

---

Measuring primate gene expression  
evolution using high throughput  
transcriptomics and massively parallel  
reporter assays

---

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

Lucas Esteban Wange

München 2022







---

Measuring primate gene expression  
evolution using high throughput  
transcriptomics and massively parallel  
reporter assays

---

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

Lucas Esteban Wange

München 2022

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

Erstgutachter: Professor Dr. Wolfgang Enard

Zweitgutachter: Professor Dr. John Parsch

Tag der Abgabe: 30. June 2022

Tag der mündlichen Prüfung: 21. October 2022

## Eidesstattliche Erklärung

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

München, den 30.06.2022

L.Wange

(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 30.06.2022

L.Wange

(Unterschrift)

\*) Nichtzutreffendes streichen



# Contents

<b>Abbreviations</b>	<b>xi</b>
<b>Publications</b>	<b>xv</b>
<b>Declarations</b>	<b>xx</b>
<b>Summary</b>	<b>1</b>
<b>1 Introduction</b>	<b>5</b>
1.1 Reading the Code of Life - A brief history of genomics . . . . .	5
1.1.1 The genetic code . . . . .	6
1.1.2 Sequencing technologies . . . . .	6
1.1.3 From genome to phenotype . . . . .	11
1.2 Gene expression quantification - the era of RNA-sequencing . .	12
1.2.1 What we talk about when we talk about RNA-sequencing	13
1.2.2 The next level - single cell RNA-sequencing . . . . .	18
1.2.3 Computational analysis of gene expression data . . . . .	22
1.2.4 Optimizing cost-efficiency of RNA-sequencing protocols .	26
1.3 Regulation of gene expression - The next frontier . . . . .	28
1.3.1 Measuring activity of regulatory regions is challenging . .	29

1.3.2	Massively Parallel Reporter Assays measure CRE activity efficiently . . . . .	29
1.4	Applying RNA-seq and MPRA to measure gene expression evolution across primates . . . . .	32
1.4.1	Gene expression is encoded in the genome . . . . .	34
1.4.2	Induced pluripotent stem cells are an invaluable tool for comparative gene expression evolution . . . . .	35
1.4.3	Brain size and folding are among the most diverged traits in mammals . . . . .	36
1.4.4	Co-evolution of regulatory sequences of <i>TRNP1</i> with cortical folding . . . . .	37
1.4.5	Measuring primate gene expression evolution using high throughput transcriptomics and massively parallel reporter assays . . . . .	38
<b>2</b>	<b>Discussion</b>	<b>41</b>
2.1	Developing and Improving RNA-sequencing methods . . . . .	41
2.1.1	The issue of barcode swapping . . . . .	42
2.1.2	Using statistical power analysis for method development	43
2.1.3	The perfect RNA-seq method does not exist . . . . .	44
2.1.4	Beyond RNA-seq . . . . .	45
2.2	Measuring and linking enhancer activity at scale . . . . .	46
2.2.1	Limitations and strength of MPRA . . . . .	47
2.2.2	New tools that can help derive rules of gene regulation by enhancers . . . . .	50

---

2.2.3	If we can't crack the regulatory code, we have to learn it from the data . . . . .	53
2.3	Comparing gene expression between species relies on gene orthology	55
2.3.1	Orthologs, Paralogs and functional conservation . . . . .	55
2.3.2	Potential ways out, beyond single gene comparisons . . . . .	56
<b>3</b>	<b>Conclusion and Outlook</b>	<b>57</b>
<b>4</b>	<b>References</b>	<b>59</b>
<b>5</b>	<b>Appendices</b>	<b>93</b>
5.1	Sensitive and powerful single-cell RNA sequencing using mcSCRBSeq . . . . .	94
5.2	Benchmarking single-cell RNA-sequencing protocols for cell atlas projects . . . . .	127
5.3	Prime-seq, efficient and powerful bulk RNA-sequencing . . . . .	201
5.4	A non-invasive method to generate induced pluripotent stem cells from primate urine . . . . .	255
5.5	TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals . . . . .	279
	<b>Acknowledgements</b>	<b>313</b>
	<b>Curriculum Vitae</b>	<b>317</b>





# List of Figures

1	Three generations of DNA Sequencing Technologies . . . . .	7
2	Flavours of RNA-seq library preparation methods . . . . .	14
3	Single cell RNA-seq innovations . . . . .	20
4	Workflow of Massively Parallel Reporter Assays . . . . .	30
5	Comparative genomics on three levels . . . . .	33



# Abbreviations

Abbreviation	Definition
AGBT	Advances in Genome Biology and Technology Meeting
AID	Auxin-Inducible Degron
ATAC-seq	Assay of Transposase Accessible Chromatin and Sequencing
BAC	Bacterial Artificial Chromosome
CAGE	Cap Analysis of Gene Expression
cDNA	complementary DNA
ChIA-PET	Chromatin interaction analysis with paired-end tag sequencing
ChIP-seq	Chromatin Immuno Precipitation Sequencing
CPM	Counts Per Million
CRE	<i>cis</i> -Regulatory Element
CRISPR	Clustered Regularly Interspersed Short Palindromic Repeats
CRISPRi	CRISPR mediated inhibition
ddNTP	Dideoxynucleosidtriphosphat
DE	Differential Expression
DEA	Differential Expression Analysis
DNA	Deoxyribonucleic Acid
dTTP	Deoxythymidine Triphosphate
dUTP	Deoxyuridine Triphosphate
eQTL	expression Quantitative Trait Loci
ERCC	External RNA Controls Consortium
FACS	Fluorescence Activated Cell Sorting
FDR	False Discovery Rate
GFP	Green Fluorescent Protein
GI	Gyrification Index
GLM	General Linear Model
gRNA	guide RNA
GSEA	Gene Set Enrichment Analysis
GTEEx	Genotype Tissue Expression Consortium
HCA	Human Cell Atlas
HCA	Human Cell Atlas
HEK293T	Human Embryonic Kidney cell line
HiC	Genomewide Chromosome Confirmation Capture protocol
IFC	Integrated Fluidic Circuits
iPSC	induced Pluripotent Stem Cells
IVT	<i>In Vitro</i> Transcription
kb	kilobases
lncRNA	long non coding RNAs
M-MLV RT	Moloney Murine Leukemia Virus Reverse Transcriptase

<b>Abbreviation</b>	<b>Definition</b>
mcSCRB-seq	molecular crowding SCRb-seq
MOI	multiplicity of infection
MPRA	Massively Parallel Reporter Assay
MR	Median of Ratios
mRNA	messenger Ribonucleic Acid
NGS	Next Generation Sequencing
NPC	Neural Progenitor Cells
PBMCs	Peripheral Blood Mononuclear Cells
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PEG	Polyethylene Glycol
PGLS	Phylogenetic Generalized Least Squares
PHRED	Sequence Quality Score
RNA-seq	Ribonucleic Acid sequencing
RRBS	Reduced Representation Bisulfite Sequencing
rRNA	ribosomal RNA
RT	Reverse Transcription
RT-qPCR	Reverse Transcription quantitative PCR
SAGE	Serial Analysis of Gene Expression
SCRb-seq	Single Cell RNA Barcoding and Sequencing
scRNA-seq	single cell RNA-seq
sncRNAs	small non-coding RNAs
SPRI	Solid Phase Reversible Immobilisation
STARR-seq	Self Transcribing Active Regulatory Region Sequencing
TAD	Topologically Associating Domain
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
TMM	Trimmed Mean of M-values
TopGO	Gene Ontologies
TPR	True Positive Rate
tRNA	transfer RNA
TRNP1	TMF-Regulated Nuclear Protein 1
tSNE	t-distributed Stochastic Neighbor Embedding
TSO	Template Switch Oligo
TSS	Transcription Start Sites
TSS	Transcription Start Site
UMAP	Uniform Manifold Approximation and Projection
UMIs	Unique Molecular Identifiers
UTR	Untranslated Region



# 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. Mereu E., Lafzi A., Moutinho C., ... **Wange L.E.** et al.  
"Benchmarking single-cell RNA-sequencing protocols for cell atlas projects"  
*Nature Biotechnology* 38, 747–755 (2020)  
doi: <https://doi.org/10.1038/s41587-020-0469-4>
- III. 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>
- IV. 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>
- V. 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>





