR samples.
10 pg of UHRR where used as input for eight replicates for each of the four protocol variants (Supplementary Table 1).
a) cDNA yield (ng) after PCR amplification per method. Each dot represents a replicate and each box represents the median and first and third quartiles per method.
b) Libraries were generated and sequenced from the above cDNA, downsampled to one million reads per library and mapped. Shown are the percentage of sequencing reads that cannot be mapped to the human genome (red), mapped to ambiguous genes (brown), mapped to intergenic regions (orange), inside introns (teal) or inside exons (blue).
Note the higher fraction of reads mapping to intergenic regions, especially in the molecular crowding condition. As UHRR is provided as DNAse-digested RNA, these reads are likely derived from endogenous transcripts, although it is unclear why these are proportionally more detected than annotated transcripts only in the molecular crowding protocol. This is also not generally observed for molecular crowding conditions, as SCRB-seq and mcSCRB-seq protocols have the same fraction (~25%) of intergenic reads mapped when single mouse ES cells are used (Supplementary Figure 7c).

Supplementary Figure 5



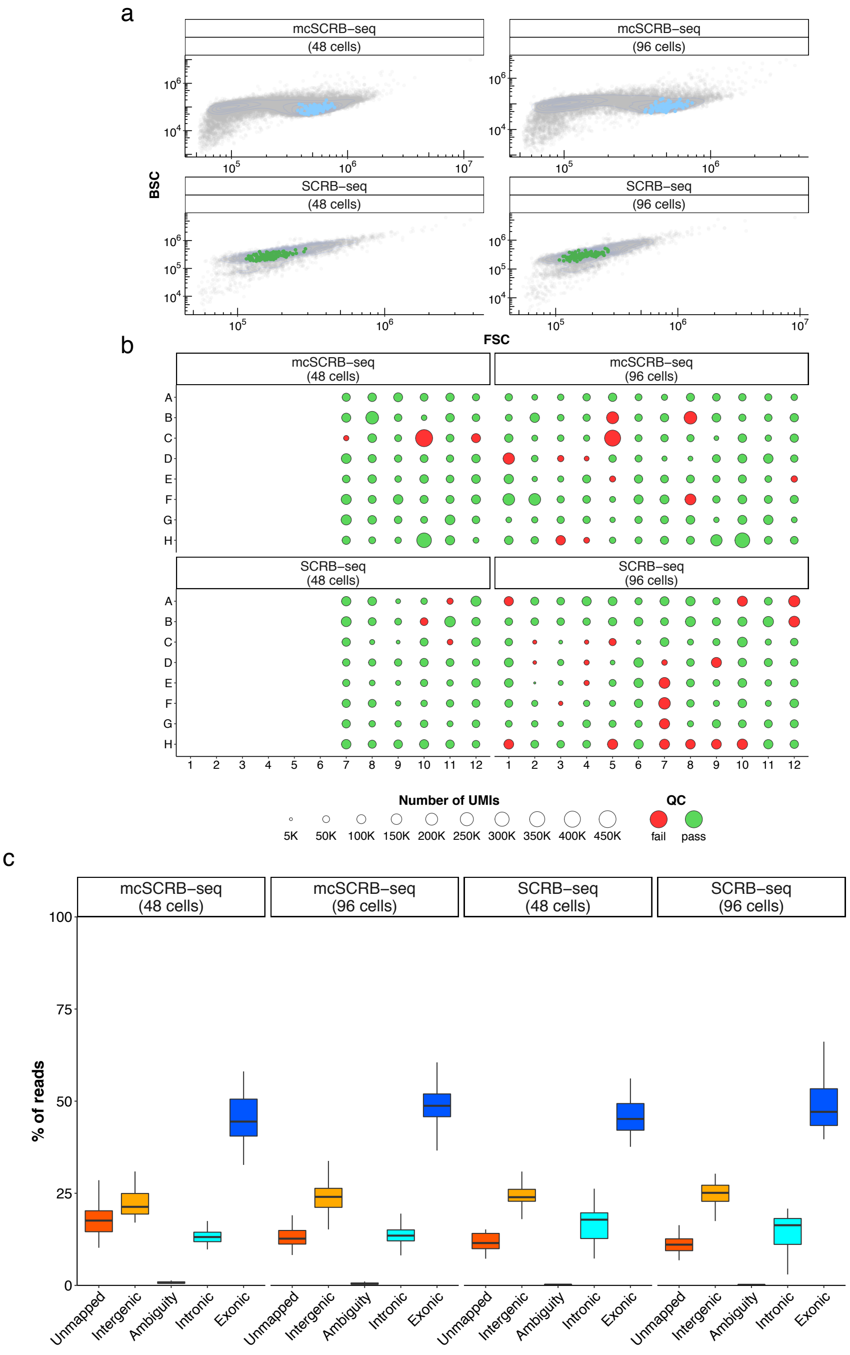
Supplementary Figure 5: Optimization of PCR amplification.
a) Relative cDNA yield after reverse transcription of 1 ng UHRR and amplification using different polymerase enzymes or ready mixes. All values are relative to the median of KAPA HiFi which is indicated by a dashed vertical line, as this was used in the SCRB-seq protocol variant of Ziegenhain et al.[2]. Solid vertical lines indicate the median for each polymerase.
b) Top: Representative length quantification of cDNA libraries amplified with KAPA HiFi (green) or SeqAmp (purple) as quantified by capillary gel electrophoresis (Agilent Bioanalyzer). Solid vertical lines depict the ranked mean length for each library within the region marked with dashed vertical lines. Bottom: Depiction of time length model (spline fit) used to analyze capillary gel electrophoresis via the ladder. Each dot represents a ladder peak with known length (bp) and measurement time (sec).
c) Relative amount of detected UMIs in single mESCs (J1) downsampled to 1 million reads using KAPA-HiFi or Terra for cDNA amplification. For both conditions, molecular crowding conditions (7.5% PEG 8000) were used during reverse transcription. Each dot represents a cell and horizontal lines indicate the median per polymerase.

Supplementary Figure 6



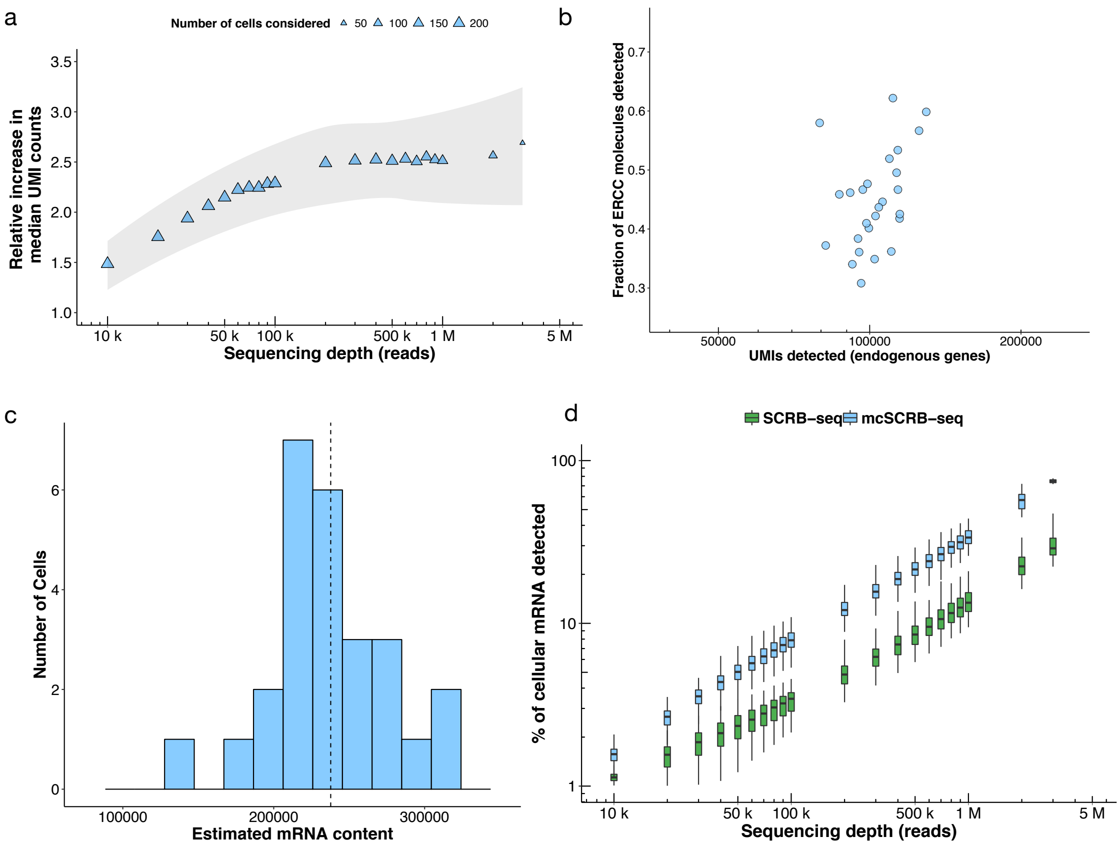
Supplementary Figure 6: Species mixing experiment for mcSCRB-seq
Human induced pluripotent stem cells and Mouse embryonic stem cells were mixed and sorted in a 96-well plate. cDNA was synthesized using the mcSCRB-seq protocol in absence and presence of PEG.
a) For each cell barcode, uniquely aligning reads to human or mouse gene features are shown in a dot plot. No doublets were observed, as expected from single-cell purity FACS sorting.
b) Each cell barcode was classified to be a human or mouse cell. Shown are the number of reads aligning to the wrong species for each of the cell barcodes. There is no significant difference between the protocols with and without PEG (two-sided t-test, p-value=0.81).

Supplementary Figure 7



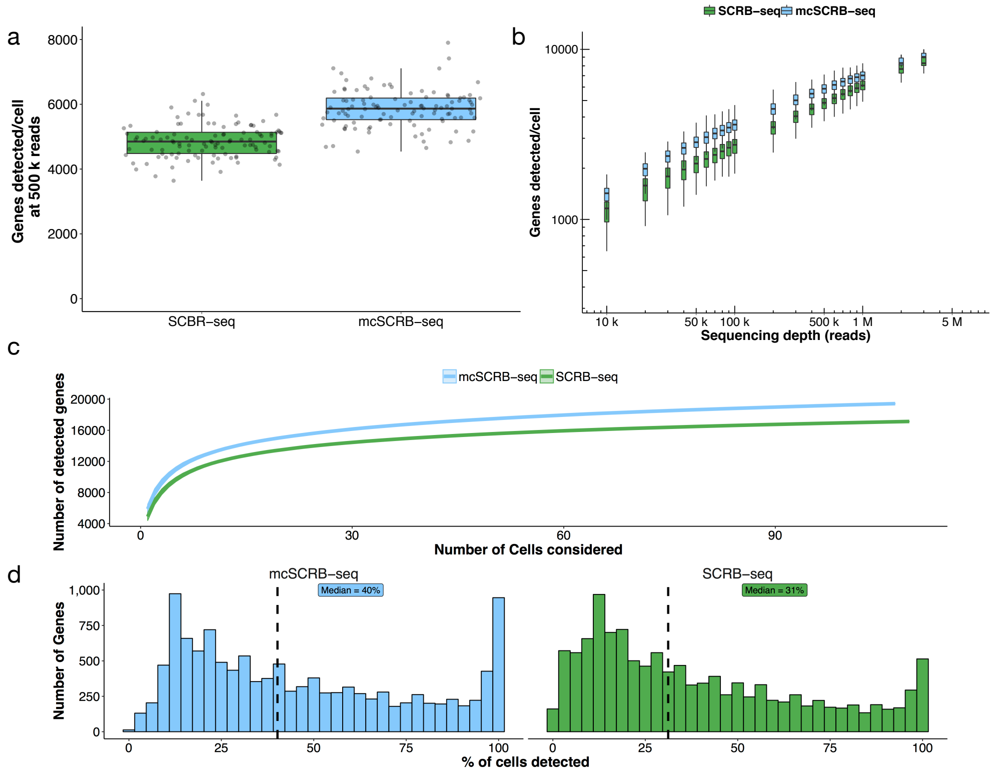
Supplementary Figure 7: Libraries from single mESCs generated with mcSCRB-seq and SCRB-seq protocols.
a) Scatter plots showing FACS data with forward (FS(c) and backward (BS(c) scatter intensities of one vial of mESCs (JM8) resuspended in PBS (mcSCRB-seq) or resuspended in RNAProtect Cell Reagent (SCRB-seq). Each dot represents an event. Coloured dots represent events that were sorted for scRNA-seq libraries in the four plates as depicted in **b**.
b) UMI counts for each cell by method (SCRB-seq/ mcSCRB-seq) and replicate (48 cells/ 96 cells) are shown in their respective position in 96-well plates. Point sizes indicate the number of detected UMIs. Colouring indicates whether a cell passed (green) or failed (red) the Quality Control (QC) as described (see Methods).
c) Percentage of reads that cannot be mapped to the human genome (red), are mapped ambiguously (brown), are mapped to intergenic regions (orange), inside introns (teal) or inside exons (blue). Each box represents the median and first and third quartiles of cells that passed QC for each method.

Supplementary Figure 8



Supplementary Figure 8: Sensitivity of SCRB-seq and mcSCRB-seq protocols.
a) Relative increase in the median of detected UMIs dependent on raw sequencing depth (reads) using mcSCRB-seq compared to SCRB-seq. Each symbol represents the median over all cells at the given sequencing depth. The size of symbols depicts the number of cells (SCRB-seq + mcSCRB-seq) that were considered to calculate the median. The 95% confidence interval of a local regression model is depicted by the shaded area.
b) For each mcSCRB-seq cell that could be downsampled to 2 million reads, the number of UMIs from endogenous genes is plotted on the x axis (median at 102,282 UMIs per cell) and the fraction of UMI- ERCCs from the total amount of spiked-in ERCCs (70,000) is plotted on the y-axis (median 0.49). These values where used to calculate the histogram shown in
c) where for each cell the number of endogenous UMIs is divided by the fraction of ERCCs that were detected in that cell. Using the median of this distribution (dotted line) was set at 100% for the graph in
d) in which the percentage of cellular mRNAs is plotted for each cell at different sequencing depths.

Supplementary Figure 9



Supplementary Figure 9: Sensitivity of SCRB-seq and mcSCRB-seq protocols by genes.

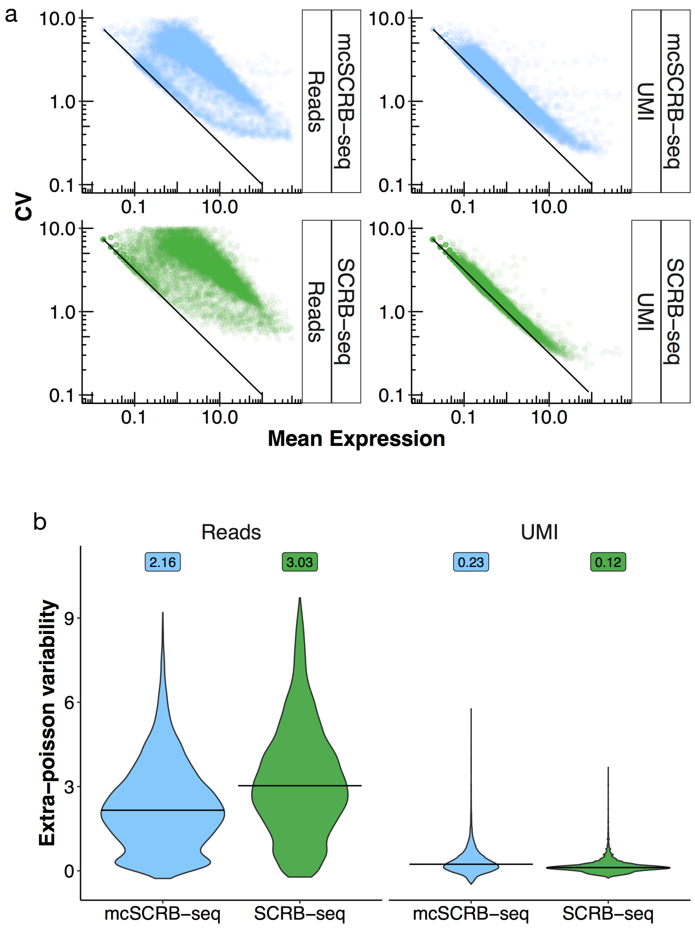
a) Number of detected genes per cell and method (SCRB-seq/mcSCRB-seq) at a sequencing depth of 500,000 reads per cell (downsampled). Each dot represents a cell and each box represents the median and first and third quartiles.

b) Number of detected genes per cell and method (SCRB-seq/mcSCRB-seq) dependent on sequencing depth (reads). Each box represents the median and first and third quartiles per sequencing depth and method. Sequencing depths and genes are plotted on a logarithmic axis (base 10).

c) Number of detected genes at a sequencing depth of 500,000 reads per cell (downsampled) dependent on the number of cells considered.

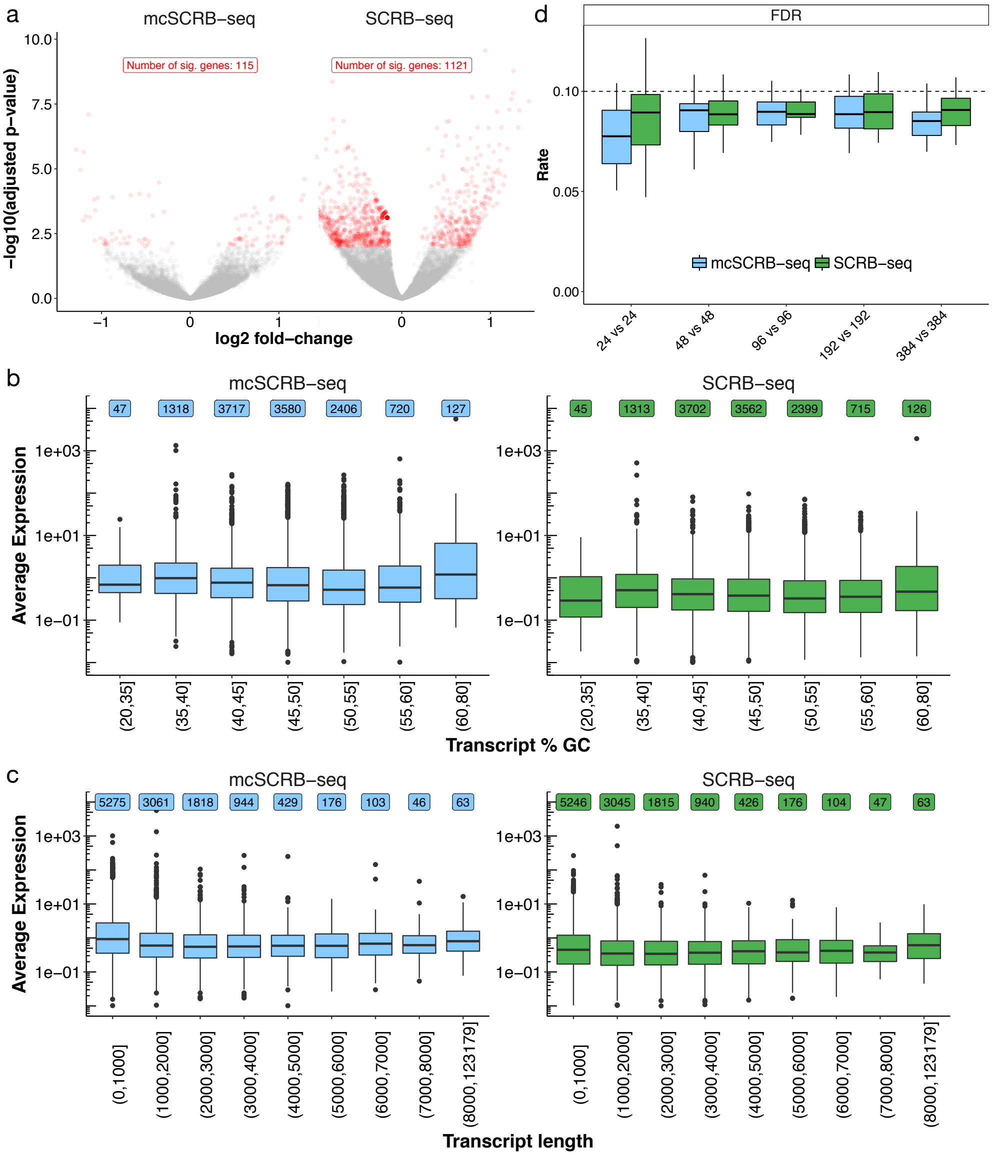
d) Gene detection reproducibility is displayed as the fraction of cells detecting a given gene. Dashed line and label indicate the median of the distribution.