# Additional Publications

- VI. Bagnoli J.W., **Wange L.E.**, Janjic A., Enard W.  
"Studying Cancer Heterogeneity by Single-Cell RNA Sequencing"  
In: Küppers R. (eds) Lymphoma. Methods in Molecular Biology, vol 1956.  
Humana Press, New York, NY. (2019)  
doi: [https://doi.org/10.1007/978-1-4939-9151-8\\_14](https://doi.org/10.1007/978-1-4939-9151-8_14)
- VII. Kempf J.M., Weser S., Bartoschek M.D., ,... **Wange L.E.** et al.  
"Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance  
in AML"  
Scientific Reports 11, 5838 (2021)  
doi: <https://doi.org/10.1038/s41598-021-84708-6>
- VIII. Porquier A., Tisserant C., Salinas F.,...**Wange L.E.** et al  
"Retrotransposons as pathogenicity factors of the plant pathogenic fungus *Botrytis cinerea*"  
Genome Biology 22, 225 (2021)  
doi: <https://doi.org/10.1186/s13059-021-02446-4>
- IX. Pekayvaz K., Leunig A., Kaiser R, ... **Wange L.E.** et al.  
"Protective immune trajectories in early viral containment of non-pneumonic SARS-CoV-2 infection"  
Nature Communications 13, 1018 (2022)  
doi: <https://doi.org/10.1038/s41467-022-28508-0>



# Declarations of contribution as first-author

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

This study was conceived by Aleksandar Janjic, me, Christoph Ziegenhain and Wolfgang Enard. Johanna Geuder, Aleksandar Janjic and Phong Nguyen prepared iPSC, HEK293T, and tissue samples. Johanna Geuder performed neural differentiation experiments. Binje Vick and Irmela Jeremias generated and supplied AML-PDX samples. Daniel Richter and Johannes Walter Bagnoli designed barcoded primers. Aleksandar Janjic, I, Johannes Walter Bagnoli and Phong Nguyen performed the RNA-seq experiments. Aleksandar Janjic and I performed sensitivity and gene expression analysis. I performed power analyses with computational and statistical support from Beate Vieth and Ines Hellmann. Aleksandar Janjic, I and Wolfgang Enard wrote the manuscript.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, we confirm the substantial contributions to this publication as described above.



---

Aleksandar Janjic

---

Lucas Esteban Wange

---

Wolfgang Enard

## TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals

Magdalena Götz proposed the project Wolfgang Enard and Ines Hellmann conceived the approaches of this study. Beate Vieth designed all initial sequence acquisitions. I, Matthias Hülsmann, Daniel Richter and Jessica Radmer conducted the MPRA. I conducted the RNA-seq experiment and performed primary data processing. Miriam Esgleas designed and conducted the proliferation assay. Jessica Radmer, Johanna Geuder and Mari Ohnuki were responsible for all primate cell culture work. Zane Kliesmete collected, integrated and analyzed all data. Zane Kliesmete, Ines Hellmann, and Wolfgang Enard wrote the manuscript.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, we confirm the substantial contributions to this publication as described above.



---

Zane Kliesmete



---

Lucas Esteban Wange

---

Wolfgang Enard

# Declarations of contribution as a co-author

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

Wolfgang Enard and Christoph Ziegenhain conceived the study. Optimization experiments and sequencing library preparations were done by Johannes Bagnoli, Aleksandar Janjic, Christoph Ziegenhain and me. Sequencing data processing and analysis was done by and Christoph Ziegenhain Johannes Bagnoli, Aleksandar Janjic, Beate Vieth and Swati Parekh. Johannes Bagnoli, Aleksandar Janjic, Wolfgang Enard and Christoph Ziegenhain wrote the manuscript.

## **Benchmarking single-cell RNA-sequencing protocols for cell atlas projects**

This study was conceived and supervised by Holger Heyn. Catia Moutinho, Adrian Alvarez and Eduard Batlle prepared the reference sample. Elisabetta Mereu and Atefeh Lafzi performed all data analyses. Aleksandar Janjic, I and Johannes Walter Bagnoli performed fluorescence activated cell sorting (FACS), scRNA sequencing library preparations and primary data analysis for gmcSCRB-seq. Holger Heyn, Elisabetta Mereu and Atefeh Lafzi wrote the manuscript with contributions from all co-authors.

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

Mari Ohnuki, Wolfgang Enard and Johanna Geuder had the idea for this work. Johanna Geuder established iPSC lines and conducted differentiation experiments. EB differentiation and immunostaining experiments were performed by Johanna Geuder with help from Jessica Radmer. I, Aleksandar Janjic, Johannes W. Bagnoli and Philipp Janssen and Johanna Geuder generated and analyzed RNA-seq data. The manuscript was written by Wolfgang Enard and Johanna Geuder.

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

---

Wolfgang Enard



# Summary

A key question in biology is how one genome sequence can lead to the great cellular diversity present in multicellular organisms. Enabled by the sequencing revolution, RNA sequencing (RNA-seq) has emerged as a central tool to measure transcriptome-wide gene expression levels. More recently, single cell RNA-seq was introduced and is becoming a feasible alternative to the more established bulk sequencing. While many different methods have been proposed, a thorough optimisation of established protocols can lead to improvements in robustness, sensitivity, scalability and cost effectiveness.

Towards this goal, I have contributed to optimizing the single cell RNA-seq method "Single Cell RNA Barcoding and sequencing" (SCRB-seq) and publishing an improved version that uses optimized reaction conditions and molecular crowding (mcSCRB-seq). mcSCRB-seq achieves higher sensitivity at lower cost per cell and shows the highest RNA capture rate when compared with other published methods. We next sought the direct comparison to other scRNA-seq protocols within the Human Cell Atlas (HCA) benchmarking effort. Here we used mcSCRB-seq to profile a common reference sample that included heterogeneous cell populations from different sources.

Transfer of the acquired knowledge on single cell RNA sequencing methods to bulk RNA-seq, led to the development of the prime-seq protocol. A sensitive, robust and cost-efficient bulk RNA-seq protocol that can be performed in any molecular biology laboratory. We compared the data generated, using the prime-seq protocol to the gold standard method TruSeq, using power simulations and found that the statistical power to detect differentially expressed genes is comparable, at 40-fold lower cost.

While gene expression is an informative phenotype, the regulation that leads to the different

phenotypes is still poorly understood. A state-of-the-art method to measure the activity of cis-regulatory elements (CRE) in a high throughput fashion are Massively Parallel Reporter Assays (MPRA). These assays can be used to measure the activity of thousands of *cis*-Regulatory Elements (CRE) in parallel.

A good way to decode the genotype to phenotype conundrum is using evolutionary information. Cross-species comparisons of closely related species can help understand how particular diverging phenotypes emerged and how conserved gene regulatory programs are encoded in the genome. A very useful tool to perform comparative studies are cell lines, particularly induced Pluripotent Stem Cells (iPSCs). iPSCs can be reprogrammed from different primary somatic cells and are per definition pluripotent, meaning they can be differentiated into cells of all three germ layers. A main challenge for primate research is to obtain primary cells. To this end I contributed to establishing a protocol to generate iPSCs from a non-invasive source of primary cells, namely urine. By using prime-seq we characterized the primary Urine Derived Stem Cells (UDSCs) and the reprogrammed iPSCs.

Finally, I used an MPRA to measure activity of putative regulatory elements of the gene *TRNP1* across the mammalian phylogeny. We found co-evolution of one particular CRE with brain folding in old world monkeys. To validate the finding we looked for transcription factor binding sites within the identified CRE and intersected the list with transcription factors confirmed to be expressed in the cellular system using prime-seq. In addition we found that changes in the protein coding sequence of *TRNP1* and neural stem cell proliferation induced by *TRNP1* orthologs correlate with brain size.

In summary, within my doctorate I developed methods that enable measuring gene expression and gene regulation in a comparative genomics setting. I further applied these methods in a cross mammalian study of the regulatory sequences of the gene *TRNP1* and its association with brain phenotypes.







# 1 | Introduction

## 1.1 Reading the Code of Life - A brief history of genomics

Ever since desoxyribonucleic acid (DNA) has been identified as the bio molecule holding the heritable instructions for building a functioning organism (Avery et al. 1944), much effort was put into reading and understanding this encrypted information. It was clear early on that DNA consisted of four chemically distinct building blocks, the so called nucleotides, each consisting of a sugar, a phosphate group and four nitrogen bases Adenine, Guanine, Thymine and Cytosine (Hammarsten 1895; Levene and Mori 1929; Levene and Tipson 1935). How only four "letters", A,T,G and C could encode the information necessary to build not only humans, but all living creatures on this planet was a long standing question and is still not fully understood despite incredible advances.

A first important step however, was to understand the structure of DNA and how the four bases interacted. Many groups worked on this particular problem and made pivotal contributions, first and foremost Rosalind Franklin's group (Franklin and Gosling 1953). Finally, James Watson and Francis Crick put together the pieces of evidence and published their model of the DNA double helix in 1953 (Watson and Crick 1953). Briefly, the DNA exists as two complementary strands that are held together by hydrogen bonds between a pair of bases. The pairing of the bases is non-random in a way that Adenine always pairs with Thymine and Guanine always pairs with Cytosine. Watson and Crick postulated, based on a picture taken in Rosalind Franklin's laboratory, that these two strands form a so called

double helix.

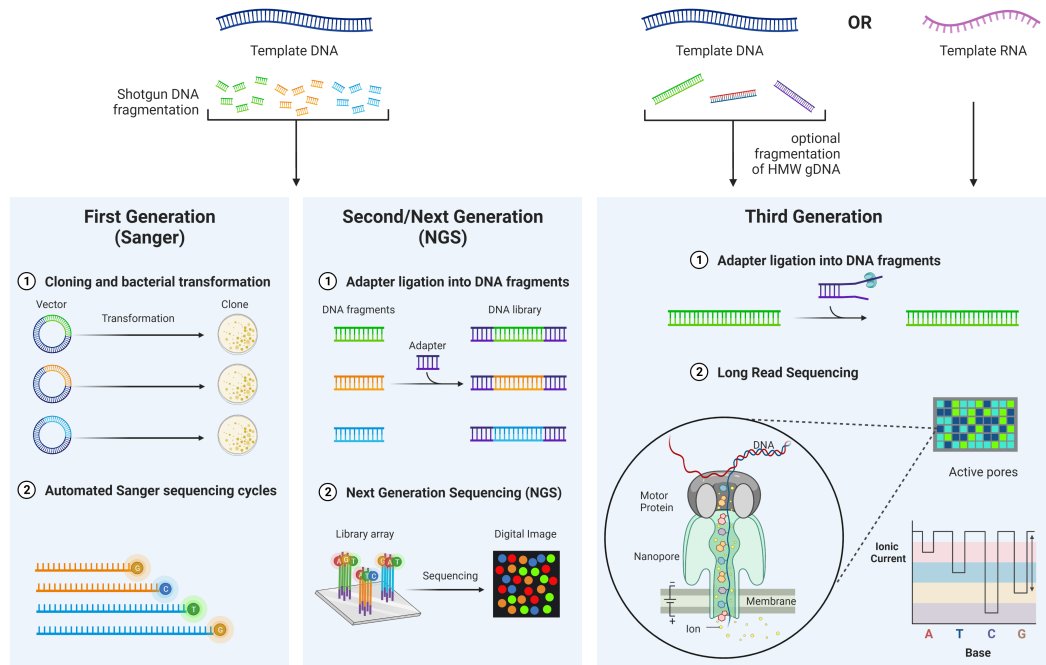
### 1.1.1 The genetic code

While this finding was remarkable, the next step in understanding the information stored in the entirety of our DNA, our genome, was to identify different functional units and how they are translated into proteins that execute the information in our genome. Those functional units were first identified by Gregor Mendel (Mendel 1866) and later termed "genes". A genetic code, that relates the four bases to 20 amino acids that form all proteins, was proposed by Crick and others (Crick et al. 1961), experimentally confirmed with the discovery of the first "word" or codon by Marshall Nirenberg and colleagues (Nirenberg and Matthaei 1961; Nirenberg and Leder 1964) and finished in 1966 by Har Gobind Khorana and colleagues (Khorana et al. 1966).

However, the first step to understand the information stored in our genome is to be able to read its sequence efficiently. Researchers worked on developing technologies to read the sequence of individual genes of interest already for decades (Sanger et al. 1973; Fiers et al. 1976), when in 1990 an international consortium started to read and assemble the entire human genome (National Research Council et al. 1988; National Center for Human Genome Research (U.S.) et al. 1990). The project was termed the Human Genome Project and the international consortium aimed to complete the entire genome sequence within 15 years. A key limitation in this endeavor was the ability to read long stretches of DNA efficiently and at sufficient throughput.

### 1.1.2 Sequencing technologies

The first generation sequencing methods were based on incorporation of radioactively or later fluorescently labeled nucleotides at the end of a DNA fragment (Maxam and Gilbert 1977; Sanger and Coulson 1975; Sanger et al. 1977). Several developments led to big improvements in usefulness and throughput of those initial technologies, namely synthetic dideoxy nucleotides (ddNTPs) (Chidgeavadze et al. 1984), Polymerase Chain Reaction (PCR)



**Figure 1. Three generations of DNA Sequencing Technologies.** The first two generations of sequencing technologies displayed here (Sanger and Illumina) are based on sequencing by synthesis of DNA and incorporation of fluorescent nucleotides. The main difference however is the massive parallelization achieved by second generation technologies. Third generation sequencing technologies like Nanopore sequencing don't share these features and enable longer read lengths often at single molecule resolution. (A) Sanger sequencing-based shotgun sequencing uses introduction of DNA fragments by cloning (1) and one-by-one sequencing of the clones using capillary gel electrophoresis and fluorescently labeled ddNTPS (2). (B) Next Generation Sequencing by Synthesis as commercialized by Illumina requires the ligation of universal adapter sequences (1) that enable hybridization of the fragments to a sequencing flow cell coated with complementary adapter sequences, followed by cluster amplification of molecules on the flow cell and reading of fluorescent nucleotides by imaging (2). (C) Third Generation Sequencing using a Nanopore as displayed here enables reading of longer fragments and native RNA. Similar to NGS, Nanopore requires the addition of adapter sequences that are in this case coupled to a motor protein (1) This motor protein attaches to Nanopore proteins bound to a membrane, unwinds the DNA and pushes the DNA molecule through the pore. Sequencing is performed by measuring the ionic current passing through the pore where each base leads to a characteristic change in ionic current(2). Created with BioRender.com

(Saiki et al. 1988) and Capillary Gel Electrophoresis (Swerdlow and Gesteland 1990; Luckey et al. 1990; Hunkapiller et al. 1991). The limitation of this technology was the maximum length of the sequenced fragment (~1,000 nucleotides) (Heather and Chain 2016) and was overcome by shotgun sequencing (Anderson 1981), i.e. by cloning DNA fragments into bacterial plasmids, sequencing the inserts and assembling larger contiguous sequences (contigs) using computational methods (Staden 1979) (Fig. 1 A). Together with the automatization and miniaturization of Sanger sequencing machines, these approaches increased the throughput immensely and led to the publication of the first human genome draft in 2001 (Lander et al. 2001; Venter et al. 2001) and the full assembly in 2004 (International Human Genome Sequencing Consortium 2004).

### **Second Generation Sequencing - more reads, lower costs**

While Sanger sequencing was essential for the success of the human genome project, sequencing continued to evolve, leading to a second wave of technologies that increased throughput and decreased costs (Mardis 2008; Dijk et al. 2014). This was fueled by the goal to sequence a whole human genome for less than 1,000 US\$ (Mardis 2006). Even though technologies differed in their approaches, the main improvement they shared was the massive parallelization that increased throughput from 96 simultaneously measured sequences to millions at a time. While other technologies, namely 454 Pyro Sequencing (Margulies et al. 2005) and SOLiD's ligation-based approach performed comparable in regards to throughput (Dijk et al. 2014), Illumina (formerly Solexa) took over the sequencing market and is currently the gold standard for most sequencing applications (Greenleaf and Sidow 2014) (Fig. 1 B).

### **Illumina Sequencing by Synthesis**

Illumina started out as one of many companies at the beginning of the next generation sequencing revolution and now is the most popular solution for most applications. This was enabled by continuously improving data quality, throughput and cost. Despite the technological improvements, the basic principle of Illumina Sequencing By Synthesis (SBS)

has not changed considerably. It features three main concepts, (1) clonal amplification of DNA fragments that are immobilized on a sequencing flow cell by hybridisation to grafted oligos, (2) fluorescent reversible terminator nucleotides, and (3) bridge amplification on the flow cell leading to DNA strand exchange. Using these main principles Illumina is able to overcome several challenges (Bentley et al. 2008). By spatially separating DNA fragments and amplifying them locally in so called clusters on the flow cell, they are able to increase the signal and distinguish different fragments. In a second step the fluorescently labeled nucleotides are incorporated by a polymerase one base at a time enabled by the reversible terminator technology. After each sequencing cycle the newly incorporated nucleotide is detected using high resolution imaging and identified by its characteristic fluorescent signal. As the molecules are immobilized on the flow cell, this process happens in a massively parallel manner and information on millions to billions of fragments is collected simultaneously. After each cycle the fluorescent group is cleaved off and the terminator is removed. This continues for typically between 35 and 300 cycles resulting in sequencing reads of the respective lengths. Next, a process called bridge amplification leads to reversing the direction of the molecules relative to the flow cell. Now, the sequencing by synthesis process is repeated reading the other end of the DNA fragment. This paired end sequencing is an important part of the method as it allows for better mapping to the genome, estimation of fragment sizes, and error correction when the two reads overlap. In 2017 Illumina produced over 90 % of world wide sequencing data, according to their own estimates (Illumina Inc. 2017)

### **Third generation sequencing - single molecules, long reads**

Illumina SBS's main limitation is the relatively short read length with a feasible maximum of ~300 nucleotides. Longer reads of up to ~600 bp are available on the MiSeq system, however, at a substantially higher cost. This might change with Illumina's announcement at the 2022 "Advances in Genome Biology and Technology Meeting" (AGBT) of a 600 cycle kit for the NextSeq system (Illumina Inc. 2022). While this is an improvement, it is still rather short. Read length might not seem like a big problem at first sight and, especially with the availability of a complete genome sequence, it is sufficient for many sequencing applications.