Supplementary Figure 10



Supplementary Figure 10: Variation parameters of SCRB-seq and mcSCRB-seq protocols by genes.
Variation and mean were calculated for each gene and method in cells downsampled to 500,000 reads using either UMIs per gene or reads per gene.
a) Gene-wise mean and coefficient of variation (standard deviation/mean) from all cells are shown as scatterplots for all methods based on read counts or UMIs. The black line indicates variance according to the poisson distribution.
b) Extra-Poisson variability across 12,086 reliably detected genes (detected in > 10% of cells) was calculated by subtracting the expected amount of variation due to Poisson sampling from the coefficient of variation (CV) measured in read-count or UMI quantification. Distributions are shown as violin plots and medians are shown as bars. Numbers indicate the median for each distribution.

Supplementary Figure 11



Supplementary Figure 11: Batch effects, biases and power analysis of SCRB-seq and mcSCRB-seq protocols

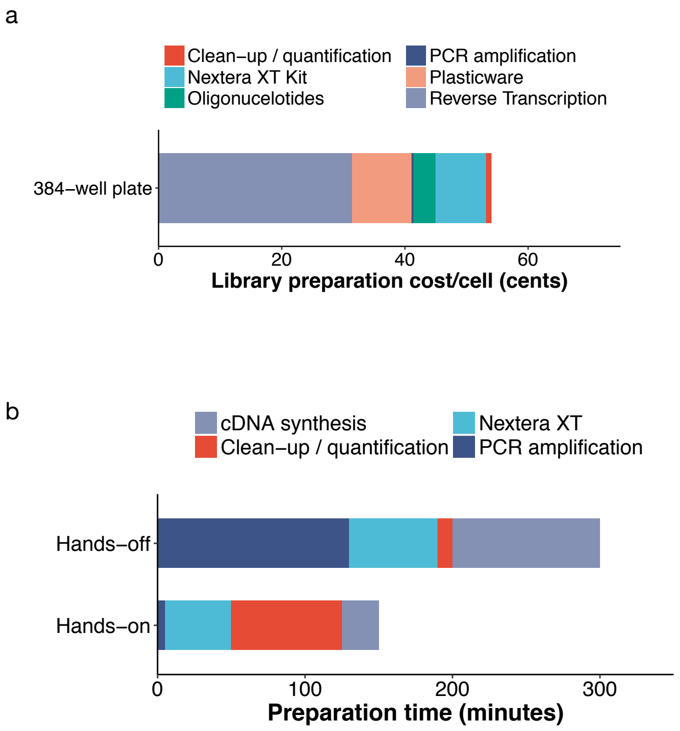
a) Volcano plots show differentially expressed genes between plates for each method. Points in red depict significantly differentially expressed genes (limma-voom; FDR < 0.01). Red labels show the number of differentially expressed genes between batches.

b) Average detected gene-wise expression levels (log normalized UMI) dependent on GC content of each transcript. Transcripts are grouped in 7 bins of GC content. Each dot represents an outlier and each box represents the median and first and third quartiles.

c) Average detected gene-wise expression levels (log normalized UMI) dependent on transcript length. Transcripts lengths are grouped in 7 bins and number of genes in each bin are indicated. Each dot represents an outlier and each box represents the median and first and third quartiles.

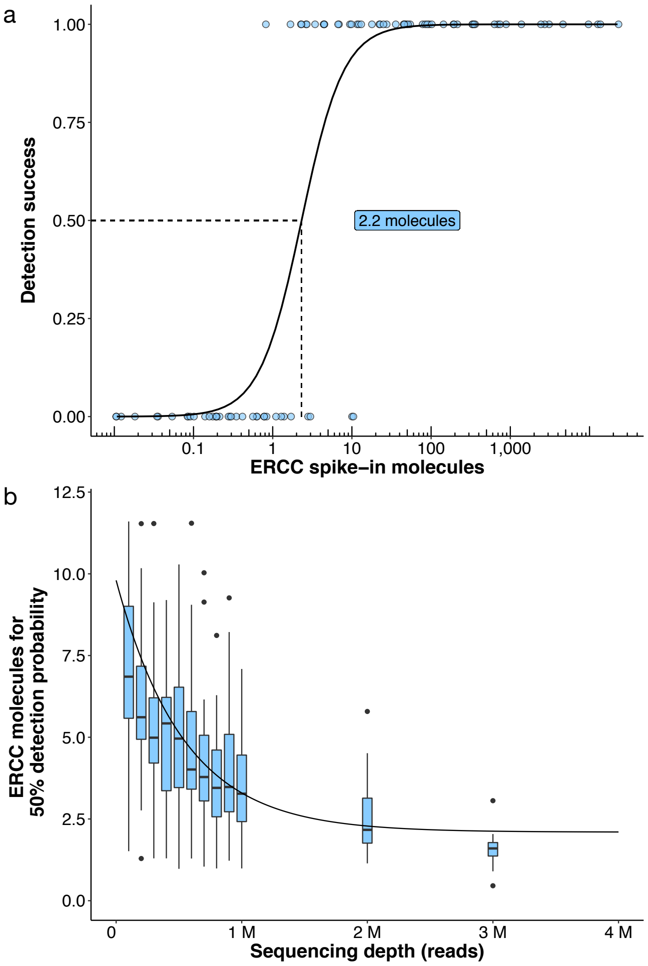
d) Power simulations were performed using the powsimR package[3] from empirical parameters estimated at 500,000 raw reads per cell. For SCRB-seq and mcSCRB-seq, we simulated n-cell two-group differential gene expression experiments with 10% differentially expressed genes. Shown is the false discovery rate (“FDR”) for sample sizes
n = 24, n = 48, n = 96, n = 192 and n = 384 per group. The corresponding true positive rate is shown in Figure 2b. Boxplots represent the median and first and third quartiles of 25 simulations. Dashed lines indicate the desired nominal level.

Supplementary Figure 12



Supplementary Figure 12: Costs and preparation time of mcSCRB-seq
a) Library preparation costs (Eurocents) per cell. Colors indicate the consumable type based on list prices (see Supplementary Table 3). Costs also apply if four 96-well plates are pooled for PCR amplification and Nextera
b) Library preparation time for one 96-well plate of mcSCRB-seq libraries was measured for bench times (“Hands-on”) and incubation times (“Hands-off”). Colors indicate the library preparation step. The total time was 7.5 hours. (see Supplementary Table 4)

Supplementary Figure 13

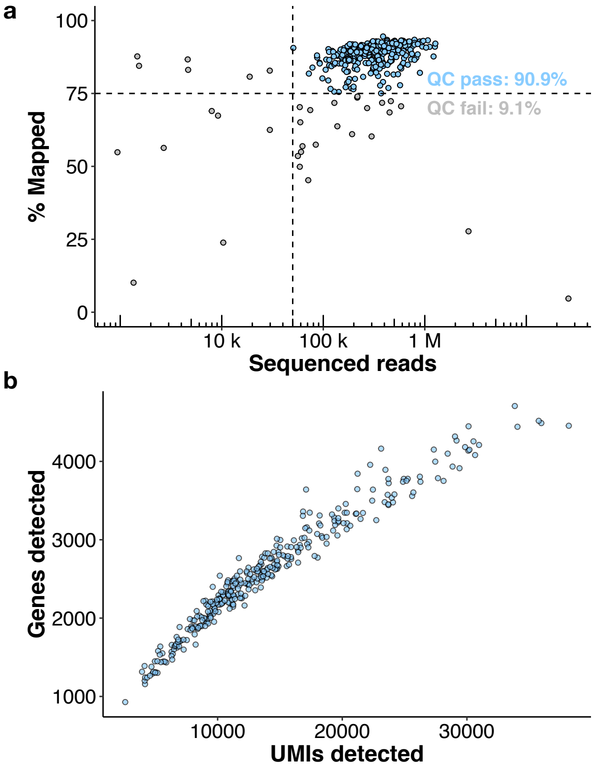


Supplementary Figure 13 : Comparison of mcSCRB-seq to other scRNA-seq data based on ERRCC spike-in detection probability

a) Shown is the detection (0 or 1) of the 92 ERCC transcripts in an average cell processed with mcSCRB-seq at 2 million reads coverage. Points and solid line represent the ERCC genes with their logistic regression model. Dashed lines and label indicate the number of ERCC molecules required for a detection probability of 50%.

b) Number of ERCC molecules required for 50% detection probability dependent on the sequencing depth (reads) for mcSCRB-seq. Each each box represents the median, first and third quartiles of cells per sequencing depth with dots marking outliers. A non-linear asymptotic fit is depicted as a solid black line.

Supplementary Figure 14



Supplementary Figure 14: Quality control of PBMC data

a) Scatter plot shows each of the 384 sequenced PBMC cells with the number of sequenced reads and the % of those reads mapped to the human genome. Dashed lines indicate quality filtering cut-offs chosen. Colors indicate QC passed cells (blue) or discarded cells (grey).

b) Cell-wise detected genes (>=1 UMI) and detected UMIs are shown for all cells that passed quality control (n=349).

Supplementary Table 1

protocol variant	**Soumillon**	**Ziegenhain**	**SmartScribe**	**molecular crowding**
Reverse transcriptase	Maxima H-	Maxima H-	SmartScribe	Maxima H-
Buffer enhancer	none	none	none	7.5% PEG
PCR polymerase	Advantage2	KAPA HiFi	KAPA HiFi	KAPA HiFi

Supplementary Table 1: Overview of used enzymes and enhancers in UHRR based experiments.

Supplementary Table 2

	SCRB-seq	**mcSCRB-seq**
Lysis	Phusion HF	Phusion HF + Proteinase K + oligo-dT primers
Cell suspension	RNAprotect	PBS
Proteinase K	Ambion	Clontech
oligo-dT concentration	1 µM	0.2 µM
reverse transcription volume	2 µl	10 µl
RT amount	25 U	20 U
RT enhancer	none	7.5% PEG
TSO modification	5’-blocking	none
TSO concentration	1 µM	2 µM
Pooling	Zymo Clean & Concentrator	magnetic beads
PCR polymerase	KAPA HiFi	Terra direct
PCR cycles	18-21	13-15
Protocol speed	2 days	1 day
Cost per cell	1-2 €	0.4-0.6 €

Supplementary Table 2: Overview of the key differences between SCRB-seq as used in Ziegenhain et al.[2] and mcSCRB-seq (this work).

Supplementary Table 3

consumable	**price/unit**	**# 384 plates**	**price/384 plate**
Barcode oligo-dT	24.000,00 €	5000	4,80 €
TSO E5V6unblocked	453,40 €	50	9,07 €
Maxima RT	554,00 €	5	110,80 €
Exonuclease I	327,00 €	1000	0,33 €
Clontech Terra	551,00 €	800	0,69 €
Nextera XT	3.002,00 €	96	31,27 €
dNTPs	1.236,00 €	125	9,89 €
Beads	20,00 €	10	2,00 €
Picogreen	542,00 €	400	1,36 €
PCR Seal	500,00 €	1000	0,50 €
PCR Plate/96	140,00 €	0	0,00 €
PCR Plate/384	195,00 €	25	7,80 €
Tips/96	36,50 €	0	0,00 €
Robotic tips/384	290,00 €	10	29,00 €
Total			207,50 €
Total/cell			**0,54 €**

Supplementary Table 3. Detailed overview of costs for mcSCRB-seq.

Supplementary Table 4

Task	**Hands-on (min)**	**Hands-off (min)**	**suggested start time**	**Stopping point?**	**Note**
Prepare workplace	10		09:00		
Proteinase K digest	10	10	09:10		Meanwhile prepare RT Master-Mix
Dispense RT Mix	5		09:30		
RT		90	09:35		
Pool + Clean-up	35	10	11:05	<72h @ 4°C	
ExoI		30	11:50		
PCR set-up	5,00		12:20		
PCR		100	12:25		
PCR clean-up	20,00		14:05	1 week @ 4°C or long-term @ -20 °C	
Quantify cDNA	5,00		14:25		
Nextera: Transposition + PCR set-up	20	10	14:30		
Nextera XT PCR		40	15:00		
PCR clean-up	15,00		15:40	1 week @ 4 °C or long-term @ -20 °C	
Gel-excision & clean-up	25	10	15:55	1 week @ 4 °C or long-term @ -20 °C	
			16:30		
total time	**150**	**300**			

Supplementary Table 4. Detailed overview of hands-on and hands-off time necessary to create a sequenceable mcSCRB-seq library from one single cell plate.

Supplementary References

1. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv* (2014). doi:10.1101/003236

2. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* 65**,** 631–643.e4 (2017)

3. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: Power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx435

2.4 A comparative study of neural differentiation in primates

Johanna Geuder, Zane Kliesmete, Lucas E. Wange, Aleksandar Janjic, Mari Ohnuki,Paulina Spurk, Christopher Alford, Johannes W. Bagnoli, Philipp Janssen, Ines Hellmann, Wolfgang Enard

Unpublished manuscript. Currently only available within this work

Abstract

Studying dynamic processes across the primate phylogeny using transcriptomics can help to better understand the molecular basis of phenotypes and diseases. Conserved gene expression patterns can hint to functional importance and might serve as a basis for further studies, for example by identifying sets of genes or pathways that are essential for the process of interest. Here, we aim to further strengthen this hypothesis by investigating a dynamic process in a cross-species differentiation context. We investigated early neural differentiation in three primate species across six time points using single cell RNA-sequencing. After aligning temporal expression trajectories across species to make them comparable, we identified a set of genes that are consistently and constantly upregulated during differentiation, in all species. We found that this set of genes is significantly enriched for transcription factors that are known to play a role during early neural differentiation. Moreover, the genes with conserved expression upregulation show a higher probability of being mutation intolerant than genes with less conserved patterns and a substantial fraction of these genes are associated with neurodevelopmental disorders. A deeper understanding of the link between regulatory conservation and functional relevance will strengthen the confidence when addressing less commonly investigated cellular processes and help to prioritize particular dysregulated genes in the context of disease.

A comparative study of neural differentiation in primates

Johanna Geuder[1], Zane Kliesmete[1], Lucas E. Wange[1], Aleksandar Janjic[1], Mari Ohnuki[1], Paulina Spurk[1], Christopher Alford[1], Johannes W. Bagnoli[1], Philipp Janssen[1], Ines Hellmann[1], Wolfgang Enard[1+]

[1]Anthropology & Human Genomics, Faculty of Biology, Ludwig-Maximilians-University, Großhaderner Straße 2, 82152 Martinsried, Germany

Abstract

Studying dynamic processes across the primate phylogeny using transcriptomics can help to better understand the molecular basis of phenotypes and diseases. Conserved gene expression patterns can hint to functional importance and might serve as a basis for further studies, for example by identifying sets of genes or pathways that are essential for the process of interest. Here, we aim to further strengthen this hypothesis by investigating a dynamic process in a cross-species differentiation context. We investigated early neural differentiation in three primate species across six timepoints using single cell RNA-sequencing. After aligning temporal expression trajectories across species to make them comparable, we identified a set of genes that are consistently and constantly upregulated during differentiation, in all species. We found that this set of genes is significantly enriched for transcription factors that are known to play a role during early neural differentiation. Moreover, the genes with conserved expression upregulation during cell-state transition towards neural progenitors show a higher probability of being mutation intolerant than genes with less conserved patterns and a substantial fraction of these genes are associated with neurodevelopmental disorders. A deeper understanding of the link between regulatory conservation and functional relevance will strengthen the confidence when addressing less commonly investigated cellular processes and help to prioritize particular dysregulated genes in the context of disease.

Background

Linking genetic to phenotypic changes is an essential, yet a highly challenging task in biology. An evolutionary perspective can shed light on this relationship through the traces that adaptation and constraint have left in different biological systems. This has been shown especially on the DNA level, however the emerging technologies and systems allow to start addressing this question also on a more cellular-system and context specific level, for example expression. Hence, a comparative approach enables us to identify rules of regulatory changes and help to differentiate between important and more spurious links and nodes in regulatory networks (Fair et al., 2020; Wunderlich et al., 2014). However, investigating early embryonic developmental processes in humans and our closest relatives, the non-human primates (NHPs), is challenging, not only because of the difficult acquisition of the primary material, but also because of substantiated ethical concerns. Induced pluripotent stem cells (iPSCs) of a wide range of primates have the potential to bridge this gap, as they can be differentiated in culture to almost any desired cell type and therefore mimic early developmental processes *in vitro* (Blake et al., 2018; Fair et al., 2020; Kanton et al., 2019; Marchetto et al., 2013; Wunderlich et al., 2014).

Here we have utilized previously generated iPSCs of human, gorilla and cynomolgus macaques (Geuder et al., 2021) to investigate the neural induction process using single cell transcriptomics. In order to identify conserved gene expression patterns to infer functional relevance, we analyzed cell compositions, reconstructed the differentiation trajectories and compared these between the species. We found that the cynomolgus cells differentiated at a faster rate than human and gorilla cells. To account for these differences we ordered the cells along a common pseudotime and identified a set of differentially expressed genes between early and late differentiation stages shared across the species. Furthermore, we found that genes that are constantly and consistently upregulated along the differentiation trajectory are significantly enriched for transcriptional regulators, show low tolerance to mutations and a substantial fraction of them are associated with neurodevelopmental disorders. With this approach we demonstrate the usefulness and strength of the evolutionary perspective when investigating dynamic processes to identify relevant patterns through conserved gene expression and regulation.

Results

Study design and data collection of the iPSC-based system

To investigate the similarities and differences of gene expression during neural differentiation between human, gorilla and cynomolgus macaque, we differentiated iPSCs via dual-SMAD inhibition (Chambers et al., 2009; Ohnuki et al., 2014) and sampled at six different timepoints during the course of neural maturation (Figure 1A). To validate the differentiation on the protein level we cryosectioned and stained the spheres for *OCT4* and *PAX6*, markers commonly used to identify pluripotent cells and neural stem cells, respectively. In addition we performed plate based single cell RNA-seq at these six timepoints to investigate the transcriptional landscapes of the different primate iPSCs during this dynamic process.

At each timepoint we obtained data of at least one clone per species. Cynomolgus reads were mapped to mmul10 and human and gorilla reads were mapped to hg38 in accordance with the findings in (Parekh, Vieth, et al., 2018). Possible mapping issues through bad annotations were dealt with by extending the ends and only 1:1 human-macaque orthologs were considered resulting in a total of 9,522 genes and 4,169 cells. We aligned the differentiation trajectories between the species using a pseudo temporal-approach and intersected differentially expressed genes between early and late pseudotime stages between the species and defined conserved patterns of gene expression along the differentiation trajectory from iPSCs to NPCs. Eventually we identify conserved expression patterns between the species and characterize these lists of genes further (Supplementary Figure S1).

Modeling early neural differentiation in human, gorilla and cynomolgus cells

We stained the generated neural spheres at six timepoints (Fig. 1B) using *OCT4* as a pluripotency marker and *PAX6* as a neural marker. Although the observed timing seems to differ between species and clones (Supplementary Figure S2), all samples showed a comparable progression from *OCT4* positive pluripotent cells to *PAX6* positive NPCs, equal fractions of *OCT4/PAX6* positive cells occurred around day 3 in all species, corresponding to the respective RNA expression dynamics derived from single cell RNA-seq data (Supplementary Figure S2). To verify the identity of the cells we classified them using SingleR (Aran et al.,

2019) and the human primary cell atlas (Mabbott et al., 2013) as reference. As expected, cells from day 0 and 1 were classified as pluripotent cells, either as iPSCs or embryonic stem cells. From day 3 on the majority of the cells of each species was classified as either neuroepithelial cells, neurons or astrocytes, although some pluripotent cells could still be detected. Between day 5 and day 9 neural cells could be detected almost exclusively in gorilla and human. The cynomolgus, however, contained a fraction of cells that were neither pluripotent, nor of neural origin (Figure 1C); these 188 cells (4.5% of cynomolgus cells) were excluded from all further analyses, resulting in a total of 3981 cells for further analyses.

Principal component analysis of the clone corrected data shows a clear separation of the different sampling days for all of the species (Figure 1C). Day 0 and 1 cluster closely together and clearly separate from the further differentiated cells from day 3 to day 9, the former cluster shows a high expression of the pluripotency factor *NANOG*, whereas the latter expresses high levels of *PAX6* in all 3 species (Figure 1E).

When looking at the sampling of the three species separately, it is apparent that human and gorilla cells are continuously distributed along the trajectory, whereas the cynomolgus sampling is sparse for the intermediate stage. A potential explanation is that the cynomolgus cells progress faster into the NPC state.