However especially for repetitive sequences and structural variation this poses a considerable computational challenge. This is where so called third generation sequencing approaches come into play. Namely the two approaches commercialized by Oxford Nanopore (Fig. 1 C) and Pacific Biosciences (Roberts et al. 2013; Branton et al. 2008; Lee et al. 2016). The two technologies share two main features that set them apart from second generation technologies. Firstly, both are single molecule readouts in contrast to Illumina where a cluster of clonal molecules is measured. Secondly, they are long read technologies with read length of up to 100,000 bases. The long reads are a big advantage for *de novo* genome assemblies of more species, but have also enabled to determine the full human genome sequence including the repetitive regions that could not be assembled in the human genome project (Nurk et al. 2022). Also, single molecule sequencing has clear advantages over previous technologies. The main draw back to this day however, remains the substantially higher cost and higher error rates (Lee et al. 2016; Cui et al. 2020). In addition, long reads and single molecule resolution are superfluous for most standard applications where accurate counting of short sequence tags is sufficient.

Today, sequencing technologies are essential tools for molecular biology, evolutionary biology and biomedicine (Koboldt et al. 2013). Large scale projects like the Vertebrate Genome Project aim to capture more of the diversity of life (Genome 10K Community of Scientists 2009; Rhie et al. 2021), cancer genome sequencing guides targeted therapies (Hutter and Zenklusen 2018) and long read sequencing of viral genomes has been an essential tool to monitor viral genome evolution during the COVID 19 pandemic (Oude Munnink et al. 2021). Sequencing costs still continue to decrease and new technologies emerge. Especially with the first Illumina patents running out, some companies try to challenge them by lowering the prices dramatically. Recently Ultima Genomics claimed to be able to sequence a whole human genome for 100 US\$, however, their claims remain to be independently validated (Almogly et al. 2022). Another example is Roswell Biotech who make use of semiconductor technology and aim to apply these molecular electronics to DNA sequencing (Fuller et al. 2022).



### 1.1.3 From genome to phenotype

The primary goal of the human genome project was to determine the sequence, but its promise has always been that with the complete sequence we would be able to understand human biology on a molecular level. This means to understand how a single genome can lead to the development of over 200 different cell types that our body consists of (National Center for Human Genome Research (U.S.) et al. 1990).

A first step in understanding the genome is to annotate it. The foundation to this was laid with Francis Crick's hypothesis that DNA is transcribed into RNA which in turn is translated into proteins and became known as the central dogma of biology (Crick 1970). Using the genetic code, it was thus possible to identify the over 20,000 protein coding genes and based on their transcription and translation we can measure where and when a particular gene or its product, a protein, is active.

While the central dogma still holds true and is widely accepted there are exceptions to the rule, particularly non-coding RNAs that perform functions without being translated (Stefani and Slack 2008). The ENCODE consortium that aimed to annotate the entire human genome found almost three quarters of the sequence to be transcribed (Djebali et al. 2012). However, it should be noted that expression does not equal function and many of the transcribed regions of the genome that they found are thought to be only spuriously expressed at very low levels (Graur et al. 2013). Many of the remaining sequences that are not associated with a particular function, albeit show evolutionary constraint, are thought to be gene regulatory elements. While each cell contains the same genome, they express different genes at different levels. These different sets of expressed genes lead to the production of proteins that ultimately functionally distinguish different cell types. A cellular phenotype is thus encoded in the genome and defined by the expression of different proteins and non-coding RNAs. Arguably, proteins have been studied for much longer than non-coding RNAs and are more interpretable in terms of their function. The most relevant phenotypic readout would thus be protein levels. However while DNA sequencing technology has improved immensely in the past decades and is a robust, flexible, sensitive and cost efficient technology, protein sequencing is much less optimized. This may in part be due to the pressure to innovate that

DNA sequencing has faced, but is likely also due to the more challenging biochemical nature of proteins over DNA and the lack of protein amplification methods (Alfaro et al. 2021). A good proxy for protein expression are, however, messenger ribonucleic acid (mRNA) levels.

## 1.2 Gene expression quantification - the era of RNA-sequencing

Gene expression quantification has been performed for many decades already and just like DNA sequencing it has developed from a single gene approach to whole transcriptome sequencing over the years. One of the first methods to quantitatively measure the amount, identity and length of a particular RNA in a sample was Northern Blot (Alwine et al. 1977). In this method, RNA is size separated using gel electrophoresis, transferred onto a membrane and hybridized to labeled DNA probes. Later, real time measurement of DNA amplification by PCR enabled accurate quantification of starting molecules (Higuchi et al. 1993). Coupled with reverse transcription prior to PCR, this can be used to accurately quantify RNA expression (Heid et al. 1996). Even though Reverse Transcription quantitative PCR (RT-qPCR) was a big improvement in terms of reproducibility and sensitivity (Wong and Medrano 2005) it is still a per gene approach. The first high throughput gene expression measurements of the whole transcriptome, were enabled by microarray technology, where RNAs are reverse transcribed and hybridized to complementary oligonucleotides spotted on a surface and quantitated by measuring the signal intensity of each spot (Schena et al. 1995; Schena et al. 1995; Wodicka et al. 1997). Finally, first sequencing-based approaches emerged that relied on counting short gene fragments called tags. Serial analysis of gene expression (SAGE) used restriction enzymes to create small cDNA fragments and concatenated them randomly by ligation. Sequencing these tags using Sanger sequencing provided quantitative data on gene expression, based on tag abundance and assignment to a gene (Velculescu et al. 1995). SAGE is thus the first occurrence of Tag sequencing for gene expression and can be seen as early predecessor of current end counting RNA-seq methods. Another iteration of this approach is Cap analysis of Gene Expression (CAGE) which particularly captures the 5'

## **1.2 Gene expression quantification - the era of RNA-sequencing 13**

---

ends of the transcript (Shiraki et al. 2003). CAGE was used extensively to map transcription start sites (TSS) (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014).

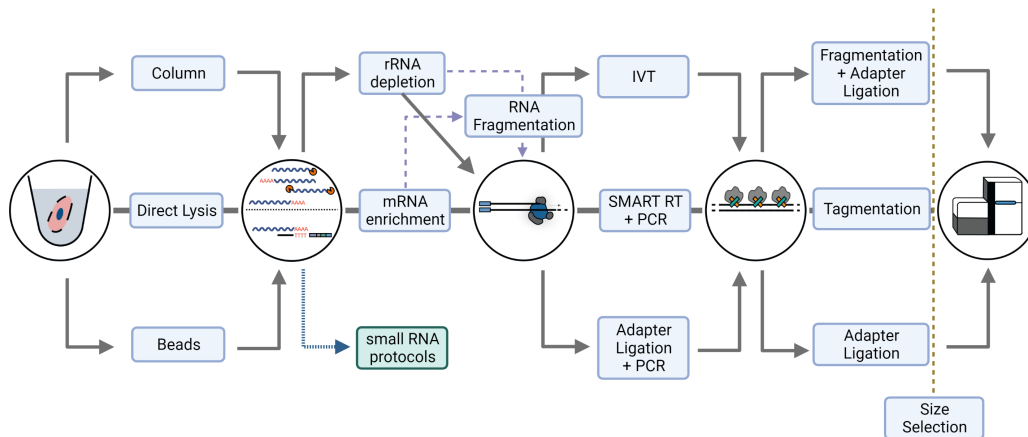
Whole transcriptome complementary DNA (cDNA) sequencing or more commonly known as RNA sequencing (RNA-seq) is based on reverse transcription (RT) of RNA into cDNA using a viral reverse transcription enzyme, subsequent second strand synthesis and sequencing adapter ligation (Fig. 2). This general workflow leads to a library of cDNA molecules that can be sequenced, aligned to the genome and finally quantified per gene or transcript. The general assumption being that the frequency of a given cDNA molecule in the library is reflective of its expression level. Advantages of RNA-seq over the previous gold standard technology expression microarrays are numerous: it is unbiased, does not require prior knowledge of expressed genes, is species agnostic, sensitive and allows for extremely high throughput (Marioni et al. 2008). As RNA-seq is sequencing-based, its cost have dropped rapidly with sequencing costs. In addition, optimized more cost-efficient methods for library preparation have been developed (Li et al. 2020; Alpern et al. 2019). In summary, RNA-seq has become the primary cellular phenotyping tool in recent years (Stark et al. 2019).