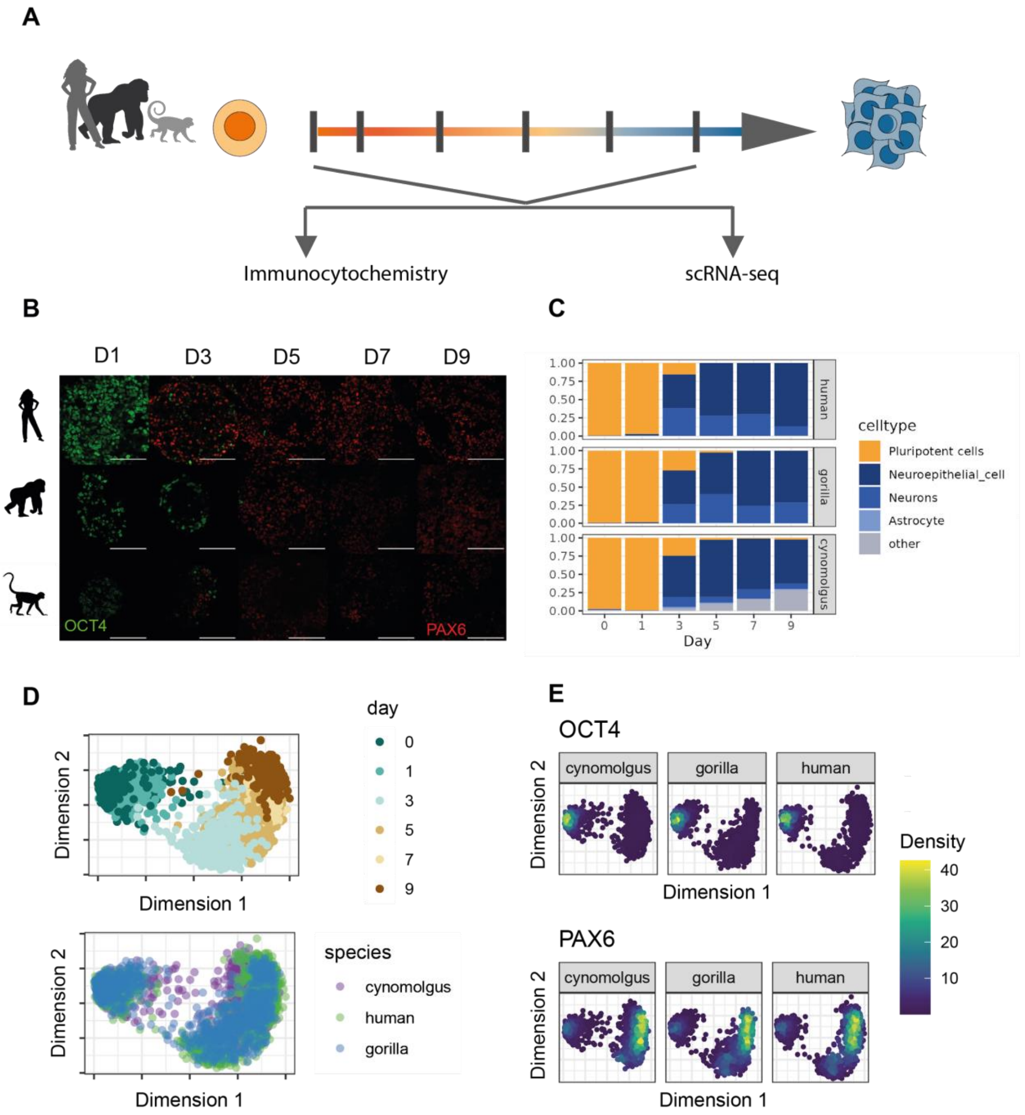


Figure 2: Differentiation validation
A) iPSCs from 3 different species were differentiated to neural precursor cells (NPCs). At six distinct timepoints, cells were sampled for Immunocytochemistry to verify the differentiation to NPCs and for scRNA-seq. B) Immunocytochemistry staining shows that all lines differentiated from pluripotent iPSC (*Oct4* positive) to NPCs (*Pax6* positive) C) Classification of scRNA seq results using singleR (Aran et al., 2019)and HPCA (Mabbott et al., 2013). Some cynomolgus cells differentiated in another or further direction than neural progenitorsD) Clone corrected PCA. iPSCs from day 0 and 1 separate clearly from cells that differentiated already further E) POU5F1 (OCT4) is expressed in the day 0/ day 1 cluster, whereas PAX6 expression is predominantly present in the later stages of differentiation.

Cynomolgus iPSCs differentiate faster than human and gorilla iPSCs

To compare the different cell states and account for potential differences in the differentiation pace, we ordered the cells along pseudotime (pt). Consistent with previous studies (Field et al., 2019) we find the cynomolgus cells to progress faster than human and gorilla cells (Figure 3A). We scaled and binned the pseudotime in three stages (hereby referred to as the stage model) which we call early (containing mostly pluripotent cells, pt: 0 - 0.25) intermediate (containing a mix of pluripotent and neural cells ; pt: >0.25 - 0.75) and late stage (containing mostly neural cells, pt: >0.75 - 1) (Figure 3A). While on day 0 and 1 almost all cells are classified as pluripotent cells and fall in the early stage, day 3 cells of human and gorilla are a heterogeneous mix of pluripotent and neural cells and therefore placed in the intermediate stage whereas the cynomolgus cells seem to have progressed further already (72% of day 3 cells have a pseudotime score higher than 0.75, corresponding to the late stage). Day 5 cells of all species have progressed further and are placed between the boundaries of intermediate and late stage. Day 7 and 9 of all species however, predominantly contain neural cells that fall in the late pseudotime stage (Figure 3A). Figure 3B shows that the sampling days correspond well with the pseudotime scores and that the inferred trajectory follows the sampling days as expected. Figure 3C shows genes that are differentially expressed (DE) between late and early stage cells in all species. For a fair comparison we downsampled the cells per stage to the same number between the species (i.e. early stage contains 378 of each species; see methods). We observe a marginal statistical power of 80% to detect DE genes for all comparisons in every species with our experimental setup (Supplementary Figure S3). We also find that constantly downregulated show an overall higher gene expression in comparison to the constantly upregulated genes. The biggest group of DE genes between early and late stage are shared among all species (821 genes). The number of DE genes shared between two species decreases with phylogenetic distance (Figure 3C), i.e. human and gorilla share more DE genes than human and cynomolgus.

To investigate the trajectory of a given gene over pseudotime, we used a mixed effects model (Law et al., 2014). We determined the best fitting model for each gene based on alternative nested models (species; splined pseudotime; species+splined pseudotime; the full model with an interaction term: species + splined pseudotime + species:splined pseudotime) using AIC (Law et al., 2014). The full model fits best for the highest number of genes (6590, compared to 2490 best explained by the model containing species and pseudotime but no interaction term, 521 best explained by the model containing only species as a predictor, 1 best explained by pseudotime only).

We used this model to accurately trace the trajectory of the gene expression (Figure 3D). Figure 3D shows the trajectory of genes that were identified as up-regulated in the early, intermediate or late stage, as identified before.

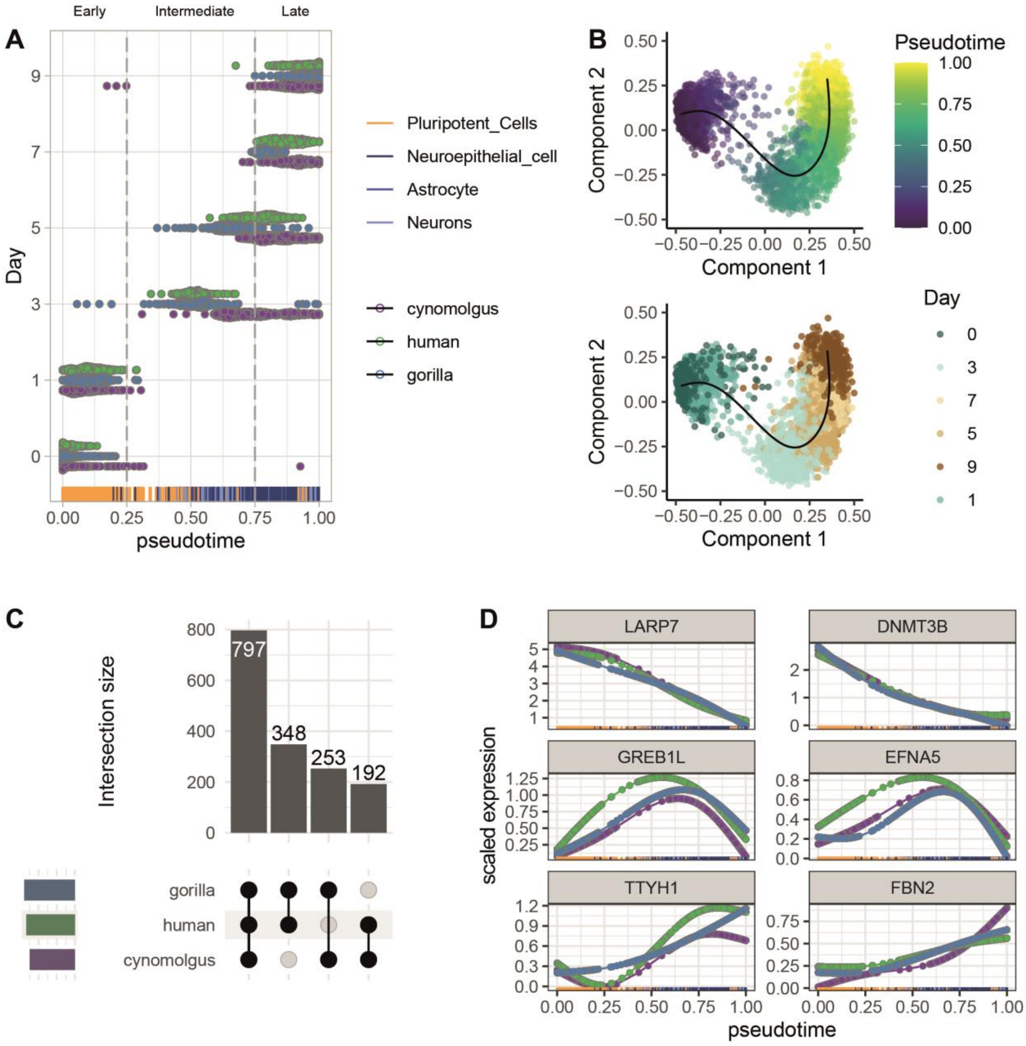


Figure 3: Pseudotime and differentiation trajectories

A) Scorpius pseudotime scores shown against sampling days. A sampling gap between 0.25 and 0.5 pt can be observed in all species. In addition, the cells from cynomolgus seem to be overall faster and show almost no cells in the intermediate stage. B) Principal component analysis and inferred trajectory by Scorpius (Cannoodt et al., 2016). C) Species-overlap of DE genes between late and early stage of differentiation, only genes that change in the same direction are considered as shared between the species. Cells are downsampled to the same number of genes per stage and species. D) Expression progression over pseudotime based on the spline model. Upper panel: genes that are significantly higher in the early stage than in the intermediate and late stage in all species. Middle panel: genes that are significantly higher in the intermediate stage than in the early stage in all species and significantly higher in the intermediate stage than in the late stage for human and gorilla (Cynomolgus was not considered in here because of the sparse sampling in the intermediate stage). Lower panel: genes that are significantly higher in the late stage than in early or intermediate stage in all species.

Constantly rising genes in all species are enriched for transcriptional regulators

We identified genes that are constantly and consistently rising or falling in mean expression over the three differentiation stages and quantified the overlap between the species to determine core neural differentiation factors. We term constantly and consistently regulated genes (CCRGs) . We defined a gene as CCRG if the mean expression in the intermediate and late stages were significantly higher or lower than in the early stage.

Figure 4A shows that 69 genes constantly increase in mean expression in all species across the stages, whereas 352 genes constantly decrease between the three stages in all species. We find pathways to be significantly enriched related to early neural development (Yu & He, 2016) (hypergeometric test, $p\text{-value}<0.1$, terms: Nervous system development, Axon guidance and EPH related pathways, Supplementary Figure S4).

Furthermore the constantly rising genes are enriched for transcriptional regulators (i.e. transcription factors, cofactors, surface proteins, signaling molecules(Obradovic et al., 2022)) (chisq p-value = 7.039e-12), when considering all genes that are differentially expressed between the early and late stage. In contrast, we find less transcriptional regulators in the group of constantly falling genes than expected by chance (chisq p-value =1.554e-08). Investigating the association between the groups (Figure 4A upper panel), constantly falling/rising transcriptional regulators in one species or shared across species, we again find positive residuals in the groups of rising genes. Transcriptional regulators are upregulated during differentiation, and we find an especially high number of transcriptional regulators in the group of rising genes shared across all species. In contrast, all genes in the group with expression decrease over the stages have negative residuals, meaning we see fewer TRs in this group than expected. The group of upregulated genes over all species alone contributes to more than 25% to the total Chi-square score and thus accounts for a large fraction of the difference between expected and observed values.

The 69 constantly rising genes shared across species contain 26 transcriptional regulators of which many are well known to play a role during early neural differentiation supporting the hypothesis that conservation can hint towards functional relevance (Supplementary Figure S5).

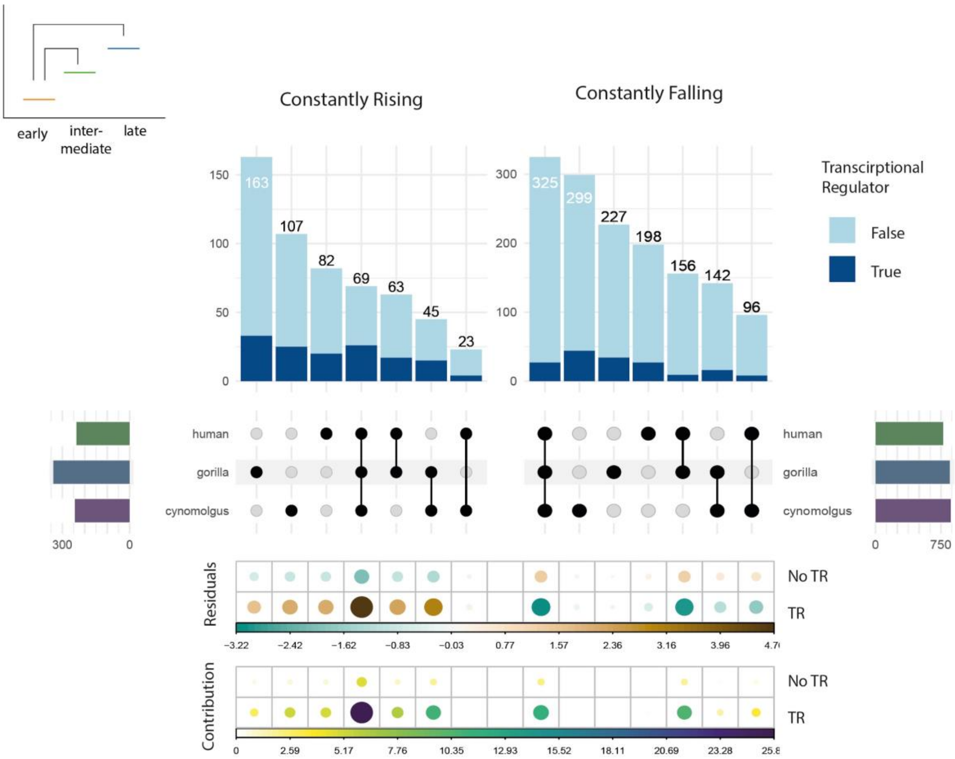


Figure 4: Association between constantly rising/falling genes and transcriptional regulators

Constantly rising/falling genes shared across and specific per species and the fraction of transcriptional regulators for each of the groups. Chisq Residuals and Contributions to the result are displayed below for each of the groups. The genes that are constantly rising in all species show a significant enrichment for transcriptional regulators.

Constantly rising genes with conserved expression patterns have a higher probability of being mutation intolerant and associated with neurodevelopmental disorders

We further assessed the importance and influence of the genes we identified by investigating their probability of being loss of function intolerant (pLI) (Lek et al., 2016) and their association with neurodevelopmental disorders (NDDs) (Leblond et al., 2021). In the case of the constantly rising genes, we see a clear correlation between the number of species that share this pattern and the pLI score (Figure 5A; Pearson's rho = 0.15, p-value = 0.001). The proportion of genes that are very likely intolerant of loss-of-function ($\text{pLI} \geq 0.9$) is highest in the group of genes that is shared across all three species in the constantly rising genes. Whereas the number of genes that are LoF tolerant ($\text{pLI} \leq 0.1$) decreases with increasing number of species that share this expression pattern (Figure 5B). The constantly falling genes do not show an association between LoF tolerance or intolerance with the number of species that share this feature. When we investigate the association of the consistently regulated genes with NDDs we find a similar pattern when looking at the TRs. Constantly upregulated TRs tend to show higher fractions of NDD associated genes than constantly falling genes or genes that are not constantly up- or down-regulated. The more species share the pattern of constantly rising genes, the higher the percentage of TRs that have a known association with NDDs (Figure 5C). On the other hand, non-TR genes do not show this trend, they even show a decrease when comparing shared between two to shared between three species. Since this trend is only present within the group of TRs, the association between the overall number of species in which the genes are constantly rising/falling and being associated with NDDs is not significant (chisq test p=0.1604, Figure 5D).

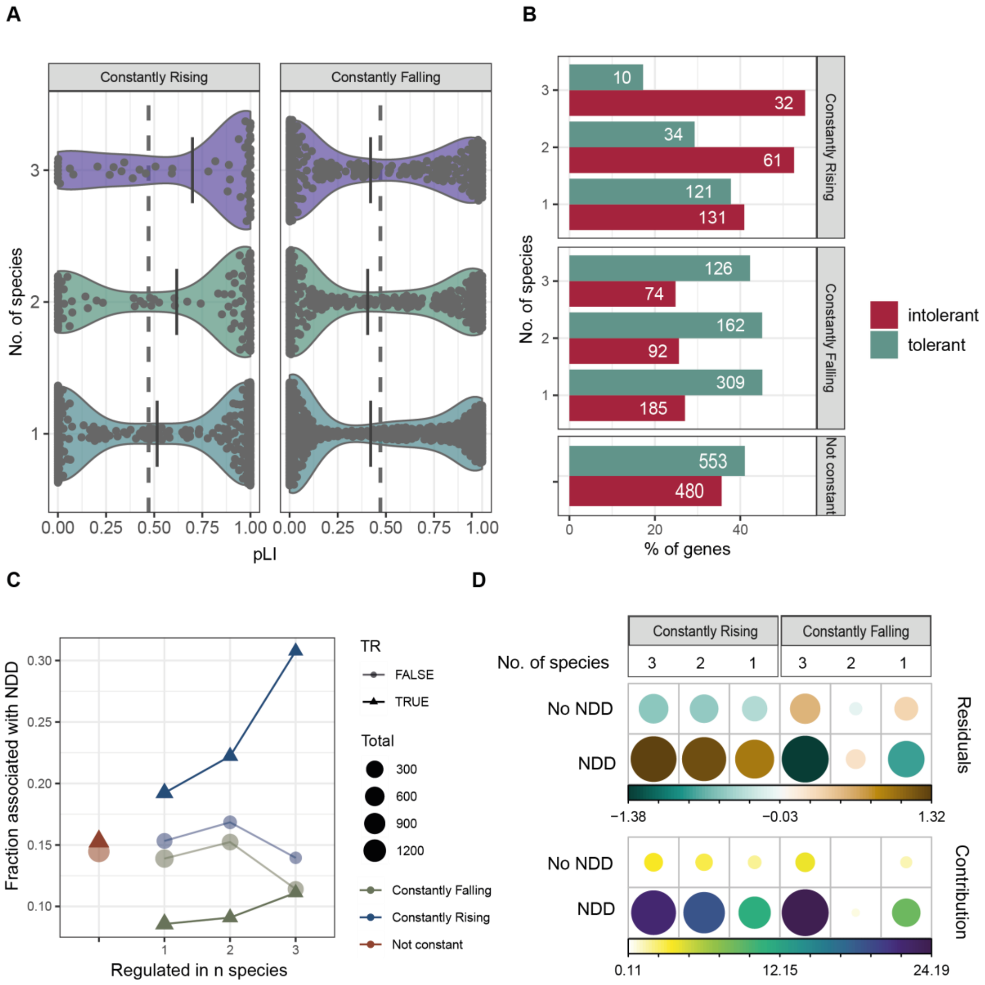


Figure 5: Characterization of conserved genes.

A) pLI scores of genes that are shared between three, two or one species. Vertical line represents the mean of each group. pLI and number of species are significantly correlated for constantly rising genes with a Pearso's rho of 0.15 and a p value of 0.001. No significant correlation is observed across the constantly falling genes. The dashed line represents the mean of genes that are not constantly regulated. B) pLI >= 0.9 was defined as mutation intolerant and <=0.1 as tolerant. In the constantly rising genes the % of genes per group that is intolerant decreases with the number of species it is shared with. The numbers in the bars represent the absolute number of genes. C) Fraction of constantly rising/falling genes shared in all or less species associated with NDDs (Leblond et al., 2021). D) Correlation of constantly rising/falling genes with NDDs.

Discussion

Comparing humans to their closest relatives, the primates, can in many ways help to better understand human evolution, as shown in important studies before (Blake et al., 2018; Field et al., 2019; Kanton et al., 2019). Not only differences between the species can give a valuable insight, but investigating conservation on the transcriptome level during a dynamic process can help us to infer functional relevance on a molecular level.

We here sought to use such a comparative approach by investigating early neural differentiation. We chose a fine grained timeline and sampled single cells at six timepoints during a ten day process of differentiation from iPSCs to NPCs. We ordered the cells along a common pseudotime trajectory, binned them into three stages and modeled the expression trajectory per gene over the time course of this dynamic differentiation process. We identified conserved expression profiles of a set of essential transcriptional regulators (TRs) for NPC differentiation across all three species and underlined the importance of these genes by showing that they have a high probability of being loss of function intolerant and for a large fraction of them an association with NDDs.

One important consideration in this study to note is the small number of cynomolgus cells in the intermediate stage. From our experiment it is not clear if they simply differentiate at a faster pace or if they follow a different differentiation route, skipping the intermediate phase that we observe in human and gorilla. This also led to the necessity to exclude the intermediate vs. late stage comparison in the identification of constantly rising or falling genes. Therefore, we also cannot identify conserved transiently expressed genes in our experiment.

Moreover, when we look at the PAX6 and OCT4 protein in immunocytochemistry and expression levels as classic markers and compare the fractions of positive cells, it seems as if the human cells are the fastest. However, in the course of our analysis we show that these marker genes alone are not sufficient to model the differentiation process. We find the gorilla and human cells to be well comparable, whereas the cynomolgus has progressed further in pseudotime already on day three. The faster differentiation time of macaques was shown in several studies and different differentiation systems before (Field et al., 2019; Jacobo Lopez et al., 2022; Kanton et al., 2019) and we emphasize here that for complex dynamic systems classic marker gene expression alone can not be used to determine differentiation timing.

Nevertheless it could be of interest to investigate if the cynomolgus cells differentiate at a faster pace or if they follow a different differentiation route, skipping the intermediate phase. One possible explanation why they are faster could be the chromatin state of cynomolgus iPSCs, i.e. these could possibly be more primed into the neural differentiation, speeding up the differentiation process. Investigation of the essential genes that we identified for neural differentiation using ATAC-seq data could show if this is the reason for a faster differentiation in macaques. Furthermore, macaques have a shorter gestation time in general. Either way, more timepoints in a shorter time period of the first days of differentiation would be necessary to gain a more fine-grained expression trajectory for macaques.

We identified conserved expression patterns across the three different species during early neural differentiation. We also found a small set of ten genes that show a constant upregulation in one and constant down regulation in another species; these genes could also be of interest. However, we here focus on conservation of expression patterns and did, therefore, not further investigate these genes.

Genes that are constantly upregulated in all species are significantly enriched for transcriptional regulators (TRs) and the majority of these TRs are known to play a role during early neural differentiation from previous studies. Finding many regulators of transcriptional processes during differentiation can be expected and these processes seem to be very conserved across primates. Using this approach we identified, among others, well known and investigated genes like POU3F2, PAX, ZEB1 and ZEB2, which were shown to play a role in early neural differentiation pathways (Inoue et al., 2019; Liu et al., 2018; Shang et al., 2018; Wen et al., 2008), underlining the informational relevance of these conserved expression patterns. In line with this notion, we find the mean pLI score (probability of being loss of function intolerant) is higher in the more species the set of genes is constantly upregulated. We find a significant correlation between the number of species that share this expression pattern and the pLI score, hinting towards functional relevance of these genes during the process of early neural differentiation. In line with this, more than 30% of the constantly rising TRs shared across species are associated with neurodevelopmental disorders (NDDs). We do not find the same trend for the constantly falling genes, we neither find an association between the number of species that share these genes, nor an association with NDDs. However, this can also be expected as the differentiation process is highly specific and genes that are involved, and hence constantly upregulated, must be highly specific for this process, too. Constantly falling genes include all genes that are expressed during the pluripotency stage but should not be expressed anymore in NPCs. As known from previous studies substantially more genes are expressed in the pluripotency stage, when the cell is less specialized (Gulati et al., 2020), so many more genes need to be downregulated during this process than upregulated. Furthermore, genes that are constantly downregulated show an overall higher gene expression and therefore we have more power to detect constantly downregulated than upregulated genes.

Our results show the importance and usefulness of investigating conservation of gene expression during a dynamic process to identify target genes of interest for further studies and ultimately infer functional relevance using this information. This approach could, in a similar way as shown here, help to identify relevant and functional genes in other, less well studied processes Perspectively this list of essential genes for early neural differentiation could be functionally validated using a perturbation screen in the different primate species. The essential genes in all species should have similar impacts on all of the species, the cells might be differentiation deficient, differentiate into a different germ layer or even apoptosis might occur. In contrast, the species specific genes, for example genes that are upregulated over differentiation in the cynomolgus but not human or gorilla, might have an impact on the cynomolgus but not influence, or differently influence, the other species. Similarly one could overexpress these genes in iPSCs and investigate whether they differentiate to NPCs and if there is a difference between shared or specific gene sets.

Methods

Cell culture maintenance and NPC Differentiation

Primate iPSCs were cultured in StemFit + bFGF as described previously (Geuder et al., 2021). For Differentiation 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). Medium was changed every second day.

Immunohistochemistry

Spheres were collected in a reaction tube, washed with PBS and fixed with 4% PFA. After the spheres were incubated in 10%, 20% and 30% sucrose solution, they were embedded in ... and stored at -80°C until further processing. Embedded spheres were then cut using a cryostat to slices a 20 µm. For staining the slides were thawed, rinsed with PBSand incubated with 0.5% TritionX100 for one hour. Primary antibodies in blocking solution were subseqently incubated overnight at 4°C. After washing with 0.05 % TritonX100/PBS at 37°C for 30 minutes, secondary antibody and DAPI were added and incubated overnight. After another 0.05 % TritonX100/PBS treatment at 37°C for 30 minutes, slides were mounted using vectashield and imaged on a confocal microscope.

Library preparation and sequencing

On day 0,1,3,5,7 and 9 cells were dissociated using accumax and single cell sorted using a BD FACS Aria II. The cells were sorted into 96-well plates containing lysis buffer consisting of Phusion buffer, proteinase K and barcoded oligo-dT primers. Libraries were then prepared using mcSCRB-seq (Bagnoli et al., 2018). The method was followed exactly as outlined in the step-by-step protocol (Bagnoli et al., 2018) with the exception of using 17 cycles for the pre-amplification.

Libraries were paired-end sequenced on an Illumina HiSeq 1500 instrument. Sixteen bases were sequenced with the first read to obtain cellular and molecular barcodes and 100 bases were sequenced in the second read into the cDNA fragment

Mapping and Quality Control

Fastq data were processed with zUMIs (Parekh, Ziegenhain, et al., 2018) using bbmap (Bushnell, 2014). The Genomes used for mapping were Homo sapiens HG38, GENCODE release 32 and Macaca mulatta ensembl version 10, release-98. For mapping to Macaca mulatta, an additional flag of *-da* was added to the BBMAP command. Explanations of the usage of these flags can be found at https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/usageguide/. Because many of the cynomolgus reads accumulated downstream of the 3' end of the existing gene model, we added a 3' extension option to zUMIs (Parekh, Ziegenhain, et al., 2018), comparable to ESAT (Derr et al., 2016), to capture reads that fell outside of the gene model using the classic pipeline. Filtering was performed based on distributions of UMIs, Genes, percent of mitochondrial reads after mapping. Distributions of each covariate were visualized and cutoffs manually decided to filter out cells. Counting and quality control was performed separately for each species. After this we restricted the count matrices to 1-to-1 orthologs between human and cynomolgus, combined the data and normalized the read counts of all species together using the package scran (Lun et al., 2016). Orthologous genes were defined as 1-to-1 orthologs per the definition from Ensembl, accessed using biomaRt tool, additionally genes with the same gene name, as pulled from the gtf files were added to the list. Clustering was performed by the quickCluster function prior to normalization, with a min.size of 50, and the option method = "hclust" selected. Log-transformed counts adjusted by size factors were then output for further downstream analysis.

Trajectory Inference

Differences between clones and hence also species were treated as batch effects and were removed prior to pseudotime estimation and clustering. Trajectory Inference was performed with SCORPIUS (Cannoodt et al., 2016). Trajectory inference was performed on counts after filtering cells for quality, normalizing counts with SCRAN, and batch correcting for clone within species with Batchelor (Haghverdi et al., 2018). Dimensionality reduction in SCORPIUS was called with “spearman” and ndim=30.

Differential Gene Expression analysis (stage model)

To compare mean expression differences between the different stages of differentiation, the cells were binned into three stages based on their pseudotime assignment. Boundaries were set at 0.25 and 0.75. For comparisons between the species stage differences we applied an additional filtering, so that the confidence intervals of a gene must overlap between the species in at least one stage. For differential gene expression we used limma-trend (Law et al., 2014; Ritchie et al., 2015) and blocked the clone as a random effect in the model. We defined a gene as rising/falling if the mean expression in the intermediate and late stage was significantly higher/lower than in the early stage.

Spline model for assessing expression over pseudotime

To model gene expression over pseudotime we used a limma-trend model, including the species and splined pseudotime and an interaction term. The knots were set as the boundaries of the stage model to 0.25 and 0.75. Also here we treated the clone as a random effect in the mixed effects model.

Power simulations

A posteriori power analysis was performed using the powsimR package (v. 1.2.4) (Vieth et al., 2017). Mean variance relationships were inferred for the subsampled data for each stage - early, intermediate, late - and species. As we did not observe a difference in this relationship for the different bins, we combined data from all time points per species for power analysis. To test for differences in power to detect DE between the stages we simulated 378 cells for early vs. 114 cells for intermediate stage as well as 215 cells for late according to the real numbers per stage in this study. Secondly, to test for differences in power to detect constantly falling or constantly rising genes, we accounted for the expression levels in the two groups by simulating 7500 genes each, based on the expression of the genes found to follow a comparable temporal pattern in all species. The particular parameters used in powsimR were as follows: fraction of genes simulated as DE, pDE=0.1; LFC cutoff for expression changes to be deemed biologically important, |logFC| >= 0.25; LFC distribution to be sampled from was a gamma distribution form -1.5 to 1.5 with shape=1 and rate=2; number of genes to be tested was 7500. powsimR was run individually per species and expression range (Constantly Rising or Constantly Falling) and marginal power was compared.

Functional relevance comparisons

pLI scores

For the pLI score comparisons we used the gene-wise pLI scores from the Exome Aggregation Consortium (ExAC) (Lek et al., 2016). The scores can be downloaded from: broadinstitute.org/pub/ExAC_release/release1/manuscript_data/forweb_cleaned_exac_r03_march16_z_data_pLI.txt.gz. As described in (Lek et al., 2016) we grouped the probability of a gene for being loss-of-function (LoF) intolerant (pLI) into LoF intolerant (($\text{pLI} \geq 0.9$) and LoF tolerant ($\text{pLI} \leq 0.1$).

Neurodevelopmental Disorder (NDD) association

We extracted high confidence NDD genes (HC-NDD) ($N = 1{,}586$) from (Leblond et al., 2021).

Acknowledgement

We acknowledge the Core Facility Flow Cytometry at the Biomedical Center, Ludwig Maximilian University of Munich for providing equipment and expertise. We are grateful to Stefan Krebs and Helmut Blum for sequencing. Finally, we would like to thank Stefanie Färberböck for all the help in cell culture, Karin Bauer, Ming Zhao, and Ines Bliessener for lab maintenance and assistance.

Availability of data and materials

The generated data can and all scripts to perform data analysis can be obtained upon request. Additionally, the analysis has been briefly outlined in the methods.

Contributions

WE, IH, MO and JG conceived the study. JG and MO cultured, differentiated and performed FACsorting of the cells. PS performed staining experiments. AJ, LEW, JWB and JG generated RNA-seq data. CA, ZK and PJ carried out preprocessing, mapping and QC analyses. IH and ZK provided guidance in data analysis on all steps. LEW carried out power analysis. JG performed DE analysis and species comparisons. JG wrote the manuscript.

Competing interests

The authors declare that they have no competing interests.

Bibliography

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. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, *20*(2), 163–172.

Bagnoli, J. W., Ziegenhain, C., Janjic, A., Wange, L. E., Vieth, B., Parekh, S., Geuder, J., Hellmann, I., & Enard, W. (2018). Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nature Communications*, *9*(1), 2937.

Blake, L. E., Thomas, S. M., Blischak, J. D., Hsiao, C. J., Chavarria, C., Myrthil, M., Gilad, Y., & Pavlovic, B. J. (2018). A comparative study of endoderm differentiation in humans and chimpanzees. *Genome Biology*, *19*(1), 162.

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner* (No. LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/servlets/purl/1241166

Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B., De Preter, K., & Saeys, Y. (2016). SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. In *bioRxiv* (p. 079509). https://doi.org/10.1101/079509

Chambers, S. M., Fasano, C. A., Papapetrou, E. P., Tomishima, M., Sadelain, M., & Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nature Biotechnology*, *27*(3), 275–280.

Derr, A., Yang, C., Zilionis, R., Sergushichev, A., Blodgett, D. M., Redick, S., Bortell, R., Luban, J., Harlan, D. M., Kadener, S., Greiner, D. L., Klein, A., Artyomov, M. N., & Garber, M. (2016). End Sequence Analysis Toolkit (ESAT) expands the extractable information from single-cell RNA-seq data. *Genome Research*, *26*(10), 1397–1410.

Fair, B. J., Blake, L. E., Sarkar, A., Pavlovic, B. J., Cuevas, C., & Gilad, Y. (2020). Gene expression variability in human and chimpanzee populations share common determinants. *eLife*, *9*. https://doi.org/10.7554/eLife.59929

Field, A. R., Jacobs, F. M. J., Fiddes, I. T., Phillips, A. P. R., Reyes-Ortiz, A. M., LaMontagne, E., Whitehead, L., Meng, V., Rosenkrantz, J. L., Olsen, M., Hauessler, M., Katzman, S., Salama, S. R., & Haussler, D. (2019). Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes. *Stem Cell Reports*, *12*(2), 245–257.

Geuder, J., Wange, L. E., Janjic, A., Radmer, J., Janssen, P., Bagnoli, J. W., Müller, S., Kaul, A., Ohnuki, M., & Enard, W. (2021). A non-invasive method to generate induced pluripotent stem cells from primate urine. *Scientific Reports*, *11*(1), 3516.

Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., Cai, S., Zabala, M., Scheeren, F. A., Lobo, N. A., Qian, D., Yu, F. B., Dirbas, F. M., Clarke, M. F., & Newman, A. M. (2020). Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, *367*(6476), 405–411.

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, *36*(5), 421–427.

Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N., & Yosef, N. (2019). Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell*, *25*(5), 713–727.e10.

Jacobo Lopez, A., Kim, S., Qian, X., Rogers, J., Stout, J. T., Thomasy, S. M., La Torre, A., Chen, R., & Moshiri, A. (2022). Retinal organoids derived from rhesus macaque iPSCs undergo accelerated differentiation compared to human stem cells. *Cell Proliferation*, e13198.

Kanton, S., Boyle, M. J., He, Z., Santel, M., Weigert, A., Sanchis-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. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature*, *574*(7778), 418–422.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29.

Leblond, C. S., Le, T.-L., Malesys, S., Cliquet, F., Tabet, A.-C., Delorme, R., Rolland, T., & Bourgeron, T. (2021). Operative list of genes associated with autism and neurodevelopmental disorders based on database review. *Molecular and Cellular Neurosciences*, *113*, 103623.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., ... Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291.

Liu, Y., Yu, C., Daley, T. P., Wang, F., Cao, W. S., Bhate, S., Lin, X., Still, C., 2nd, Liu, H., Zhao, D., Wang, H., Xie, X. S., Ding, S., Wong, W. H., Wernig, M., & Qi, L. S. (2018). CRISPR Activation Screens Systematically Identify Factors that Drive Neuronal Fate and Reprogramming. *Cell Stem Cell*, *23*(5), 758–771.e8.

Lun, A. T. L., McCarthy, D. J., & Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, *5*, 2122.

Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C., & Hume, D. A. (2013). An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics*, *14*, 632.

Marchetto, M. C. N., Narvaiza, I., Denli, A. M., Benner, C., Lazzarini, T. A., Nathanson, J. L., Paquola, A. C. M., Desai, K. N., Herai, R. H., Weitzman, M. D., Yeo, G. W., Muotri, A. R., & Gage, F. H. (2013). Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature*, *503*(7477), 525–529.

Obradovic, A., Vlahos, L., Laise, P., Worley, J., Tan, X., Wang, A., & Califano, A. (2022). PISCES: A pipeline for the Systematic, Protein Activity-based Analysis of Single Cell RNA Sequencing Data. In *bioRxiv* (p. 2021.05.20.445002). https://doi.org/10.1101/2021.05.20.445002

Ohnuki, M., Tanabe, K., Sutou, K., Teramoto, I., Sawamura, Y., Narita, M., Nakamura, M., Tokunaga, Y., Nakamura, M., Watanabe, A., Yamanaka, S., & Takahashi, K. (2014). Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(34), 12426–12431.

Parekh, S., Vieth, B., Ziegenhain, C., Enard, W., & Hellmann, I. (2018). Strategies for quantitative RNA-seq analyses among closely related species. In *bioRxiv* (p. 297408). https://doi.org/10.1101/297408

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2018). zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *GigaScience*, *7*(6), 1–9.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47.

Shang, Z., Chen, D., Wang, Q., Wang, S., Deng, Q., Wu, L., Liu, C., Ding, X., Wang, S., Zhong, J., Zhang, D., Cai, X., Zhu, S., Yang, H., Liu, L., Fink, J. L., Chen, F., Liu, X., Gao, Z., & Xu, X. (2018). Single-cell RNA-seq reveals dynamic transcriptome profiling in human early neural differentiation. *GigaScience*, *7*(11). https://doi.org/10.1093/gigascience/giy117

Vieth, B., Ziegenhain, C., Parekh, S., Enard, W., & Hellmann, I. (2017). powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* , *33*(21), 3486–3488.

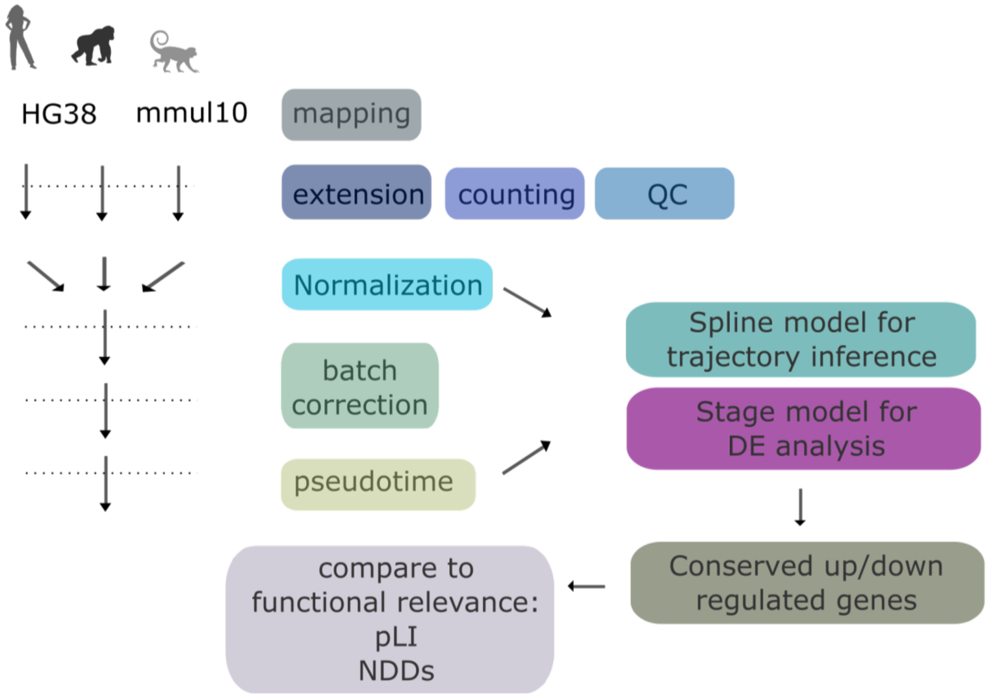
Wen, J., Hu, Q., Li, M., Wang, S., Zhang, L., Chen, Y., & Li, L. (2008). Pax6 directly modulate Sox2 expression in the neural progenitor cells. *Neuroreport*, *19*(4), 413–417.

Wunderlich, S., Kircher, M., Vieth, B., Haase, A., Merkert, S., Beier, J., Göhring, G., Glage, S., Schambach, A., Curnow, E. C., Pääbo, S., Martin, U., & Enard, W. (2014). Primate iPS cells as tools for evolutionary analyses. *Stem Cell Research*, *12*(3), 622–629.