### **1.2.1 What we talk about when we talk about RNA-sequencing**

Most RNA-seq protocols follow the general workflow described above, however, a multitude of different protocol variants has been developed. For each step of the protocol different strategies are available that have their own set of advantages and disadvantages (Fig. 2).

#### **RNA isolation**

The first step in any RNA sequencing experiment is obtaining RNA. RNA can be purified before subjecting it to RNA-seq, using for example spin columns with silica membrane or extracted, using solid phase reversible immobilisation (SPRI) beads. Column-based isolation is the commercial standard, however, bead-based isolation is increasing in popularity (Oberacker et al. 2019). SPRI beads are paramagnetic particles coated with carboxyl residues



**Figure 2. Flavours of RNA-seq library preparation methods.** The "standard" RNA-seq workflow with different variations of RNA sequencing protocols. Key differences between protocols are the way RNA is isolated, how is dealt with rRNA and other RNA species, at which step the full length molecules are fragmented, how second strand synthesis is achieved and how sequencing adapters are added. Created with BioRender.com

that reversibly bind nucleic acids in the presence of Poly ethylene glycol (PEG) and high salt concentrations. In the absence of the the high salt concentrations binding is reversed and nucleic acids are released (DeAngelis et al. 1995). Thus, SPRI beads by default isolate total nucleic acids, which makes digesting DNA, by using a DNA specific nuclease, necessary to obtain pure RNA. A different approach is to start with crude lysate and skip RNA isolation, however this might have adverse effects on RT efficiency with higher cell numbers (Svec et al. 2013; Le et al. 2015).

### The whole transcriptome - total RNA, small RNA and mRNA

Although often referred to as whole transcriptome sequencing, most RNA-seq protocols either enrich for polyadenylated mRNA or deplete ribosomal RNA (rRNA). This step is necessary as only roughly 3-5 % of total cellular RNA in eukaryotes is mRNA, with the majority of it being ribosomal RNA followed by transfer RNA (tRNA). While rRNA and tRNA are an interesting field of study by themselves, they carry little to no information in regard to the cellular phenotype. Spending 95 % of sequencing reads on these molecules is thus to be avoided. A very useful feature of eukaryotic mRNAs and many long non coding RNAs

## 1.2 Gene expression quantification - the era of RNA-sequencing 15

---

(lncRNA) (Sun et al. 2018) is that they are polyadenylated upon transcription. This so called poly(A) tail can be used to selectively enrich for these molecules. However, other RNA species like small non-coding RNAs (sncRNAs) and some lncRNAs are not polyadenylated and thus missed with poly(A) enrichment methods. In addition their short length makes sncRNAs a challenging target for Illumina sequencing that usually requires fragments to be >200 bases long to be sequenced efficiently. Despite these limitations, the most commonly used RNA-seq methods use either poly(A) priming in the RT or hybridisation of the RNA to oligo(dT) coated beads and subsequent removal of unbound RNA.

Another option, that is especially important for prokaryotes that do not feature a poly(A) tail, is rRNA depletion. Two main options exist for removing rRNA, enzymatic digestion or hybridisation and removal of the most common rRNA species (O'Neil et al. 2013). While efficient protocols for this purpose have been developed (Wangsanuwat et al. 2020), the most commonly used rRNA depletion kits are prohibitively expensive (Janjic et al. 2022). On top of that, most researchers are still interested in the coding transcriptome and would thus have to sequence much deeper to get to the same coverage for protein coding transcripts (Zhao et al. 2018).

Finally, there is a set of methods for small RNA sequencing available that overcome the aforementioned size limitations and additional challenges in small RNA sequencing like 2'-O-methyl modifications at the 3' end of some sncRNAs that inhibit adapter ligation (Dard-Dascot et al. 2018).

### Second strand synthesis

A common step that is present in all RNA-seq protocols, even direct RNA sequencing which is possible only with third generation sequencing technologies (Garalde et al. 2018), is reverse transcription (RT). While transcription is a one way road in all DNA-based organisms, as postulated in the central dogma of molecular biology, some RNA viruses (retroviridae) have an RNA-dependent DNA Polymerase, also referred to as Reverse Transcriptase. Since its discovery in 1970 (Temin and Mizutani 1970; Baltimore 1970) it has earned their discoverers the Nobel price and has transformed molecular biology (Coffin 2021). In RNA sequencing,

RT is used to convert RNA into complementary DNA (cDNA). Usually reverse transcription is either primed using oligo(dT) primers that hybridize to the poly(A) tail or random hexamer primers, random six nucleotide long DNA fragments that bind all over the transcript. Like DNA-dependent DNA polymerases, RT enzymes move along their template and incorporate matching dNTPs into the growing DNA strand. This results in a double stranded RNA-DNA hybrid. Unlike DNA-dependent polymerases, RT enzymes do not generate a new template during this process that would enable synthesizing the second strand of DNA. However, DNA sequencing technologies require double stranded DNA as input, necessitating second strand synthesis. One traditional technique for second-strand synthesis is the removal of the template RNA using RNaseH and synthesis of the second strand using DNA polymerase I (Gubler and Hoffman 1983). This is followed by fragmentation, end repair, dA-tailing and sequencing adapter ligation (Agarwal et al. 2015). This however, leads to loss of strand specificity which is often overcome by incorporating dUTP instead of dTTP in the second strand, followed by dUTP specific digestion of the second strand after adapter ligation. A method that avoids conventional second strand synthesis ligates adapters specifically to either the 3' or 5' end of the to first strand cDNA, followed by PCR (Agarwal et al. 2015).

An alternative that has become popular in recent years is harnessing the terminal transferase activity of Moloney Murine Leukemia Virus Reverse Transcriptase (M-MLV RT). M-MLV RT adds non-template nucleotides at the 5' end of the transcripts (Zajac et al. 2013) which can be used for a process called template switching, where a template switching oligo (TSO), consisting of DNA and RNA bases, binds to the non-template cytosines and subsequently serves as a template for extension of the cDNA, adding a defined 5' end (Wulf et al. 2019). This Switch Mechanism at the 5' End of RNA Templates (SMART) enables the generation of full-length cDNA libraries and is highly sensitive (Zhu et al. 2001).

## **cDNA amplification**

While initial techniques used several micrograms of RNA as input, the need to generate RNA-seq data from small amounts of RNA became clear even before the advent of single cell RNA-seq (scRNA-seq). While cDNA synthesis can be performed with small amounts

## **1.2 Gene expression quantification - the era of RNA-sequencing 17**

---

of RNA, the resulting library would be too lowly concentrated for further processing and sequencing. To enable cDNA sequencing from small amounts of input, several strategies were developed that usually involve PCR amplification of the cDNA prior to library preparation. Generally this requires a universal sequence for primer binding on both ends of the cDNA. While the 3' sequence is usually introduced with the oligodT primer, a 5' sequence can be added by ligation or via the Template switch oligo in the SMART™ approach.

Another option is using a combination of RT followed by *in vitro* transcription (IVT) to amplify input RNA and a final RT with second strand synthesis. As IVT is a linear amplification process, this aims to introduce less bias compared to PCR (Eberwine et al. 1992). The main limitation here are the high input amount required for IVT (Hashimshony et al. 2012).

### **Illumina Sequencing library preparation**

Often already combined with second strand synthesis in standard RNA-seq methods, the goal of sequencing library preparation is to (1) add sequencing adapters that enable binding to the p5 and p7 grafted oligo nucleotides of Illumina sequencing flow cells and priming of the sequencing reads; (2) select for fragments between 200 and 1,000 base pairs length to enable efficient sequencing. Additionally while initial RNA-seq methods used at least one lane of sequencing per RNA-seq library (Marioni et al. 2008), with increases in data obtained from one lane and improvements in RNA-seq library composition such an approach would be excessive and multiplexing of many samples on one lane has become the standard. To demultiplex the individual samples after sequencing, sample specific DNA Barcode sequences on both ends of the fragments are introduced using PCR at this step.