Yu, G., & He, Q.-Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular bioSystems*, *12*(2), 477–479.

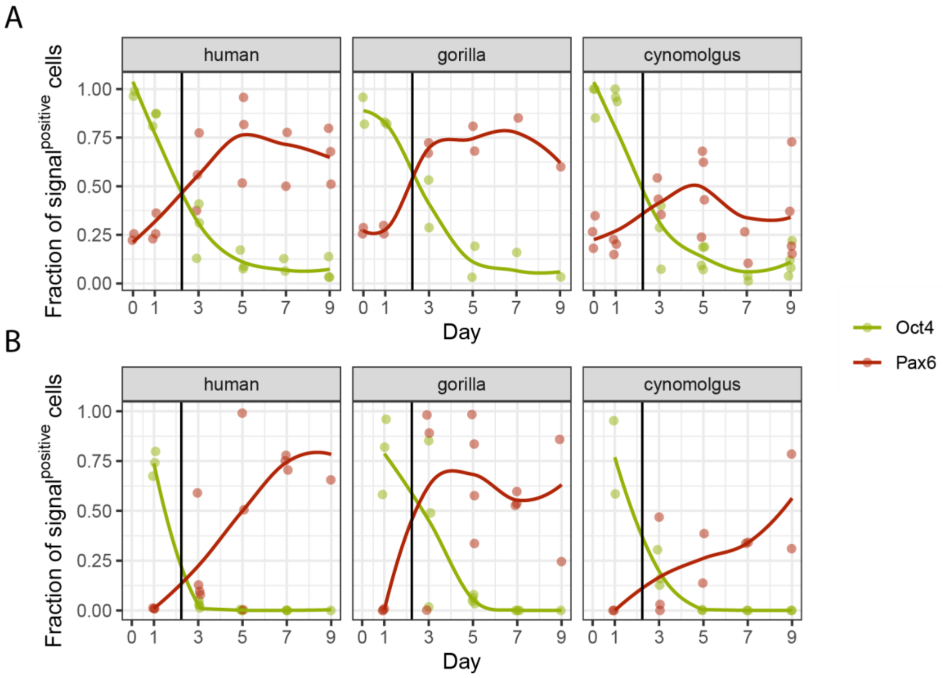
Supplementary Figures and Tables

Supplementary Material

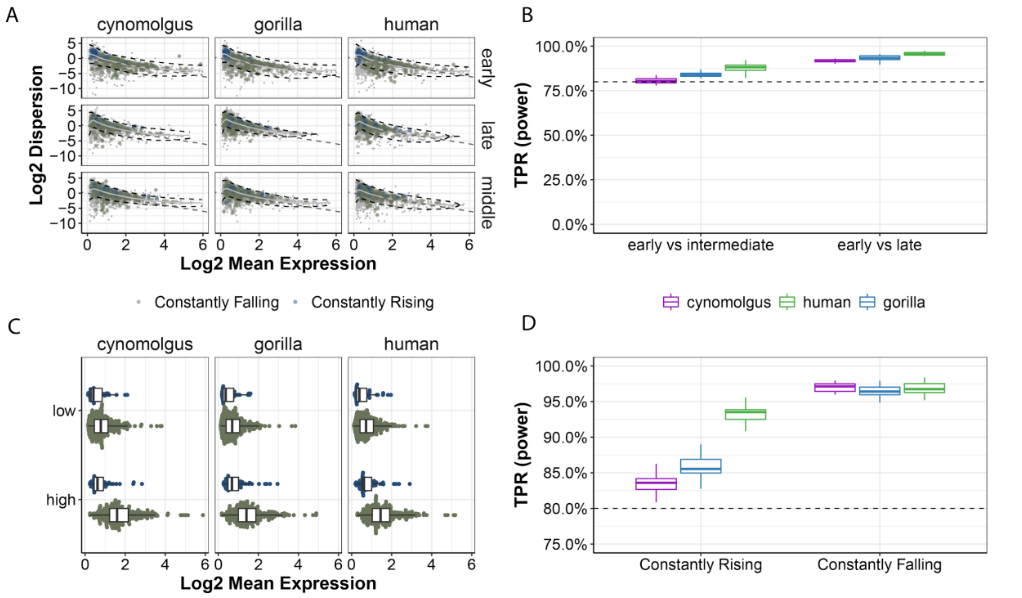

HG38
mmul10
mapping
extension
counting
QC
Normalization
batch correction
pseudotime
Spline model for trajectory inference
Stage model for DE analysis
Conserved up/down regulated genes
compare to functional relevance:
pLI
NDDs

Supplementary Figure 1: Analysis overview

Reads were mapped separately to hg38 in the case of human and gorilla and to the cynomolgus was mapped to mmul10. 3' Extension, counting and QC analysis were performed per species. Normalization using scran was done for all species together and for visualization and pseudotime assignment purposes a batch correction was performed. The normalized counts together with the pseudotime estimation were used for the differential expression analysis, trajectory inference and analysis of conserved expression patterns.

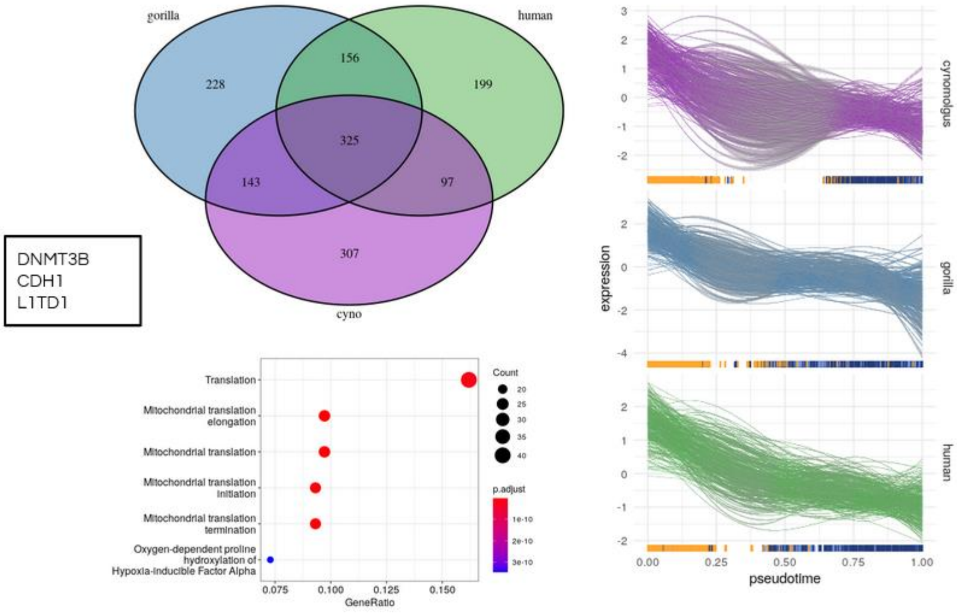


Supplementary Figure 2: Comparison of RNA and Protein levels of OCT4 and PAX6
A) Fraction of cells that express OCT4/PAX6 in RNA seq data (one point per clone) B) Fraction of cells positive for OCT4/PAX6 in staining (2 Clones per species and 2 replicates per clone)

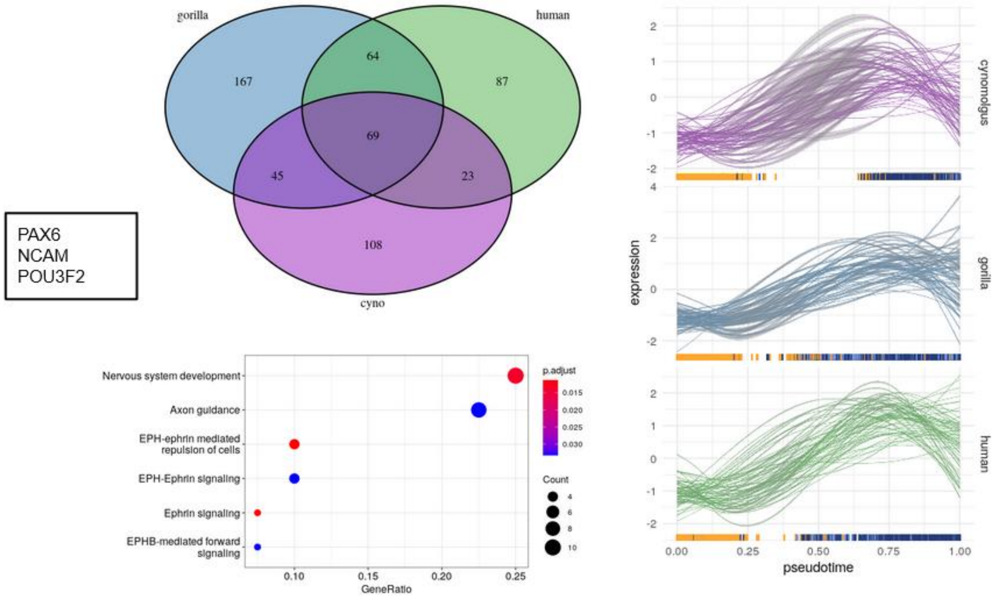


Supplementary Figure 3: Power analysis for the different stages and for the two groups of genes that show constant expression changes in all three species. A) Mean-Variance trends per stage and species. Black dashed lines variance bands as estimated by powsimR; dark gray dashed lines Poison expectation. Points are coloured according to the group of genes they belong to, Constantly Rising/Falling. B) Marginal power (True Positive Rate) per species for comparisons between the different stages. Dashed line marks 80% power C) Mean expression of constantly Rising/Falling genes at their peak stage (high) and at their valley stage (low). D) Marginal power per species for the different sets of genes. Dashed line marks 80% power.

A Constantly falling genes

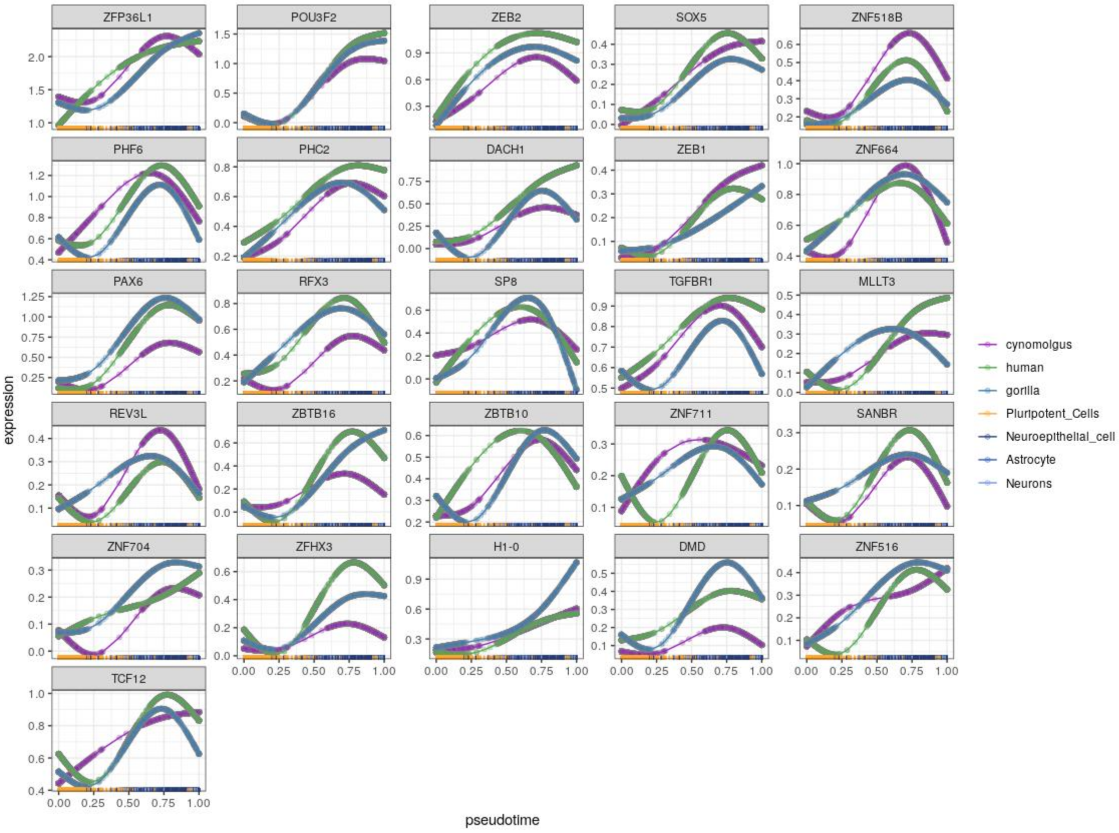


B Constantly rising genes



Supplementary Figure 4: Characteristics of constantly falling and rising genes

Upper panel: Top Reactome Pathways enriched in the constantly falling genes shared across species. lower panel: Top Reactome Pathways enriched in the constantly rising genes shared across species



Supplementary Figure 5: Trajectory of the conserved constantly rising genes
Expression trajectory of 26 Transcriptional regulators that are significantly upregulated in all species over pseudotime.

3 | Discussion

In my thesis I have contributed towards leveraging a comparative molecular approach in primates, by optimizing methods to generate and characterize iPSCs as well as optimizing a scRNA-seq and a bulk RNA-seq method. Further, I combined these technologies to compare gene expression profiles during neural differentiation of primate iPSCs. In the following I will discuss three aspects of this work, namely that bulk RNA-seq is an efficient way to characterize primate iPSCs, aspects on generating iPSCs from more primate species and the prospects of the comparative approach with these resources.

3.1 Bulk RNA sequencing is an efficient way to characterize primate iPSCs

Since the emergence of induced pluripotent stem cells in 2006, researchers have worked on optimizing reprogramming protocols especially focusing on human and mouse cells, as the main model organism. The efficiency and safety of reprogramming has drastically improved during the last decade. An equally important step, the process of verifying the pluripotency state of the cells and the distinction of *bona fide* iPSCs from partially reprogrammed cells (Chan et al. 2009), has not changed that drastically. Although there are some important checkpoints on which the scientific community largely agreed, which assays are actually performed depends on the lab and the subsequent application.

3.1.1 The classic way to characterize pluripotent cells

These strategies involve several steps. A classic first step is the confirmation of the expression of pluripotent cell-specific marker genes via immunocytochemistry or qPCR, often testing the reprogramming factor OCT3/4 or the surface marker TRA-1-60 (Nichols et al. 1998; Andrews et al. 1984). Additionally, the potential to differentiate into cells of all three germ layers needs to be verified. Historically this trilineage differentiation capacity has been tested *in vivo* as a teratoma assay. For that, iPSCs were injected into immune-deficient mice to show the ability to form a teratoma in which cells of the three germ layers can be detected (Nelakanti et al. 2015). However, this strategy is not only very costly and time consuming but also under criticism because of the variability of protocols, for ethical reasons and because its significance is furthermore questionable as the cells are placed in a non-physiological environment (Buta et al. 2013; Vallier et al. 2009; Bouma et al. 2017). A widely used *in vitro* alternative is the embryoid body formation, during which the cells are placed in general differentiation media and differentiate randomly in a 3D structure and cells of the three germlayers can be identified via immunocytochemistry (Höpfl et al. 2004). Both methods perform equally in determining the differentiation potential of the iPSCs, however, only a teratoma assay can give information about the malignant potential of the cell line, which is relevant for clinical applications (Bouma et al. 2017; International Stem Cell Initiative 2018). Moreover, cells can be subjected to a directed differentiation, which requires more specialized protocols and supplements. Of central importance is furthermore the verification of pluripotency at high passage number (at least 50 passages), as well as a karyotype analysis to demonstrate the absence of recurrent numerical or structural aberrations.

3.1.2 Peculiarities of non-human primate iPSCs

All these guidelines in principle also apply for the characterization of non-human primate (NHP) iPSCs. However, as usually when working with non-model organisms, some difficulties and species specific differences can be expected. While many of the classic tools used for human pluripotency testing also work for NHP species, not all of the knowledge from hiPSCs can be easily transferred. For example the gene REX1 was shown to be indispensable for human and mouse pluripotency, but this may not be the case for chimpanzee PSCs (Gallego Romero et al. 2015). While REX1 is not a commonly used marker for pluripotency, this finding still highlights the differences between closely related species and caveats of focusing solely on the expression of single genes when assessing pluripotency in different species. So, while the zoo of available species and clones is growing, it might be highly relevant to standardize procedures and add more efficient/high throughput processes that are applicable to a wider range of primates.

3.1.3 prime-seq is an efficient way to characterize primate iPSCs and their derivatives

Bulk RNA-seq is an ideal method to further characterize iPSCs of different species. With prime-seq we have a highly sensitive and affordable method at hand which is predestined for these types of approaches. First of all, prime-seq does not rely on isolated RNA as input, cells can either be sorted or directly lysed, due to its sensitivity single colonies can simply be picked from the plate and directly be transferred to the tube containing the lysis buffer, making it easy and fast to collect cells at each step during an experiment. Only few cells are needed as input, 1000 cells work well, which is especially interesting when precious samples or rare cell types are to be investigated, for example when picking single colonies or after a differentiation and sorting step. The subsequent bead clean up is not only time and cost-saving compared to the classic RNA isolation with silica membranes, also a high throughput of samples is possible. Overall prime-seq generates high-quality expression profiles at low costs and can be highly beneficial in characterizing iPSCs and their derivatives of different primate species. We show that the cells cluster according to cell type and species and can efficiently and reliably be classified by correlating the expression profiles to reference datasets using methods like SingleR (Aran et al. 2019) and reference datasets containing human embryonic stem cell derived progenitor states (Chu et al. 2016). As previously described (Wunderlich et al. 2014), the expression distance between species is far larger than the distance between individuals, clones or technical differences during the reprogramming process. prime-seq therefore not only poses the advantage of working comparably well for many different species, as shown for samples across 17 species, we furthermore generate comprehensive global gene expression data of the cell lines, valuable information which can can give deeper insights into the cell state and can also be beneficial for subsequent down-stream applications. With the growing body on data on many cell lines like in the HipSci project, characteristics like lineage biases can be derived directly from the expression profiles (Jerber et al. 2021). This valuable information and the knowledge from previous methods like ScoreCard which uses qPCR expression profiles from specific sets of genes to quantify functional pluripotency (Bock et al. 2011; Tsankov et al. 2015), could be integrated with the knowledge we gain from global expression data from different primates and therefore expand the possibilities to reliably and efficiently quantify pluripotency across related species.

Using prime-seq we highlight the usefulness of our generated iPSCs in a straight forward, sensitive and cost as well as time efficient manner. Especially in comparison to commercially available methodologies, utilizing classifiers comparing gene expression profiles, this is a major advantage.

In parallel, we sequenced and analyzed single colony derived primary UDSCs. As these UDSC expression profiles were derived from single colonies, they represent a homogeneous population of cells that can therefore also be efficiently analyzed and classified using prime-seq. We showed that different types of cells can be isolated from urine. All of the three types express pluripotency markers like *KLF4* and *OCT3/4*, explaining their high reprogramming efficiency compared to other primary cells. We also classified the UDSCs with a human microarray reference covering 38 human cell types (Mabbott et al. 2013) and found them to be most similar to either mesenchymal stem cells, epithelial cells or smooth muscle cells respectively, indicating different tissues of origin. These results highlight how prime-seq can help us to better characterize cell states and types, which in turn can provide information on the efficiency of reprogramming and important insights for the comparisons of iPSCs and the cells they are derived from. Large scale characterization of somatic cells and classification based on their reprogramming efficiency can further be used to select for cells with high chances of successful reprogramming. This will be especially important when moving away further from humans and might help to adapt human optimized culture conditions to be more generalizeable.

In the future bulk RNA-seq and especially prime-seq will be a valuable tool in tackling the task of developing streamlined, efficient and straight forward methods to reprogram, characterize and compare primary cells and iPSCs of different species.

3.2 Extension towards more primate species

In recent years many efforts have been made to create large panels of well characterized iPSCs. However, most of these focus on human iPSCs, for example the HipSci consortium (Leha et al. 2016; Streeter et al. 2017) or the STEMBANCC (Cader et al. 2019; Morrison et al. 2015). As primate iPSCs become more important and widely used, it becomes clear that also here a large panel of well characterized and comparable iPSCs is needed, as for example already started by Romero and colleagues who generated a panel of fully characterized chimpanzee (iPSC) lines (Gallego Romero et al. 2015). Especially for evolutionary, cross species analyses it is highly relevant to extend the zoo of species and clones to as many as possible (Kelley and Gilad 2020). Three major steps are important for this venture: a reliable source for the acquisition of somatic cells, an efficient reprogramming procedure and culture conditions that work for a broad spectrum of species, resulting in a panel of comparable iPSCs.