Essentially two methods exist for library preparation, Fragmentation (either enzymatic or chemical), end repair, d(A)Tailing, phosphorylation and ligation of adapters, or Tn5 Transposase mediated fragmentation and ligation. These steps are usually followed by a PCR amplification of fragments carrying both 3' and 5' adapters, that adds sample barcodes via overhang primers to be read in the index reads of Illumina sequencing. Size selection is achieved using either SPRI beads, silica columns or agarose gel excision.

### 1.2.2 The next level - single cell RNA-sequencing

While the first wave of RNA-seq and RNA-seq development was driven by bulk sequencing it was shortly followed by the first single cell method (Tang et al. 2009). The usefulness of having single cell information is manifold and was recognized early on (Eberwine et al. 1992). Cells are the fundamental operational units of life and as such they are the most relevant level at which transcriptomes should be compared. Bulk information can be very useful for investigating differences between homogeneous cell populations but as soon as cellular heterogeneity is to be expected, single cell information is necessary to make the right comparisons. As expression levels in bulk RNA-seq represent mean expression over the whole population, subtle differences in a small fraction of cells or changes in the composition can often be missed (Sandberg 2014).

Two limitations for single cell RNA-seq are the minute input amounts that are usually in the low picogram range and the necessity to dissociate tissues into intact single cells before processing. While the latter is specific to the cell type or tissue under investigation and no general rules can be applied to overcome this challenge, low input amounts are a more general problem. Finally, the key to make single cell methods useful to the broader scientific community is cost efficiency. Only if several thousand cells can be processed at a reasonable cost, single cell RNA-seq can fulfill its promise of uncovering cellular heterogeneity. These technical challenges were overcome by several key innovations in the past years.

#### **Key innovations introduced by single cell methods**

In principle there are two ways to increase the amount of available input, amplification and pooling. While PCR amplification is generally suitable, it can introduce serious bias in the composition of molecules by preferential amplification of some molecules over the others. This problem becomes more pronounced the more cycles are performed and leads to incorrect gene expression levels and lower library complexity (Phipson et al. 2017). In addition, performing one PCR reaction per single cell is very tedious and costly (Ramsköld et al. 2012). The second strategy, pooling, brings input amounts back to bulk levels but



## 1.2 Gene expression quantification - the era of RNA-sequencing 19

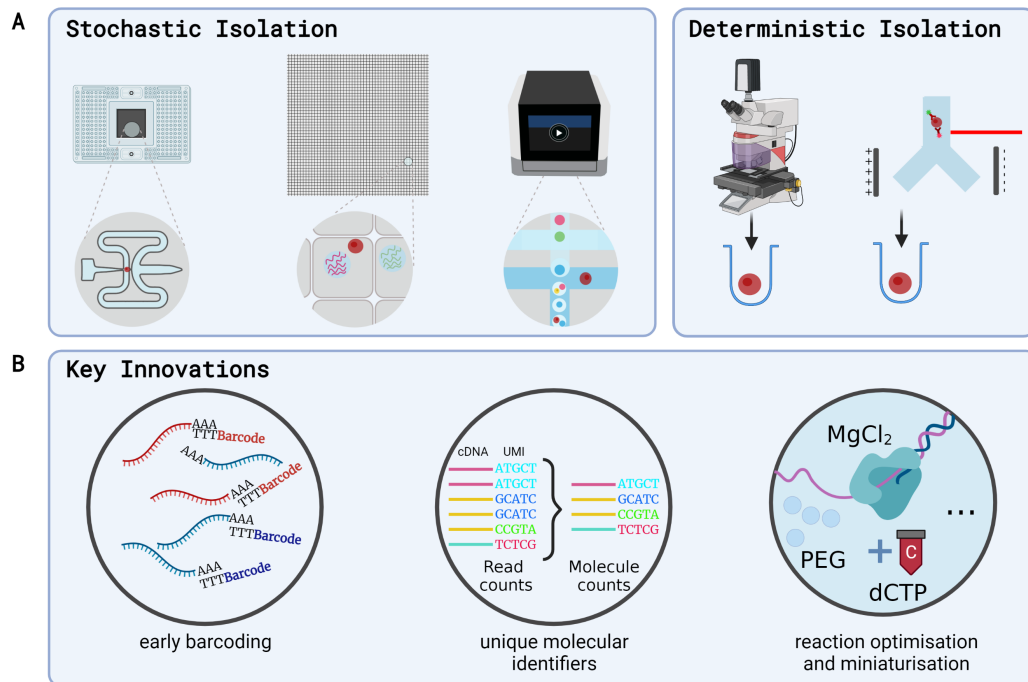
---

only makes sense if transcripts can be assigned back to the originating single cells in the end. The key innovation that made this possible is the introduction of a cell barcode during reverse transcription with the oligo(dT) primer or TSO and subsequent early pooling (Islam et al. 2011; Islam et al. 2014). In addition early pooling reduces the number of reactions to be performed per single cell drastically and thus decreased the library preparation costs immensely.

A different approach for amplifying the input is employed by the Cell Expression by Linear amplification and Sequencing (CEL-seq) methods (Hashimshony et al. 2012; Hashimshony et al. 2016). They use early barcoding of first strand cDNA as described above, followed by pooling, bulk IVT for linear amplification of the cDNA and finally a second round of RT including second strand synthesis.

To overcome the problem of amplification bias, researchers introduced a second random sequence in the reverse transcription reaction called Unique Molecular Identifiers (UMIs) (Kivioja et al. 2011). After sequencing, reads from the same cell that map to the same gene and have the same UMI sequence are assumed to originate from the same initial molecule and are thus collapsed. This reduces amplification artifacts to a minimum (Islam et al. 2014), however, it cannot recover molecules that were outcompeted during amplification and thus not detected later on. This last challenge was tackled by several groups that tried to improve existing protocols systematically by optimizing RT and PCR reaction conditions to capture more of the initial RNA molecules and ensure more even amplification (Picelli et al. 2013; Bagnoli et al. 2018; Sasagawa et al. 2018; Hughes et al. 2020).

Finally, a main difference between scRNA-seq methods that use early pooling and methods that perform one reaction per cell like the SMART-seq methods (Ramsköld et al. 2012; Picelli et al. 2013; Hagemann-Jensen et al. 2020) is the fact that early pooling methods generally only sequence either the 3' or 5' end of the transcript to capture the cell barcode. SMART-seq methods are full length, meaning reads from all over the gene body are sequenced. Both of these approaches have their own advantages and disadvantages and the choice of method depends on the scientific question.



**Figure 3. Single cell RNA-seq innovations.** (A) Different single cell isolation strategies. Integrated fluidic circuits (IFC) as commercialized by Fluidigm, trap cells in small chambers and subsequently deliver reagents directly to the wells. Nanowells are dense arrays of nanoscale wells where barcoded beads as well as cells are flushed over the surface and sink into the wells by gravity. Droplet microfluidics as commercialized by 10x Genomics aim to co-encapsulate barcoded beads with cells in water-in-oil emulsion droplets. Deterministic methods like laser capture micro dissection or Fluorescence Activated Cell Sorting (FACS) actively deposit cells in reaction vessels. (B) Key innovations introduced by single cell RNA-seq. Early barcoding of mRNA molecules by adding barcodes in the reverse transcription enables processing many cells at low cost. Unique molecular identifiers (UMI) tag each original RNA molecule with a random sequence. This enables distinguishing initial molecules from PCR duplicates and makes quantification more accurate. Adding particular reaction enhancers makes reactions more efficient in capturing more of the initial complexity and miniaturization reduces reagent costs. Created with BioRender.com

### Different strategies for isolation of single cells

Generally, two different strategies of cell isolation can be distinguished, stochastic and deterministic methods. The deterministic methods often use Fluorescence Activated Cell Sorting (FACS) or Laser-capture micro dissection to deposit single cells in reaction chambers containing lysis buffer (Ziegenhain et al. 2018). This ensures efficient use of resources, as each reaction chamber contains exactly one cell and enables enrichment for specific rare cell types. Conversely, these approaches are time consuming, require additional specialized equipment as well as expertise and, due to the lengthy cell handling, might introduce unwanted transcriptome changes and batch effects (Massoni-Badosa et al. 2020). The stochastic methods come in a variety of different forms, namely integrated fluidic circuits (IFC), droplet-based microfluidics, nanowell microfluidics and split-pool barcoding approaches. IFC have similar limitations in terms of throughput and cost as the deterministic approaches. Conversely the other stochastic methods often require a relatively high number of cells as input, don't enable enriching for certain populations and suffer from multiplets, empty reaction chambers and background RNA contamination. The great advantage of those methods is the high number of cells that can be processed in a single experiment, their mostly unbiased nature and the extremely low cost per cell (Ziegenhain et al. 2018).