3.2.1 The importance of a non-invasive somatic cell source

In order to increase the number of available species, it is important to consider the strict laws with respect to animal welfare. Therefore, as a very first step in the process it is crucial to establish methods to isolate primary cells in a non-invasive way. To this end, we show that urine, even unsterile urine from the zoo floor, can be used to isolate reprogrammable cells. The only practical drawback is, that compared to other invasive methods the success rate of isolation is rather low due to the few cells that can attach and proliferate in culture (Lang et al. 2013) in combination with contamination issues that come with the unsterile primate urine. We could overcome the problem of contaminated samples to a large extent by the addition of Normocure, a broad-spectrum antibacterial agent, to our cell cultures. We confirmed that the addition of Normocure does not have an impact on the number of colonies that can be isolated from human urine samples and were able to isolate proliferating cells from two orangutan and one gorilla sample collected from the zoo floor. It has to be mentioned that although our method worked for two great ape species, this was not the case for chimpanzee, despite a rather large total sample size, indicating that the process works for some, but not for all species. Luckily, due to the minimal hands on time required for the isolation process and its low costs, it is practically feasible and worthwhile to try isolating UDSCs from urine anytime a sample is available. However, if this method is found not to work for some species or no urine samples are available, alternative opportunities for somatic cell isolation should be used, such as material from health checkups or during surgeries of zoo animals. Nevertheless, the non-invasive sampling of somatic cells from urine allows us to establish iPSCs as a renewable source of cells from various primates, with many advantages and possibilities to expand the zoo of available primate iPSC species.

3.2.2 Human reprogramming factors for non-human primate reprogramming

The classic reprogramming factor cocktails OCT3/4, SOX2, KLF4, and MYC (commonly referred as to OSKM) (Takahashi and Yamanaka 2006) and OCT3/4, SOX2, NANOG and LIN28A (commonly referred as to OSNL) (Yu et al. 2007) have already widely been used to reprogram cells of various taxonomic groups, including many different primates (Endo et al. 2020). The efficiency of human transcription factor sequences for other species is probably attributed to the high degrees of genetic conservation (Endo et al. 2020; Watanabe et al. 2019). However, some species, like marmosets, seem to be hard-to-reprogram and researchers use different strategies to try to solve these problems. Debowski et al. for example utilized a six factor approach with marmoset specific reprogramming factors to overcome this difficulty (Debowski et al. 2015), while other groups tried to enhance the reprogramming efficiency by adding additional factors to the reprogramming cocktail (Tomioka et al. 2010; Watanabe et al. 2019). Importantly, while some species display difficult-to-reprogram characteristics, the field is constantly advancing and we learn more and more about what makes these processes so difficult and how to circumvent these issues. Moreover for a multitude of species, the usage of the classic human reprogramming factors and strategies works efficiently and robustly. Given these successes and the high degrees of conservation of the transcription factor sequences it is likely that human reprogramming factors can be successfully used to generate iPSCs from a wide range of primates.

3.2.3 Finding culture conditions that work for all primates

An additional challenge one is facing when trying to establish comparable iPSCs from different species are the culture conditions. Most protocols focus on the optimization of human iPSCs. While for example human pluripotent stem cells relied on feeder cells for quite some time, this issue was overcome and almost all labs use feeder and xeno-free, defined culture conditions for their standard hiPSC culture by now (Xu et al. 2005a; Xu et al. 2005b; Ludwig et al. 2006; Chen et al. 2011; Nakagawa et al. 2014). Many major advances also have been made in the field of NHP iPSCs in recent years, but not all inventions from hiPSCs were readily applicable to all NHP iPSCs.

The culture of most NHP iPSCs relied on the co-culture with feeder cells or xenogenic medium for longer than hiPSCs (Aron Badin et al. 2019; Hong et al. 2014; Nakai et al. 2018; Navara et al. 2018; Navara et al. 2013). However, as these additional factors complicate downstream applications, limit reproducibility and are very time consuming, more and more work is put into facilitating different solutions. For example, rhesus macaque and marmoset iPSCs were, for a long time only possible to be kept on a feeder cell layer, or in conditioned MEF medium (Yada et al. 2017; Wu et al. 2010), the classic medium components known from feeder-free hiPSC culture seemed to not be sufficient to keep the pluripotent state of these species. In 2020 Stauske and colleagues developed chemically defined conditions for a feeder-free culture of rhesus macaque and baboon by the simultaneous use of Wnt-activation by GSK-inhibition and Wnt-inhibition and called this medium Universal Primate Pluripotent Stem cell (UPPS) medium (Stauske et al. 2020). Other groups describe the usefulness of a defined medium that was previously reported to induce naive human iPSCs and marmoset ESCs (Yoshimatsu et al. 2021; Shiozawa et al. 2020), or a feeder-free culture systems for marmoset using two small molecule inhibitors and customized marmoset iPSC medium (Petkov and Behr 2021).

The diversity of available protocols, aiming to optimize the conditions for NHP iPSCs highlights the difficulties of working with cell cultures of different species. Especially if comparable cells between different species are essential for the success of the experiment and the interpretability of the results, as it is the case for comparative approaches. Nevertheless, the ongoing research and constant improvements are a promising first step towards the establishment of culture conditions that work for many primate species.

And while the somatic cell isolation, reprogramming and feeder-free culture of some species seem to be more challenging, a variety of promising methods have proven to be able to establish comparable cells from different primate species. We contribute to this endeavor with an easy and efficient protocol, which opens the door to more primary material, uses a non-integrating reprogramming approach and has so far been demonstrated to work for human, gorilla and orangutan (Geuder et al. 2021). We compared our iPSCs to a previously reported and well characterized human iPS cell line. This cell line was generated from PBMCs using episomal vectors and subsequently adapted to the same feeder-free culture conditions that we use for our primate cells. Importantly, we found the expression distances between clones are comparable to those between individuals and by far smaller than differences introduced by technical factors like reprogramming method or laboratory the cells were generated in. Furthermore, in an experiment using different reprogramming factor delivery strategies for chimpanzee cells Hemmi et al. reported that no clear association between deficits of iPSC lines and the vectors that were used could be observed (Hemmi et al. 2017). Leading to the conclusion that it might be not so important how the reprogramming factors are introduced into the cell, it is mainly crucial that the cells can be kept in the same medium under the same conditions after acquiring a stable pluripotent state. Also opening up the possibility to reprogram cells of hard-to-reprogram species on feeder cells or in special medium and only after getting a stable cell line adapting all cells to be investigated to universal culture conditions.

3.3 Leveraging the information from comparative approaches

The field of functional comparative genomics is constantly progressing and helping us to better understand human specific traits, diseases and in general the basis of genotype-phenotype relationships (Enard 2012; Housman and Gilad 2020). As pointed out above, iPSCs in combination with scRNA-seq allow to study early development and also provide experimental access to the compared cells for follow up experiments. While iPSCs from human, chimpanzees and other primates have been used e.g. to model brain development and identify human-specific properties (Mora-Bermúdez et al. 2016; Kanton et al. 2019; Pollen et al. 2019), they so far have been conducted in few lines (Kelley and Gilad 2020) of few species and have not leveraged the information from conserved processes. To start establishing quantitative comparisons in this respect, we subjected iPSCs of gorilla, human and cynomolgus to a cross-species differentiation experiment. The cynomolgus iPSCs used in this experiment were established from fibroblasts using the same reprogramming method as described for urinary cells (Geuder et al. 2021). We differentiated the cells via dual-SMAD inhibition, using the same conditions for all species (Chambers et al. 2009; Ohnuki et al. 2014) and studied the transcriptomes during this early neural differentiation process, using scRNA-seq. The cells of nine different clones from three species were sampled at six distinct time points and sequencing libraries were prepared using the mcSCRB-seq protocol. Although the cynomolgus cells progressed faster along pseudotime, we were able to define comparable cell states and compared them between the species. We identified iPSC/NPC specific sets of genes shared across species and show that genes with a conserved constantly rising pattern along the differentiation trajectory are enriched for transcriptional regulators (TRs), among them many known factors important for neural differentiation. Furthermore, these conserved constantly rising TRs show a higher probability of being loss of function intolerant (Lek et al. 2016) and more than 30 % are associated with neurodevelopmental disorders (Leblond et al. 2021), showcasing how conservation of expression patterns during a dynamic differentiation process could help to infer functional relevance on a molecular level.

We concluded that the TRs that show a conserved constant up-regulation across the three species are essential during early neural differentiation. For a further assessment of the functional importance of the TRs as well as their targets in different species, one could imagine a perturbation screen as a next step. These types of screens become possible by the advantages in scRNA-seq technologies in combination with inducible CRISPRa or CRISPRi screens (Jaitin et al. 2016; Dixit et al. 2016; Datlinger et al. 2017). Comparing the difference between a perturbation of TRs which are conserved to species specific TRs in iPSCs of different species could help to better understand the underlying mechanisms, the functionality of these genes and their regulatory networks.

Moreover, to demonstrate the principle that comparing gene expression patterns between closely related species can be used as a means to infer functional relevance, we here used the well studied process of early differentiation via dual-SMAD inhibition. However, leveraging the information of conserved gene expression patterns during any process of interest, might help identifying functional relevant genes and processes or to interpret disease associated changes (Enard 2012). This study might serve as a blueprint for future research investigating expression patterns between species in a dynamic context. Eventually, more clones and available species will facilitate more and more high-throughput experiments with bigger sample sizes over a broad range of processes and this data will ultimately help to better understand the genotype-phenotype relationships among closely related species.

4 | Conclusion and Outlook

In this work focusing on the use of primate iPSCs for evolutionary analyses, I summarized the current state of the art, developed a method to generate primate iPSCs, contributed to method improvements in the field of RNA-seq and finally present a model study demonstrating the usefulness of these methods and the underlying premise.

The field of comparative primate genomics is rapidly evolving. However, one of they keys to successful studies, obtaining comparable cells of a broad spectrum of primates, is still a major hurdle. Although iPSCs in principle are an endless renewable resource, their generation can be challenging. I contributed to unlocking this resource by establishing primate iPSCs from a non-invasive cell source using a footprint free reprogramming approach. A method which might help to expand the zoo of available species, as shown for gorilla and orangutan. This has already opened up opportunities for new research projects in our lab. Furthermore, I contributed to developing a bulk RNA sequencing method, and show that it is a valuable tool to characterize iPSCs of different primates. In addition, I contributed to the systematic optimization of a scRNA sequencing method that can reliably quantify expression levels, which is indispensable for the comparison of expression patterns during dynamic processes of development in an evolutionary framework. Our experiment on a time series of early neural differentiation in primates demonstrates the power of such comparative approaches. By measuring the phenotype of gene expression we were able to infer putative functionally relevant transcriptional regulators from expression conservation across species.

Further optimizations and the availability of more species and clones will enable and facilitate more comparative experiments using a broad spectrum of technologies and in-depth analyses. Future approaches might, for example utilize co-culture of cells of different species, like it has been shown for different human cell lines already, further reducing biases in comparative studies and making it possible to increase the number of species and clones to be studied at once.

The constant progress in the field of functional comparative primate genomics allows us to investigate genotype-phenotype relationships from an evolutionary perspective which will help us to further deepen our understanding of human specific traits and ultimately create a better understanding and functional annotation of the human genome.

Bibliography

Aasen, Trond and Juan Carlos Izpisúa Belmonte (2010). Isolation and cultivation of human keratinocytes from skin or plucked hair for the generation of induced pluripotent stem cells. *Nat. Protoc.* 5.2, 371–382.

Aasen, Trond, Angel Raya, Maria J Barrero, Elena Garreta, Antonella Consiglio, Federico Gonzalez, Rita Vassena, Josipa Bilić, Vladimir Pekarik, Gustavo Tiscornia, Michael Edel, Stéphanie Boué, and Juan Carlos Izpisúa Belmonte (2008). Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nat. Biotechnol.* 26.11, 1276–1284.

Adey, Andrew, Hilary G Morrison, Asan, Xu Xun, Jacob O Kitzman, Emily H Turner, Bethany Stackhouse, Alexandra P MacKenzie, Nicholas C Caruccio, Xiuqing Zhang, and Jay Shendure (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* 11.12, R119.

Al Abbar, Akram, Siew Ching Ngai, Nadine Nograles, Suleiman Yusuf Alhaji, and Syahril Abdullah (2020). Induced Pluripotent Stem Cells: Reprogramming Platforms and Applications in Cell Replacement Therapy. *Biores. Open Access* 9.1, 121–136.

Alföldi, Jessica and Kerstin Lindblad-Toh (2013). Comparative genomics as a tool to understand evolution and disease. *Genome Res.* 23.7, 1063–1068.

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.

Andrews, P W, G Banting, I Damjanov, D Arnaud, and P Avner (1984). Three monoclonal antibodies defining distinct differentiation antigens associated with different high molecular weight polypeptides on the surface of human embryonal carcinoma cells. *Hybridoma* 3.4, 347–361.

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.

Aron Badin, Romina, Aurore Bugi, Susannah Williams, Marta Vadori, Marie Michael, Caroline Jan, Alberto Nassi, Sophie Lecourtois, Antoine Blancher, Emanuele Cozzi, Philippe Hantraye, and Anselme L Perrier (2019). MHC matching fails to prevent long-term rejection of iPSC-derived neurons in non-human primates. *Nat. Commun.* 10.1, 4357.

Bagnoli, Johannes W, Christoph Ziegenhain, Aleksandar Janjic, Lucas E Wange, Beate Vieth, Swati Parekh, Johanna Geuder, Ines Hellmann, and Wolfgang Enard (2018). Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* 9.1, 2937.

Ban, Hiroshi, Naoki Nishishita, Noemi Fusaki, Toshiaki Tabata, Koichi Saeki, Masayuki Shikamura, Nozomi Takada, Makoto Inoue, Mamoru Hasegawa, Shin Kawamata, and Shin-Ichi Nishikawa (2011). Efficient generation of transgene-free human induced pluripotent stem cells (iPSCs) by temperature-sensitive Sendai virus vectors. *Proc. Natl. Acad. Sci. U. S. A.* 108.34, 14234–14239.

Beers, Jeanette, Kaari L Linask, Jane A Chen, Lauren I Siniscalchi, Yongshun Lin, Wei Zheng, Mahendra Rao, and Guokai Chen (2015). A cost-effective and efficient reprogramming platform for large-scale production of integration-free human induced pluripotent stem cells in chemically defined culture. *Sci. Rep.* 5, 11319.

Ben-Nun, Inbar Friedrich, Susanne C Montague, Marlys L Houck, Ha T Tran, Ibon Garitaonandia, Trevor R Leonardo, Yu-Chieh Wang, Suellen J Charter, Louise C Laurent, Oliver A Ryder, and Jeanne F Loring (2011). Induced pluripotent stem cells from highly endangered species. *Nat. Methods* 8.10, 829–831.

Bernloehr, Christian, Sascha Bossow, Guy Ungerechts, Sorin Armeanu, Wolfgang J Neubert, Ulrich M Lauer, and Michael Bitzer (2004). Efficient propagation of single gene deleted recombinant Sendai virus vectors. *Virus Res.* 99.2, 193–197.

Bharadwaj, Shantaram, Guihua Liu, Yingai Shi, Rongpei Wu, Bin Yang, Tongchuan He, Yuxin Fan, Xinyan Lu, Xiaobo Zhou, Hong Liu, Anthony Atala, Jan Rohozinski, and Yuanyuan Zhang (2013). Multipotential differentiation of human urine-derived stem cells: potential for therapeutic applications in urology. *Stem Cells* 31.9, 1840–1856.

Bininda-Emonds, Olaf R P, Marcel Cardillo, Kate E Jones, Ross D E MacPhee, Robin M D Beck, Richard Grenyer, Samantha A Price, Rutger A Vos, John L Gittleman, and Andy Purvis (2007). The delayed rise of present-day mammals. *Nature* 446.7135, 507–512.

Blake, Lauren E, Samantha M Thomas, John D Blischak, Chiaowen Joyce Hsiao, Claudia Chavarria, Marsha Myrthil, Yoav Gilad, and Bryan J Pavlovic (2018). A comparative study of endoderm differentiation in humans and chimpanzees. *Genome Biol.* 19.1, 162.

Bock, Christoph, Evangelos Kiskinis, Griet Verstappen, Hongcang Gu, Gabriella Boulting, Zachary D Smith, Michael Ziller, Gist F Croft, Mackenzie W Amoroso, Derek H Oakley, Andreas Gnirke, Kevin Eggan, and Alexander Meissner (2011). Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* 144.3, 439–452.

Bouma, Marga J, Maarten van Iterson, Bart Janssen, Christine L Mummery, Daniela C F Salvatori, and Christian Freund (2017). Differentiation-Defective Human Induced Pluripotent Stem Cells Reveal Strengths and Limitations of the Teratoma Assay and In Vitro Pluripotency Assays. *Stem Cell Reports* 8.5, 1340–1353.

Brink, Susanne C van den, Fanny Sage, Ábel Vértesy, Bastiaan Spanjaard, Josi Peterson-Maduro, Chloé S Baron, Catherine Robin, and Alexander van Oudenaarden (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* 14.10, 935–936.

Buermans, H P J and J T den Dunnen (2014). Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta* 1842.10, 1932–1941.

Buta, Christiane, Robert David, Ralf Dressel, Mia Emgård, Christiane Fuchs, Ulrike Gross, Lyn Healy, Jürgen Hescheler, Roman Kolar, Ulrich Martin, Harald Mikkers, Franz-Josef

Müller, Rebekka K Schneider, Andrea E M Seiler, Horst Spielmann, and Georg Weitzer (2013). Reconsidering pluripotency tests: do we still need teratoma assays? *Stem Cell Res.* 11.1, 552–562.

Cader, Zameel, Martin Graf, Mark Burcin, Carl-Fredrik Mandenius, and James A Ross (2019). Cell-Based Assays Using Differentiated Human Induced Pluripotent Cells. *Methods Mol. Biol.* 1994, 1–14.

Cannoodt, Robrecht, Wouter Saelens, and Yvan Saeys (2016). Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.* 46.11, 2496–2506.

Chambers, Stuart M, Christopher A Fasano, Eirini P Papapetrou, Mark Tomishima, Michel Sadelain, and Lorenz Studer (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat. Biotechnol.* 27.3, 275–280.

Chan, Elayne M, Sutheera Ratanasirintrawoot, In-Hyun Park, Philip D Manos, Yuin-Han Loh, Hongguang Huo, Justine D Miller, Odelya Hartung, Junsung Rho, Tan A Ince, George Q Daley, and Thorsten M Schlaeger (2009). Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. *Nat. Biotechnol.* 27.11, 1033–1037.

Charrier, Cécile, Kaumudi Joshi, Jaeda Coutinho-Budd, Ji-Eun Kim, Nelle Lambert, Jacqueline de Marchena, Wei-Lin Jin, Pierre Vanderhaeghen, Anirvan Ghosh, Takayuki Sassa, and Franck Polleux (2012). Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* 149.4, 923–935.

Chen, F C and W H Li (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68.2, 444–456.

Chen, Guokai, Daniel R Gulbranson, Zhonggang Hou, Jennifer M Bolin, Victor Ruotti, Mitchell D Probasco, Kimberly Smuga-Otto, Sara E Howden, Nicole R Diol, Nicholas E Propson, Ryan Wagner, Garrett O Lee, Jessica Antosiewicz-Bourget, Joyce M C Teng, and James A Thomson (2011). Chemically defined conditions for human iPSC derivation and culture. *Nat. Methods* 8.5, 424–429.

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

Chu, Li-Fang, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* 17.1, 173.

Consortium, Zoonomia and Zoonomia Consortium (2020). *A comparative genomics multitool for scientific discovery and conservation.*

Dannemann, Michael and Irene Gallego Romero (2021). Harnessing pluripotent stem cells as models to decipher human evolution. *FEBS J.*

Datlinger, Paul, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* 14.3, 297–301.

Debowski, Katharina, Charis Drummer, Jana Lentes, Maren Cors, Ralf Dressel, Thomas Lingner, Gabriela Salinas-Riester, Sigrid Fuchs, Erika Sasaki, and Rüdiger Behr (2016). The transcriptomes of novel marmoset monkey embryonic stem cell lines reflect distinct genomic features. *Sci. Rep.* 6, 29122.

Debowski, Katharina, Rita Warthemann, Jana Lentes, Gabriela Salinas-Riester, Ralf Dressel, Daniel Langenstroth, Jörg Gromoll, Erika Sasaki, and Rüdiger Behr (2015). Non-viral generation of marmoset monkey iPS cells by a six-factor-in-one-vector approach. *PLoS One* 10.3, e0118424.

Derr, Alan, Chaoxing Yang, Rapolas Zilionis, Alexey Sergushichev, David M Blodgett, Sambra Redick, Rita Bortell, Jeremy Luban, David M Harlan, Sebastian Kadener, Dale L Greiner, Allon Klein, Maxim N Artyomov, and Manuel Garber (2016). End Sequence Analysis Toolkit (ESAT) expands the extractable information from single-cell RNA-seq data. *Genome Res.* 26.10, 1397–1410.