Most popular for the last years and the current standard of single cell RNA-seq are droplet-based methods, namely the 10x Genomics Chromium platform (Zheng et al. 2017). The two original droplet-microfluidics methods Drop-seq (Macosko et al. 2015) and InDrops (Klein et al. 2015) were home-brew methods that can, in theory, be set up in most molecular biology labs. The basic principle is that single cells are stochastically co-encapsulated with beads or hydrogels that carry barcoded oligo(dT) primers in a water-in-oil emulsion droplet. After this droplet generation process, reverse transcription is performed within each droplet, transcripts from a single cell are uniquely barcoded and can be pooled by breaking the emulsion. The co-encapsulation in this process follows a poisson sampling and thus, to encapsulate on average just one cell, most droplets will have to be empty and some will carry 2-n cells. Nanowell techniques like Seq-well (Gierahn et al. 2017) or Microwell-seq (Han et al. 2018) generally have similar properties but with slightly more control over the

co-encapsulation (Prakadan et al. 2017). The technology that intrinsically allows for the highest number of cells profiled in a single experiment is combinatorial *in situ* barcoding of cells by splitting and pooling (Rosenberg et al. 2018; Cao et al. 2017). These technologies led to the exponential scaling of scRNA-seq throughput from tens of cells to currently up to 100,000 cells in a single experiment within just a few years (Svensson et al. 2018). Recently different strategies have tried to overcome some of those drawbacks of stochastic methods by combining combinatorial indexing and droplet-based cell isolation (Datlinger et al. 2021) or deterministic mRNA-capture bead and cell co-encapsulation (DisCo) (Bues et al. 2022).

### **The best single cell RNA-seq method and current state of the field**

Following the large increase in single cell RNA-seq methods, several groups have compared methods in regard to cost, accuracy of the measured gene expression levels, sensitivity, power to detect differential expression, the ability to distinguish different sub populations, throughput and other factors (Ziegenhain et al. 2017; Svensson et al. 2017; Ding et al. 2020; Mereu et al. 2020; Bagnoli et al. 2018). While these benchmarks arrived at different conclusions in regard to which method is the best, they have been important guides to select the right method for the particular scientific question at hand. Researchers continue to improve existing methods (Hagemann-Jensen et al. 2020; Hagemann-Jensen et al. 2022; Hughes et al. 2020; De Rop et al. 2022; *Scalable Single Cell Sequencing* 2021) and even though 10x Genomics is the current standard (Svensson et al. 2018) the fight for the best method is not yet over. However, while the first years of single cell RNA-seq were strongly driven by method development, the technologies are now mature enough to answer the particular biological questions that motivated their development.

#### **1.2.3 Computational analysis of gene expression data**

Early on in the human genome project, researchers realized that to cope with the quantities of data being produced, computational resources and methods to analyze high dimensional data would have to keep up (National Research Council (US) Chemical Sciences Roundtable 1999). For RNA-sequencing the plethora of methods can be subdivided into different areas, pre-

## 1.2 Gene expression quantification - the era of RNA-sequencing 23

---

processing, standard analysis methods including, but not limited to, differential expression analysis and single cell specific methods.

### Pre-processing RNA-seq data

While often taken for granted, the steps leading from raw sequencing data to a normalized expression matrix are complex and have great impact on downstream analysis (Vieth et al. 2019). The first step for most sequence data is quality filtering based on PHRED scores and adapter trimming, i.e. removing unwanted sequences like poly(A) stretches or flow cell adapters. Omitting this step can lead to serious problems like mismapping and other artifacts, especially in *de-novo* assembly methods (Coker and Davies 2004). The first RNA-seq specific step is the alignment of the reads to the genome or transcriptome, often also referred to as mapping. A particular challenge when mapping to the genome is that mRNA is usually spliced, since the genome sequence contains introns a read spanning two exons would thus map to two different genomic locations. Splice aware aligners like STAR take this into account (Dobin et al. 2013). Other approaches are *de-novo* transcriptome assembly (Hölzer and Marz 2019) and pseudo alignment methods (Patro et al. 2017; Bray et al. 2016). While pseudo alignment methods already include the gene expression quantification step, for alignment-based methods a separate counting step is necessary. Commonly used tools that summarize the number of times a sequencing read overlaps a feature/gene are featureCounts and HTseq (Liao et al. 2014; Anders et al. 2015). A major factor that influences the accuracy of counting is the quality of the reference annotation (Vieth et al. 2019). This goes so far that it might be preferable to map to a well annotated genome of a closely related species rather than to the native genome (Parekh et al. 2018). Especially for less well annotated genomes it might be advisable to extend the existing annotation to capture reads falling in potentially unannotated UTRs (Derr et al. 2016). However, even for the human genome it was shown recently that these strategies are beneficial (Pool et al. 2022). All these preprocessing steps lead to a count matrix, a simple sample by gene matrix with the summed up counts per gene, as final output.

## Normalization, Filtering and Outlier detection

One of the steps that has the biggest impact on downstream data analysis is the normalization of the data (Vieth et al. 2019). Consequently many methods have been developed over the years for this purpose. The main goal of normalization is to account for technical variation like sequencing depth or compositional differences in the input RNA pool. Classical depth normalization methods like counts per million (CPM) take care of the sequencing depth but neglect compositional biases. More sophisticated methods like edgeR's trimmed mean of M-values (TMM) or Median of Ratios (MR) as implemented in the DESeq2 package take this into account (Robinson and Oshlack 2010; Anders and Huber 2010; Love et al. 2014). Additionally dedicated methods try to explicitly remove technical variation, for example by using External RNA Controls Consortium (ERCC) spike-in molecules or control genes (Risso et al. 2014; Jiang et al. 2011; Leek 2014). Removing lowly expressed genes only present in a fraction of samples with low counts and removing outlier samples is important. A common strategy to find outliers is principal component analysis (PCA) where samples affected by e.g. high technical noise often explain much of the variance in the data and thus are separated from the remaining samples in the first two principal components (Chen et al. 2016).

Recently, normalization methods specifically for single cell RNA-seq have been developed to tackle particular challenges like the sparseness of the data, big differences in RNA content per cell and high variance between cells caused by for example cell cycle or transcriptional bursting (Vieth et al. 2019; Boeshaghi et al. 2022).

## Common downstream analyses

Even though RNA-seq can be used for diverse purposes, the most common use is differential expression analysis (DEA) between two or more groups of samples (Conesa et al. 2016). To test for significant expression differences, a generalized linear model is fit, based on a specified design matrix. Next, statistical testing and multiple testing correction of raw p-values are performed. The most commonly used tools for DEA are DESeq2 (Love et al. 2014) and limma (Ritchie et al. 2015). To increase interpretability of the resulting lists of DE genes, several approaches have been devised that associate the DE genes with databases containing

## **1.2 Gene expression quantification - the era of RNA-sequencing 25**

---

functional annotation. Popular tools are Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005), using Gene Ontologies (TopGO) (Gene Ontology Consortium 2021; Alexa and Rahnenfuhrer 2022) and Reactome Pathway Analysis (Yu and He 2016). While also for DE analysis specific single cell approaches have been developed, Soneson et al. found that classical general linear models-based (GLM) methods like limma perform well with single cell data in regard to power (Sonesson and Robinson 2018). One problem of these methods however is that pseudoreplication over cells from the same sample inflates power (Zimmerman et al. 2021), with pseudo bulk aggregation of single cells and mixed models as a potential solution (Zimmerman et al. 2021; Murphy and Skene 2022).

### **Single cell analysis methods**

In addition to the aforementioned single cell adaptations of bulk methods, there is a whole field dedicated to developing methods particularly for single cell RNA-seq. The scRNA-tools database currently contains over 1200 computational analysis methods (Zappia et al. 2018). Early on, new tools for dimensionality reduction and visualisation of the high dimensional data in 2D were developed, like t-distributed stochastic neighbor embedding (tSNE) (Van Maaten and Hinton 2008) and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018). With increasing amounts of bulk and single cell RNA-seq reference data sets available, many tools are dedicated to classifying cells into cell types based on their transcriptome (Abdelaal et al. 2019). Other important areas of method development include trajectory inference (Saelens et al. 2019), data integration (Luecken et al. 2022), clustering (Yu et al. 2022) and gene regulatory network inference (Pratapa et al. 2020).