Dixit, Atray, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, Eric S Lander, Jonathan S Weissman, Nir Friedman, and Aviv Regev (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167.7, 1853–1866.e17.

Duboule, D (1994). Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev. Suppl.*, 135–142.

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

— (2016). The Molecular Basis of Human Brain Evolution. *Curr. Biol.* 26.20, R1109–R1117.

Enard, Wolfgang, Molly Przeworski, Simon E Fisher, Cecilia S L Lai, Victor Wiebe, Takashi Kitano, Anthony P Monaco, and Svante Pääbo (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 418.6900, 869–872.

Endo, Yoshinori, Ken-Ichiro Kamei, and Miho Inoue-Murayama (2020). *Genetic Signatures of Evolution of the Pluripotency Gene Regulating Network across Mammals.*

Evans, M J and M H Kaufman (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature* 292.5819, 154–156.

Field, Andrew R, Frank M J Jacobs, Ian T Fiddes, Alex P R Phillips, Andrea M Reyes-Ortiz, Erin LaMontagne, Lila Whitehead, Vincent Meng, Jimi L Rosenkrantz, Mari Olsen, Max Hauessler, Sol Katzman, Sofie R Salama, and David Haussler (2019). Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes. *Stem Cell Reports* 12.2, 245–257.

Florio, Marta, Mareike Albert, Elena Taverna, Takashi Namba, Holger Brandl, Eric Lewitus, Christiane Haffner, Alex Sykes, Fong Kuan Wong, Jula Peters, Elaine Guhr, Sylvia Klemroth, Kay Prüfer, Janet Kelso, Ronald Naumann, Ina Nüsslein, Andreas Dahl, Robert Lachmann, Svante Pääbo, and Wieland B Huttner (2015). Human-specific gene *ARHGAP11B* promotes basal progenitor amplification and neocortex expansion. *Science* 347.6229, 1465–1470.

Fraser, M J, T Clszczon, T Elick, and C Bauser (1996). *Precise excision of TTAA-specific lepidopteran transposons piggyBac (IFP2) and tagalong (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera.*

Fujie, Yasumitsu, Noemi Fusaki, Tomohiko Katayama, Makoto Hamasaki, Yumi Soejima, Minami Soga, Hiroshi Ban, Mamoru Hasegawa, Satoshi Yamashita, Shigemi Kimura, Saori Suzuki, Tetsuro Matsuzawa, Hirofumi Akari, and Takumi Era (2014). New type of Sendai virus vector provides transgene-free iPS cells derived from chimpanzee blood. *PLoS One* 9.12, e113052.

Fusaki, Noemi, Hiroshi Ban, Akiyo Nishiyama, Koichi Saeki, and Mamoru Hasegawa (2009). 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.* 85.8, 348–362.

Gallego Romero, Irene, Bryan J Pavlovic, Irene Hernando-Herraez, Xiang Zhou, Michelle C Ward, Nicholas E Banovich, Courtney L Kagan, Jonathan E Burnett, Constance H Huang, Amy Mitrano, Claudia I Chavarria, Inbar Friedrich Ben-Nun, Yingchun Li, Karen Sabatini, Trevor R Leonardo, Mana Parast, Tomas Marques-Bonet, Louise C Laurent, Jeanne F Loring, and Yoav Gilad (2015). A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. *Elife* 4, e07103.

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.

Goode, David L, Gregory M Cooper, Jeremy Schmutz, Mark Dickson, Eidelyn Gonzales, Ming Tsai, Kalpana Karra, Eugene Davydov, Serafim Batzoglou, Richard M Myers, and Arend Sidow (2010). Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Res.* 20.3, 301–310.

Greenleaf, William J and Arend Sidow (2014). The future of sequencing: convergence of intelligent design and market Darwinism. *Genome Biol.* 15.3, 303.

Hayashi, Tetsutaro, Norito Shibata, Ryo Okumura, Tomomi Kudome, Osamu Nishimura, Hiroshi Tarui, and Kiyokazu Agata (2010). Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its “index sorting” function for stem cell research. *Dev. Growth Differ.* 52.1, 131–144.

Head, Steven R, H Kiyomi Komori, Sarah A LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R Salomon, and Phillip Ordoukhanian (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* 56.2, 61–4, 66, 68, passim.

Heide, Michael, Christiane Haffner, Ayako Murayama, Yoko Kurotaki, Haruka Shinohara, Hideyuki Okano, Erika Sasaki, and Wieland B Huttner (2020). Human-specific *ARHGAP11B* increases size and folding of primate neocortex in the fetal marmoset. *Science* 369.6503, 546–550.

Hemmi, Jacob J, Anuja Mishra, and Peter J Hornsby (2017). Overcoming barriers to reprogramming and differentiation in nonhuman primate induced pluripotent stem cells. *Primate Biol* 4.2, 153–162.

Hong, So Gun, Thomas Winkler, Chuanfeng Wu, Vicky Guo, Stefania Pittaluga, Alina Nicolae, Robert E Donahue, Mark E Metzger, Sandra D Price, Naoya Uchida, Sergei A Kuznetsov, Tina Kilts, Li Li, Pamela G Robey, and Cynthia E Dunbar (2014). Path to the clinic: assessment of iPSC-based cell therapies in vivo in a nonhuman primate model. *Cell Rep.* 7.4, 1298–1309.

Höpfl, Gisele, Max Gassmann, and Isabelle Desbaillets (2004). Differentiating embryonic stem cells into embryoid bodies. *Methods Mol. Biol.* 254, 79–98.

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

Hu, Haiyang, Masahiro Uesaka, Song Guo, Kotaro Shimai, Tsai-Ming Lu, Fang Li, Satoko Fujimoto, Masato Ishikawa, Shiping Liu, Yohei Sasagawa, Guojie Zhang, Shigeru Kuratani, Jr-Kai Yu, Takehiro G Kusakabe, Philipp Khaitovich, Naoki Irie, and EXPANDE Consortium (2017). Constrained vertebrate evolution by pleiotropic genes. *Nat Ecol Evol* 1.11, 1722–1730.

Hu, Kejin (2014). All roads lead to induced pluripotent stem cells: the technologies of iPSC generation. *Stem Cells Dev.* 23.12, 1285–1300.

Hubisz, M J, K S Pollard, and A Siepel (2011). *PHAST and RPHAST: phylogenetic analysis with space/time models.*

International Stem Cell Initiative (2018). Assessment of established techniques to determine developmental and malignant potential of human pluripotent stem cells. *Nat. Commun.* 9.1, 1925.

Irie, Naoki and Shigeru Kuratani (2011). Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat. Commun.* 2, 248.

Jaitin, Diego Adhemar, Assaf Weiner, Ido Yofe, David Lara-Astiaso, Hadas Keren-Shaul, Eyal David, Tomer Meir Salame, Amos Tanay, Alexander van Oudenaarden, and Ido Amit (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* 167.7, 1883–1896.e15.

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.

Jerber, Julie, Daniel D Seaton, Anna S E Cuomo, Natsuhiko Kumasaka, James Haldane, Juliette Steer, Minal Patel, Daniel Pearce, Malin Andersson, Marc Jan Bonder, Ed Mountjoy, Maya Ghoussaini, Madeline A Lancaster, HipSci Consortium, John C Marioni, Florian T Merkle, Daniel J Gaffney, and Oliver Stegle (2021). Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* 53.3, 304–312.

Kalinka, Alex T, Karolina M Varga, Dave T Gerrard, Stephan Preibisch, David L Corcoran, Julia Jarrells, Uwe Ohler, Casey M Bergman, and Pavel Tomancak (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468.7325, 811–814.

Kanton, Sabina, Michael James Boyle, Zhisong He, Malgorzata Santel, Anne Weigert, Fátima Sanchis-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.

Karikó, Katalin, Michael Buckstein, Houping Ni, and Drew Weissman (2005). Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. *Immunity* 23.2, 165–175.

Karikó, Katalin, Hiromi Muramatsu, Frank A Welsh, János Ludwig, Hiroki Kato, Shizuo Akira, and Drew Weissman (2008). Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol. Ther.* 16.11, 1833–1840.

Karikó, Katalin and Drew Weissman (2007). Naturally occurring nucleoside modifications suppress the immunostimulatory activity of RNA: implication for therapeutic RNA development. *Curr. Opin. Drug Discov. Devel.* 10.5, 523–532.

Kelley, Joanna L and Yoav Gilad (2020). Effective study design for comparative functional genomics. *Nat. Rev. Genet.* 21.7, 385–386.

King, M C and A C Wilson (1975). Evolution at two levels in humans and chimpanzees. *Science* 188.4184, 107–116.

Kircher, Martin, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46.3, 310–315.

Klein, Allon M, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161.5, 1187–1201.

Klingenstein, Stefanie, Moritz Klingenstein, Alexander Kleger, and Stefan Liebau (2020). From Hair to iPSCs-A Guide on How to Reprogram Keratinocytes and Why. *Curr. Protoc. Stem Cell Biol.* 55.1, e121.

Lang, Ren, Guihua Liu, Yingai Shi, Shantaram Bharadwaj, Xiaoyan Leng, Xiaobo Zhou, Hong Liu, Anthony Atala, and Yuanyuan Zhang (2013). Self-renewal and differentiation capacity of urine-derived stem cells after urine preservation for 24 hours. *PLoS One* 8.1, e53980.

Leblond, Claire S, Thuy-Linh Le, Simon Malesys, Freddy Cliquet, Anne-Claude Tabet, Richard Delorme, Thomas Rolland, and Thomas Bourgeron (2021). Operative list of genes associated with autism and neurodevelopmental disorders based on database review. *Mol. Cell. Neurosci.* 113, 103623.

Leha, Andreas, Nathalie Moens, Ruta Meleckyte, Oliver J Culley, Mia K Gervasio, Maximilian Kerz, Andreas Reimer, Stuart A Cain, Ian Streeter, Amos Folarin, Oliver Stegle, Cay M Kielty, HipSci Consortium, Richard Durbin, Fiona M Watt, and Davide Danovi (2016). A high-content platform to characterise human induced pluripotent stem cell lines. *Methods* 96, 85–96.

Lek, Monkol, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, Taru Tukiainen, Daniel P Birnbaum, Jack A Kosmicki, Laramie E Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N Cooper, et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536.7616, 285–291.

Levin, Michal, Tamar Hashimshony, Florian Wagner, and Itai Yanai (2012). Developmental milestones punctuate gene expression in the Caenorhabditis embryo. *Dev. Cell* 22.5, 1101–1108.

Liebau, Stefan, Pallavi U Mahaddalkar, Hans A Kestler, Anett Illing, Thomas Seufferlein, and Alexander Kleger (2013). A hierarchy in reprogramming capacity in different tissue microenvironments: what we know and what we need to know. *Stem Cells Dev.* 22.5, 695–706.

Linta, Leonhard, Marianne Stockmann, Karin N Kleinhans, Anja Böckers, Alexander Storch, Holm Zaehres, Qiong Lin, Gotthold Barbi, Tobias M Böckers, Alexander Kleger, and Stefan Liebau (2012). Rat embryonic fibroblasts improve reprogramming of human keratinocytes into induced pluripotent stem cells. *Stem Cells Dev.* 21.6, 965–976.

Liu, Gele, Brian T David, Matthew Trawczynski, and Richard G Fessler (2020). Advances in Pluripotent Stem Cells: History, Mechanisms, Technologies, and Applications. *Stem Cell Rev Rep* 16.1, 3–32.

Liu, Jialin, Rebecca R Viales, Pierre Khoueiry, James P Reddington, Charles Girardot, Eileen E M Furlong, and Marc Robinson-Rechavi (2021). The hourglass model of evolutionary conservation during embryogenesis extends to developmental enhancers with signatures of positive selection. *Genome Res.* 31.9, 1573–1581.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15.12, 550.

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

Ludwig, Tenneille E, Veit Bergendahl, Mark E Levenstein, Junying Yu, Mitchell D Probasco, and James A Thomson (2006). Feeder-independent culture of human embryonic stem cells. *Nat. Methods* 3.8, 637–646.

Lun, Aaron T L, Davis J McCarthy, and John C Marioni (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* 5, 2122.

Mabbott, Neil A, J Kenneth Baillie, Helen Brown, Tom C Freeman, and David A Hume (2013). An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics* 14, 632.

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.

Malhis, Nawar, Steven J M Jones, and Jörg Gsponer (2019). Improved measures for evolutionary conservation that exploit taxonomy distances. *Nat. Commun.* 10.1, 1556.

Malik, Nasir and Mahendra S Rao (2013). A review of the methods for human iPSC derivation. *Methods Mol. Biol.* 997, 23–33.

Marchetto, Maria C N, Iñigo Narvaiza, Ahmet M Denli, Christopher Benner, Thomas A Lazzarini, Jason L Nathanson, Apuã C M Paquola, Keval N Desai, Roberto H Herai, Matthew D Weitzman, Gene W Yeo, Alysson R Muotri, and Fred H Gage (2013). Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* 503.7477, 525–529.

Marioni, John C, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18.9, 1509–1517.

Martin, G R (1981). *Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells.*

Mitalipov, Shoukhrat, Hung-Chih Kuo, James Byrne, Lisa Clepper, Lorraine Meisner, Julie Johnson, Renee Zeier, and Don Wolf (2006). Isolation and characterization of novel rhesus monkey embryonic stem cell lines. *Stem Cells* 24.10, 2177–2186.

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.

Morrison, Michael, Christine Klein, Nicole Clemann, David A Collier, John Hardy, Barbara Heisserer, M Zameel Cader, Martin Graf, and Jane Kaye (2015). StemBANCC: Governing Access to Material and Data in a Large Stem Cell Research Consortium. *Stem Cell Rev Rep* 11.5, 681–687.

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.

Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320.5881, 1344–1349.

Nakagawa, Masato, Michiyo Koyanagi, Koji Tanabe, Kazutoshi Takahashi, Tomoko Ichisaka, Takashi Aoi, Keisuke Okita, Yuji Mochiduki, Nanako Takizawa, and Shinya Yamanaka (2008). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat. Biotechnol.* 26.1, 101–106.

Nakagawa, Masato, Yukimasa Taniguchi, Sho Senda, Nanako Takizawa, Tomoko Ichisaka, Kanako Asano, Asuka Morizane, Daisuke Doi, Jun Takahashi, Masatoshi Nishizawa, Yoshinori Yoshida, Taro Toyoda, Kenji Osafune, Kiyotoshi Sekiguchi, and Shinya Yamanaka (2014). A novel efficient feeder-free culture system for the derivation of human induced pluripotent stem cells. *Sci. Rep.* 4, 3594.

Nakai, Risako, Mari Ohnuki, Kota Kuroki, Haruka Ito, Hirohisa Hirai, Ryunosuke Kitajima, Toko Fujimoto, Masato Nakagawa, Wolfgang Enard, and Masanori Imamura (2018). Derivation of induced pluripotent stem cells in Japanese macaque (Macaca fuscata). *Sci. Rep.* 8.1, 12187.

Navara, Christopher S, Shital Chaudhari, and John R McCarrey (2018). Optimization of culture conditions for the derivation and propagation of baboon (Papio anubis) induced pluripotent stem cells. *PLoS One* 13.3, e0193195.

Navara, Christopher S, Jacey Hornecker, Douglas Grow, Shital Chaudhari, Peter J Hornsby, Justin K Ichida, Kevin Eggan, and John R McCarrey (2013). Derivation of induced pluripotent stem cells from the baboon: a nonhuman primate model for preclinical testing of stem cell therapies. *Cell. Reprogram.* 15.6, 495–502.

Navara, Christopher S, Jocelyn D Mich-Basso, Carrie J Redinger, Ahmi Ben-Yehudah, Ethan Jacoby, Elizabeta Kovkarova-Naumovski, Meena Sukhwani, Kyle Orwig, Naftali Kaminski, Carlos A Castro, Calvin R Simerly, and Gerald Schatten (2007). Pedigreed primate embryonic stem cells express homogeneous familial gene profiles. *Stem Cells* 25.11, 2695–2704.

Nelakanti, Raman V, Nigel G Kooreman, and Joseph C Wu (2015). Teratoma formation: a tool for monitoring pluripotency in stem cell research. *Curr. Protoc. Stem Cell Biol.* 32, 4A.8.1–4A.8.17.

Nichols, J, B Zevnik, K Anastassiadis, H Niwa, D Klewe-Nebenius, I Chambers, H Schöler, and A Smith (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* 95.3, 379–391.

Oberacker, Phil, Peter Stepper, Donna M Bond, Sven Höhn, Jule Focken, Vivien Meyer, Luca Schelle, Victoria J Sugrue, Gert-Jan Jeunen, Tim Moser, Steven R Hore, Ferdinand von Meyenn, Katharina Hipp, Timothy A Hore, and Tomasz P Jurkowski (2019). Bio-On-Magnetic-Beads (BOMB): Open platform for high-throughput nucleic acid extraction and manipulation. *PLoS Biol.* 17.1, e3000107.

Ohnuki, Mari, Koji Tanabe, Kenta Sutou, Ito Teramoto, Yuka Sawamura, Megumi Narita, Michiko Nakamura, Yumie Tokunaga, Masahiro Nakamura, Akira Watanabe, Shinya Yamanaka, and Kazutoshi Takahashi (2014). Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc. Natl. Acad. Sci. U. S. A.* 111.34, 12426–12431.

Okita, Keisuke, Yasuko Matsumura, Yoshiko Sato, Aki Okada, Asuka Morizane, Satoshi Okamoto, Hyenjong Hong, Masato Nakagawa, Koji Tanabe, Ken-Ichi Tezuka, Toshiyuki

Shibata, Takahiro Kunisada, Masayo Takahashi, Jun Takahashi, Hiroh Saji, and Shinya Yamanaka (2011). A more efficient method to generate integration-free human iPS cells. *Nat. Methods* 8.5, 409–412.

Okita, Keisuke, Tatsuya Yamakawa, Yasuko Matsumura, Yoshiko Sato, Naoki Amano, Akira Watanabe, Naoki Goshima, and Shinya Yamanaka (2013). An efficient nonviral method to generate integration-free human-induced pluripotent stem cells from cord blood and peripheral blood cells. *Stem Cells* 31.3, 458–466.

Parekh, Swati, Beate Vieth, Christoph Ziegenhain, Wolfgang Enard, and Ines Hellmann (2018). Strategies for quantitative RNA-seq analyses among closely related species.

Petit, I, N Salman Kesner, R Karry, O Robicsek, E Aberdam, F J Müller, D Aberdam, and D Ben-Shachar (2012). Induced pluripotent stem cells from hair follicles as a cellular model for neurodevelopmental disorders. *Stem Cell Res.* 8.1, 134–140.

Petkov, Stoyan G and Rüdiger Behr (2021). Generation of Marmoset Monkey iPSCs with Self-Replicating VEE-mRNAs in Feeder-Free Conditions. *Methods Mol. Biol.*

Piao, Yulan, Sandy Shen-Chi Hung, Shiang Y Lim, Raymond Ching-Bong Wong, and Minoru S H Ko (2014). Efficient generation of integration-free human induced pluripotent stem cells from keratinocytes by simple transfection of episomal vectors. *Stem Cells Transl. Med.* 3.7, 787–791.

Picelli, Simone, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10.11, 1096–1098.

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.

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.

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, 768391.

Ramaswamy, Krishna, Wing Yan Yik, Xiao-Ming Wang, Erin N Oliphant, Wange Lu, Darryl Shibata, Oliver A Ryder, and Joseph G Hacia (2015). Derivation of induced pluripotent stem cells from orangutan skin fibroblasts. *BMC Res. Notes* 8, 577.

Rao, Mahendra S and Nasir Malik (2012). Assessing iPSC reprogramming methods for their suitability in translational medicine. *J. Cell. Biochem.* 113.10, 3061–3068.

Ray, Arnab, Jahnavy Madhukar Joshi, Pradeep Kumar Sundaravadivelu, Khyati Raina, Nibedita Lenka, Vishwas Kaveeshwar, and Rajkumar P Thummer (2021). An Overview on Promising Somatic Cell Sources Utilized for the Efficient Generation of Induced Pluripotent Stem Cells. *Stem Cell Rev Rep* 17.6, 1954–1974.

Regev, Aviv, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. (2017). Science forum: the human cell atlas. *Elife* 6, e27041.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43.7, e47.

Romero, Irene Gallego, Ilya Ruvinsky, and Yoav Gilad (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* 13.7, 505–516.

Rozenblatt-Rosen, Orit, Michael J T Stubbington, Aviv Regev, and Sarah A Teichmann (2017). The Human Cell Atlas: from vision to reality. *Nature* 550.7677, 451–453.

Sacco, Anna Maria, Immacolata Belviso, Veronica Romano, Antonia Carfora, Fabrizio Schonauer, Daria Nurzynska, Stefania Montagnani, Franca Di Meglio, and Clotilde Castaldo (2019). Diversity of dermal fibroblasts as major determinant of variability in cell reprogramming. *J. Cell. Mol. Med.* 23.6, 4256–4268.

Saelens, Wouter, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 37.5, 547–554.

Sasaki, Erika, Kisaburo Hanazawa, Ryo Kurita, Akira Akatsuka, Takahito Yoshizaki, Hajime Ishii, Yoshikuni Tanioka, Yasuyuki Ohnishi, Hiroshi Suemizu, Ayako Sugawara, Norikazu

Tamaoki, Kiyoko Izawa, Yukoh Nakazaki, Hiromi Hamada, Hirofumi Suemori, Shigetaka Asano, Norio Nakatsuji, Hideyuki Okano, and Kenzaburo Tani (2005). Establishment of novel embryonic stem cell lines derived from the common marmoset (Callithrix jacchus). *Stem Cells* 23.9, 1304–1313.

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.

Schlaeger, Thorsten M, Laurence Daheron, Thomas R Brickler, Samuel Entwisle, Karrie Chan, Amelia Cianci, Alexander DeVine, Andrew Ettenger, Kelly Fitzgerald, Michelle Godfrey, Dipti Gupta, Jade McPherson, Prerana Malwadkar, Manav Gupta, Blair Bell, Akiko Doi, Namyoung Jung, Xin Li, Maureen S Lynes, Emily Brookes, et al. (2015). A comparison of non-integrating reprogramming methods. *Nat. Biotechnol.* 33.1, 58–63.

Shendure, Jay (2008). The beginning of the end for microarrays? *Nat. Methods* 5.7, 585–587.

Shiozawa, Seiji, Mayutaka Nakajima, Junko Okahara, Yoko Kuortaki, Fumihiko Kisa, Sho Yoshimatsu, Mari Nakamura, Ikuko Koya, Mika Yoshimura, Yohei Sasagawa, Itoshi Nikaido, Erika Sasaki, and Hideyuki Okano (2020). Primed to Naive-Like Conversion of the Common Marmoset Embryonic Stem Cells. *Stem Cells Dev.* 29.12, 761–773.

Siepel, Adam, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W James Kent, Webb Miller, and David Haussler (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15.8, 1034–1050.

Soumillon, Magali, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S Mikkelsen (2014). Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*, 003236. eprint: `1011.1669v3`.

Stadtfeld, Matthias, Nimet Maherali, David T Breault, and Konrad Hochedlinger (2008). Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. *Cell Stem Cell* 2.3, 230–240.

Staerk, Judith, Meelad M Dawlaty, Qing Gao, Dorothea Maetzel, Jacob Hanna, Cesar A Sommer, Gustavo Mostoslavsky, and Rudolf Jaenisch (2010). *Reprogramming of Human Peripheral Blood Cells to Induced Pluripotent Stem Cells.*

Stauske, Michael, Ignacio Rodriguez Polo, Wadim Haas, Debbra Yasemin Knorr, Thomas Borchert, Katrin Streckfuss-Bömeke, Ralf Dressel, Iris Bartels, Malte Tiburcy, Wolfram-Hubertus Zimmermann, and Rüdiger Behr (2020). Non-Human Primate iPSC Generation, Cultivation, and Cardiac Differentiation under Chemically Defined Conditions. *Cells* 9.6.

Steinle, Heidrun, Andreas Behring, Christian Schlensak, Hans Peter Wendel, and Meltem Avci-Adali (2017). Concise Review: Application of In Vitro Transcribed Messenger RNA for Cellular Engineering and Reprogramming: Progress and Challenges. *Stem Cells* 35.1, 68–79.

Streeter, Ian, Peter W Harrison, Adam Faulconbridge, The HipSci Consortium, Paul Flicek, Helen Parkinson, and Laura Clarke (2017). The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Res.* 45.D1, D691–D697.

Suemori, Hirofumi and Norio Nakatsuji (2006). Generation and characterization of monkey embryonic stem cells. *Methods Mol. Biol.* 329, 81–89.

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.

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.

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.

Tavares, Lucélia, Paula M Alves, Ricardo B Ferreira, and Claudia N Santos (2011). Comparison of different methods for DNA-free RNA isolation from SK-N-MC neuroblastoma. *BMC Res. Notes* 4, 3.

Tavernier, Geertrui, Katharina Wolfrum, Joseph Demeester, Stefaan C De Smedt, James Adjaye, and Joanna Rejman (2012). Activation of pluripotency-associated genes in mouse embryonic fibroblasts by non-viral transfection with in vitro-derived mRNAs encoding Oct4, Sox2, Klf4 and cMyc. *Biomaterials* 33.2, 412–417.

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.

Thomson, J A, J Kalishman, T G Golos, M Durning, C P Harris, and J P Hearn (1996). Pluripotent cell lines derived from common marmoset (Callithrix jacchus) blastocysts. *Biol. Reprod.* 55.2, 254–259.

Thomson, James A, Joseph Itskovitz-Eldor, Sander S Shapiro, Michelle A Waknitz, Jennifer J Swiergiel, Vivienne S Marshall, and Jeffrey M Jones (1998). *Embryonic Stem Cell Lines Derived from Human Blastocysts.*

Tomioka, Ikuo, Takuji Maeda, Hiroko Shimada, Kenji Kawai, Yohei Okada, Hiroshi Igarashi, Ryo Oiwa, Tsuyoshi Iwasaki, Mikio Aoki, Toru Kimura, Seiji Shiozawa, Haruka Shinohara, Hiroshi Suemizu, Erika Sasaki, and Hideyuki Okano (2010). Generating induced pluripotent stem cells from common marmoset (Callithrix jacchus) fetal liver cells using defined factors, including Lin28. *Genes Cells* 15.9, 959–969.

Tsankov, Alexander M, Veronika Akopian, Ramona Pop, Sundari Chetty, Casey A Gifford, Laurence Daheron, Nadejda M Tsankova, and Alexander Meissner (2015). A qPCR ScoreCard quantifies the differentiation potential of human pluripotent stem cells. *Nat. Biotechnol.* 33.11, 1182–1192.

Vallier, Ludovic, Thomas Touboul, Stephanie Brown, Candy Cho, Bilada Bilican, Morgan Alexander, Jessica Cedervall, Siddharthan Chandran, Lars Ahrlund-Richter, Anne Weber, and Roger A Pedersen (2009). Signaling pathways controlling pluripotency and early cell fate decisions of human induced pluripotent stem cells. *Stem Cells* 27.11, 2655–2666.

Velculescu, V E, L Zhang, B Vogelstein, and K W Kinzler (1995). Serial analysis of gene expression. *Science* 270.5235, 484–487.

Vera, J Cristobal, Christopher W Wheat, Howard W Fescemyer, Mikko J Frilander, Douglas L Crawford, Ilkka Hanski, and James H Marden (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* 17.7, 1636–1647.

Vidal, Simon E, Bhishma Amlani, Taotao Chen, Aristotelis Tsirigos, and Matthias Stadtfeld (2014). Combinatorial modulation of signaling pathways reveals cell-type-specific requirements for highly efficient and synchronous iPSC reprogramming. *Stem Cell Reports* 3.4, 574–584.

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.

Volpato, Viola, James Smith, Cynthia Sandor, Janina S Ried, Anna Baud, Adam Handel, Sarah E Newey, Frank Wessely, Moustafa Attar, Emma Whiteley, Satyan Chintawar, An Verheyen, Thomas Barta, Majlinda Lako, Lyle Armstrong, Caroline Muschet, Anna Artati, Carlo Cusulin, Klaus Christensen, Christoph Patsch, et al. (2018). Reproducibility of Molecular Phenotypes after Long-Term Differentiation to Human iPSC-Derived Neurons: A Multi-Site Omics Study. *Stem Cell Reports* 11.4, 897–911.

Volpato, Viola and Caleb Webber (2020). Addressing variability in iPSC-derived models of human disease: guidelines to promote reproducibility. *Dis. Model. Mech.* 13.1.

Wang, Xiliang, Yao He, Qiming Zhang, Xianwen Ren, and Zemin Zhang (2021). Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Genomics Proteomics Bioinformatics* 19.2, 253–266.

Ward, Michelle C and Yoav Gilad (2019). A generally conserved response to hypoxia in iPSC-derived cardiomyocytes from humans and chimpanzees. *Elife* 8.

Warren, Luigi, Philip D Manos, Tim Ahfeldt, Yuin-Han Loh, Hu Li, Frank Lau, Wataru Ebina, Pankaj K Mandal, Zachary D Smith, Alexander Meissner, George Q Daley, Andrew S Brack, James J Collins, Chad Cowan, Thorsten M Schlaeger, and Derrick J Rossi (2010). Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* 7.5, 618–630.

Watanabe, Toshiaki, Shun Yamazaki, Nao Yoneda, Haruka Shinohara, Ikuo Tomioka, Yuichiro Higuchi, Mika Yagoto, Masatsugu Ema, Hiroshi Suemizu, Kenji Kawai, and Erika Sasaki (2019). Highly efficient induction of primate iPS cells by combining RNA transfection and chemical compounds. *Genes Cells* 24.7, 473–484.

Wen, Lu and Fuchou Tang (2016). *Single-cell sequencing in stem cell biology.*

Woltjen, Knut, Iacovos P Michael, Paria Mohseni, Ridham Desai, Maria Mileikovsky, Riikka Hämäläinen, Rebecca Cowling, Wei Wang, Pentao Liu, Marina Gertsenstein, Keisuke Kaji, Hoon-Ki Sung, and Andras Nagy (2009). piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* 458.7239, 766–770.

Wu, Yuehong, Yong Zhang, Anuja Mishra, Suzette D Tardif, and Peter J Hornsby (2010). Generation of induced pluripotent stem cells from newborn marmoset skin fibroblasts. *Stem Cell Res.* 4.3, 180–188.

Wunderlich, Stephanie, Alexandra Haase, Sylvia Merkert, Jennifer Beier, Kristin Schwanke, Axel Schambach, Silke Glage, Gudrun Göhring, Eliza C Curnow, and Ulrich Martin (2012). Induction of pluripotent stem cells from a cynomolgus monkey using a polycistronic simian immunodeficiency virus-based vector, differentiation toward functional cardiomyocytes, and generation of stably expressing reporter lines. *Cell. Reprogram.* 14.6, 471–484.

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.

Xu, Chunhui, Elen Rosler, Jianjie Jiang, Jane S Lebkowski, Joseph D Gold, Chris O'Sullivan, Karen Delavan-Boorsma, Michael Mok, Adrienne Bronstein, and Melissa K Carpenter (2005a). Basic fibroblast growth factor supports undifferentiated human embryonic stem cell growth without conditioned medium. *Stem Cells* 23.3, 315–323.

Xu, Ren-He, Ruthann M Peck, Dong S Li, Xuezhu Feng, Tenneille Ludwig, and James A Thomson (2005b). Basic FGF and suppression of BMP signaling sustain undifferentiated proliferation of human ES cells. *Nat. Methods* 2.3, 185–190.

Yada, Ravi Chandra, So Gun Hong, Yongshun Lin, Thomas Winkler, and Cynthia E Dunbar (2017). Rhesus Macaque iPSC Generation and Maintenance. *Curr. Protoc. Stem Cell Biol.* 41, 4A.11.1–4A.11.13.

Yakubov, Eduard, Gidi Rechavi, Shmuel Rozenblatt, and David Givol (2010). Reprogramming of human fibroblasts to pluripotent stem cells using mRNA of four transcription factors. *Biochem. Biophys. Res. Commun.* 394.1, 189–193.

Yanai, Itai, Leonid Peshkin, Paul Jorgensen, and Marc W Kirschner (2011). Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. *Dev. Cell* 20.4, 483–496.

Yao, Shuyuan, Tanya Sukonnik, Tara Kean, Rikki R Bharadwaj, Peter Pasceri, and James Ellis (2004). Retrovirus silencing, variegation, extinction, and memory are controlled by a dynamic interplay of multiple epigenetic modifications. *Mol. Ther.* 10.1, 27–36.

Yi, Hana, Yong-Joon Cho, Sungho Won, Jong-Eun Lee, Hyung Jin Yu, Sujin Kim, Gary P Schroth, Shujun Luo, and Jongsik Chun (2011). Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic Acids Res.* 39.20, e140.

Yoshimatsu, Sho, Mayutaka Nakajima, Aozora Iguchi, Tsukasa Sanosaka, Tsukika Sato, Mari Nakamura, Ryusuke Nakajima, Eri Arai, Mitsuru Ishikawa, Kent Imaizumi, Hirotaka Watanabe, Junko Okahara, Toshiaki Noce, Yuta Takeda, Erika Sasaki, Rüdiger Behr, Kazuya Edamura, Seiji Shiozawa, and Hideyuki Okano (2021). Non-viral Induction of Transgene-free iPSCs from Somatic Fibroblasts of Multiple Mammalian Species. *Stem Cell Reports* 16.4, 754–770.

Yu, Junying, Kejin Hu, Kim Smuga-Otto, Shulan Tian, Ron Stewart, Igor I Slukvin, and James A Thomson (2009). Human induced pluripotent stem cells free of vector and transgene sequences. *Science* 324.5928, 797–801.

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.

Zhang, Yuanyuan, Elena McNeill, Hong Tian, Shay Soker, Karl-Erik Andersson, James J Yoo, and Anthony Atala (2008). Urine derived cells are a potential source for urological tissue reconstruction. *J. Urol.* 180.5, 2226–2233.

Zhao, Shanrong, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* 9.1, e78644.

Zhao, Shanrong, Ying Zhang, Ramya Gamini, Baohong Zhang, and David von Schack (2018). Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci. Rep.* 8.1, 4781.

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.

Zhou, Ting, Christina Benda, Sarah Dunzinger, Yinghua Huang, Jenny Cy Ho, Jiayin Yang, Yu Wang, Ya Zhang, Qiang Zhuang, Yanhua Li, Xichen Bao, Hung-Fat Tse, Johannes Grillari, Regina Grillari-Voglauer, Duanqing Pei, and Miguel A Esteban (2012). Generation of human induced pluripotent stem cells from urine samples. *Nat. Protoc.* 7.12, 2080–2089.

Zhou, Ting, Christina Benda, Sarah Duzinger, Yinghua Huang, Xingyan Li, Yanhua Li, Xiangpeng Guo, Guokun Cao, Shen Chen, Lili Hao, Yau-Chi Chan, Kwong-Man Ng, Jenny Cy Ho, Matthias Wieser, Jiayan Wu, Heinz Redl, Hung-Fat Tse, Johannes Grillari, Regina Grillari-Voglauer, Duanqing Pei, et al. (2011). Generation of induced pluripotent stem cells from urine. *J. Am. Soc. Nephrol.* 22.7, 1221–1228.

Ziegenhain, Christoph, Beate Vieth, Swati Parekh, Ines Hellmann, and Wolfgang Enard (2018). Quantitative single-cell transcriptomics. *Brief. Funct. Genomics* 17.4, 220–232.

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 Sequence divergence 5
2 How comparative data can help to infer functional relevance 7
3 General RNA-seq workflow 10
4 Isolation of single cells 11
5 Generation of iPSCs 16
6 Overview of reprogramming methods 19

Acknowledgements

First and foremost, I would like to thank my advisor and mentor Wolfgang Enard. I am more than grateful that you gave me the chance to learn and grow so much in this amazing environment that you and Ines created. You are the best boss and mentor I can imagine. Thank you for your open ear on every matter, your motivation and that talking to you is always reassuring and of tremendous help every single time.

I would also like to thank Ines Hellmann, the other head of the Enard-Hellmann-WG, especially for your advice on computational and scientific questions, but also for fun social activities and for taking away my fear of dogs.

Moreover, I am very grateful to Mari Ohnuki for letting me join her projects, without your help I would have never got to see these amazing iPSCs that fascinated me so much, or your truly beautiful home country. Thank you for all your moral support and patience!

Also, a big thank you to Karin Bauer, for charming walks to the ubahn with good conversations; Ines Bliesener, for always listening and helping; Stephanie Färberböck, for the tremendous help in cell culture and good vibes; and all of you, together with Frau Zhao for always helping out with the really difficult thing throughout this PhD journey – the bureaucracy!

Special thanks to Lu and Aleks, for being the best and most helpful colleagues and amazing friends. I'm more than happy to have had you there during the difficult times but of course also all the fun times we had during the last years in the office, in apartments, on streets, rooftops and boats. And thank you for bringing Paulina and Michael into the game, making us hexlemma and the times even more cheerful. Couldn't have imagined it any better.

Thanks to Zanita for all the conversations and phone calls about important and not so important things and the fun times in and outside the lab. Also, for listening to all my complains and helping me out with everything computational and my unimaginativeness.

Thank you, Daniel and Johannes, for always being helpful and answering any question, from the beginning of my master studies until today. And of course, also all the great times we had in the sun.

Naturally thanks to all the amazing people I had the pleasure to work with during the last years: Beate, Christoph and Swati for demonstrating how great a PhD can be if you have good people supporting you and for giving patient explanations to the many stupid questions I might have had. Philipp for always being patiently helpful, your contagious laughter and the pringles I stole from you and Zane. Fiona and Jessy for your help, good coffee and nice breaks. And all of the other members and students: Ilse, Simon, Nik, Rudi, Arthur, Isabel, Veronika, Theresa, Selin, Natalia, Chris, Zeynep . . . and everyone who I might have forgotten!

Finally I want to thank my family, especially: Mama, for your unlimited support and because I can always ask you anything and you patiently answer even though I've asked a thousand times before; Dati for always cheering me up and being the most understanding person ever; Diti for always being there and helping with every little thing even if you don't know and Jui for your patience when I freak out, your persistent cheerfulness and kindness, your help and your love. Without you this wouldn't have been possible at all.

Curriculum Vitae

RAISSA JOHANNA ANDREA GEUDER

Education

February 2017 - present	**Doctoral Candidate - Human Genomics** **Ludwig-Maximilians University** Project: Generating and characterizing primate iPSCs for evolutionary analyses
October 2014 - September 2016	**Master of Science - Biology** **Ludwig-Maximilians University** Thesis: Estalishing methods for Ggnerating iPSCs from primates
October 2011 - September 2014	**Bachelor of Science - Biology** **Ludwig-Maximilians University** Thesis: Influence of Humanized Foxp2 on amphetamine-induced striatal gene expression in mice

Experience

September 2019 - October 2019	**Intership at Center for iPS Cell Research and Application (CiRA), Kyoto Japan, Woltjen group** Projects: The influence of KRAB on reversible gene exepression / Generation of dox-inducible PB-dCas9-KRAB-hiPS cell line/ Generation of an iPSC line harbouring a naturally occoring deletion variant using microhomology
March 2015 - October 2016	**Student assistant at LMU munich, Enard group** Project:Conditional knock-out of FOXP2 in the mouse brain
2012-2013	**Student assistant Becker und Kollegen** Microbiology and Diagnostics

Conferences & Courses

Single Cell Genomics Symposium (Poster)
2022 March, Barcelona

Stem Cell Network, 10th International Meeting (Elevator Pitch)
2021, Virtual conference
Pitch award: 2nd place

International Symposium Acute leukemias
Munich Germany
2017 & 2019

Visual communication webinar
March 2021

Teaching

Humanphysiologie - Practical course
2021

Supervision of Bachelor and Master Thesis and Research Course Students
2017-2021

Human Biologie I - Serology Lecture and practical course
2017-2021

Induced Pluripotent Stem Cell technologies - Seminar
2018

Publications

A non-invasive method to generate induced pluripotent stem cells from primate urine.
Geuder, J., Wange, L. E., Janjic, A., Radmer, J., Janssen, P., Bagnoli, J. W., Müller, S., Kaul, A., Ohnuki, M., & Enard, W. (2021) Scientific Reports, 11(1), 3516.
https://doi.org/10.1038/s41598-021-82883-0

Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq.
Bagnoli, J. W., Ziegenhain, C., Janjic, A., Wange, L. E., Vieth, B., Parekh, S., **Geuder, J**.,Hellmann, I., & Enard, W. (2018). Nature Communications, 9(1), 1–8.
https://doi.org/10.1038/s41467-018-05347-6

prime-seq efficient and powerful bulk RNA-sequencing.
Janjic, A., Wange, L. E.*, Bagnoli, J. W., **Geuder, J**., Nguyen, P., Richter, D., Vieth, B., Ziegenhain, C., Vick, B., Hellmann, I., & Enard, W. (2022). Genome Biology, 23, 88
https://doi.org/10.1186/s13059-022-02660-8

TRNP1 sequence , function and regulation co-evolve with cortical folding in mammals.
Kliesmete, Z., Wange, L. E.*, Vieth, B., Esgleas, M., Radmer, J., **Geuder, J.**, Richter, D., Ohnuki, M., & Magdalena, G. (2021). bioRxiv, 2021.02.05.429919.
https://doi.org/10.1101/2021.02.05.429919