### **Benchmarking and power simulations**

This vastness of RNA-seq methods and corresponding analysis tools calls for thorough comparisons to understand the strength and limitations of the different methods. One important prerequisite for meaningful benchmarks is a ground truth data set to use. This can be a real data set produced for the purpose of benchmarking (Tian et al. 2019) or a simulated data set computationally produced based on the properties of the real data (Vieth et al.

2017). One way to assess how well a particular method or tool is suited to detect differential expression are statistical power analyses. Firstly, differential expression is simulated based on the input parameters, secondly detected using DEA and finally the confusion matrix can be calculated including the False Discovery Rate (FDR) and True Positive Rate (TPR), also referred to as power. The power to detect DE depends on the amount of technical variance, reflected in the mean dispersion relationship of data from a homogeneous source, relative to the biological variance (Vieth et al. 2017).

### 1.2.4 Optimizing cost-efficiency of RNA-sequencing protocols

To develop better more robust and cost-efficient RNA-seq methods, we started from single-cell RNA barcoding and sequencing (SCRB-seq) (Soumillon et al. 2014) to develop molecular crowding SCRB-seq (mcSCRB-seq) (Bagnoli et al. 2018). The main innovation being, as the name suggests, the use of a molecular crowding agent. Molecular crowding is a process that can speed up reactions and thereby make them more efficient by increasing contact probabilities between reagents. This observation has first been described for blunt end ligation reactions (Zimmerman and Pheiffer 1983), but is also used in transposase based library preparation protocols (Picelli et al. 2014). By systematically optimizing the reverse transcription and pre-amplification PCR reactions for increased yield we improved gene detection and UMI counts compared to previous versions of the protocol (Bagnoli et al. 2018). Using power simulations and ERCC spike-in molecules we compared the sensitivity of our method to published data of other methods and found superior performance of mcSCRB-seq.

However, we wanted to further validate our method in a more realistic scenario, namely a direct comparison with other methods on a unified heterogeneous input sample (Mereu et al. 2020). The sample contained a mixture of human peripheral blood mononuclear cells (PBMCs), the Human Embryonic Kidney cell line (HEK293T) and a mouse colon sample. We and other method developers performed their protocol on these cells and the resulting data was jointly analyzed and compared to each other. While our method did not perform as well



## 1.2 Gene expression quantification - the era of RNA-sequencing 27

---

as expected due to unclear reasons this benchmark gave us directions for further improvement and highlighted the need to not only optimize for high sensitivity and yield but also for the power of a method to distinguish different sub populations. In addition, the improvements in 10x scRNA-seq methodology made the continued optimization of mcSCRB-seq less urgent and hence, we focused on adapting SCR B-seq to improve bulk RNA-seq methodologies.

Guided by the single cell method optimization and the demand for affordable efficient bulk RNA-seq we developed a version of SCR B-seq suitable for bulk inputs (Janjic et al. 2022). Namely, we introduced a direct RNA isolation step using magnetic beads, which enabled us to work with crude lysate as input. This has several advantages: It avoids costly RNA extraction using commercial kits, makes the method more amenable to high throughput processing without specialized equipment and makes input cell numbers as low as 1,000 cells feasible without particular adjustments. To show that this type of isolation does not lead to systematic biases or reduced sensitivity relative to the more standard column purification, we compared these two methods on three different input types. We found that the two methods performed similar for all parameters that we investigated, gene detection, library complexity and accuracy as measured using ERCC spike-in molecules. Finally, we compared the power to detect differential expression of our method prime-seq to the gold standard method TruSeq. We again saw comparable performance with power being a function of expression level and fold change, but most importantly the number of replicates per group. As the number of replicates used with a particular method is often restricted by the cost of library preparation and sequencing, we calculated the cost per sample for several commercial and non commercial methods. Using this information we compared the power of different methods to detect DE as a function of the budget. This metric clearly highlighted the big advantage of low cost methods like prime-seq over costly commercial methods like TruSeq. In addition the increased number of biological replicates that low cost methods enable will increase the reproducibility and statistical soundness of DE analyses (Li and Wang 2021).

## 1.3 Regulation of gene expression - The next frontier

As discussed above, quantifying gene expression levels using RNA-seq is currently the most powerful tool to measure complex cellular phenotypes, but how these phenotypes are encoded in the genome sequence can not be derived efficiently using RNA-seq alone. Unlike for coding sequences, where we can translate DNA sequence into protein sequence using the genetic triplet code, we can not readily interpret regulatory regions and predict their impact on gene expression.

A non-trivial first step is to identify regulatory regions and associate them to genes. For proximal, *cis*-regulatory regions (CRE) this is usually achieved by measuring chromatin accessibility and associating the open chromatin regions to the closest gene. Openness of chromatin can be measured genome wide by DNase-seq (Crawford et al. 2006) or Assay for Transposase Accessible Chromatin with sequencing (ATAC-seq) (Buenrostro et al. 2013). In addition, early on, regulatory sequences have been identified by sequence conservation e.g. between human and mouse (Loots et al. 2000). Another level of gene regulation are biochemical marks. Most prominently, cytosine methylation and histone modifications, measured typically using RRBS and ChIP-seq for particular biochemical marks. Even though DNA methylation and histone modifications can be relatively well interpreted, they are not predictive of enhancer function on their own. Once identified, one way to interpret putative CREs on a sequence level is by looking at transcription factor binding sites (TFBS) (Wasserman and Sandelin 2004). Transcription Factors (TF) are proteins that recognize certain DNA motifs and recruit the transcriptional machinery to the promoters of proximal genes, leading to their transcription (Lambert et al. 2018). However, looking for TF motif enrichment in a particular CRE is no definitive proof that this sequence leads to increased transcription when open and gives no information of how strongly it activates transcription. Based on chromosome conformation capture (3C) experiments, genome architecture has been linked to gene regulation and these technologies have led to a model where enhancers physically interact with promoters by looping, i.e. folding of the chromatin to bring promoter

and enhancer in close proximity. In addition chromatin domains have been found that lead to partial insulation of promoter enhancer pairs when located in different domains (Misteli 2020).

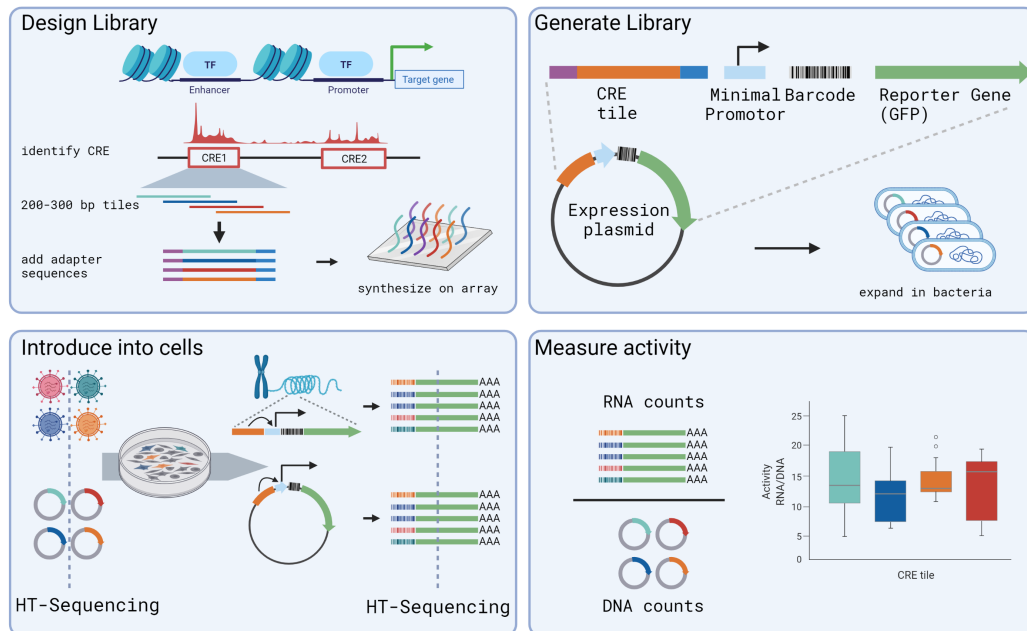
### 1.3.1 Measuring activity of regulatory regions is challenging

In order to directly measure the activity of a putative regulatory region, reporter assays have been used classically. In fact the most studied type of *cis*-regulatory elements, enhancers, were discovered using an episomal reporter assay and many of the defining properties were derived using this type of data (Banerji et al. 1981). Reporter assays use, as the name suggests, a reporter gene, often a fluorescent or luminescent protein, that indicates activity of the CRE to be measured. For this purpose, a CRE is cloned in front of e.g. a luciferase gene and the activity is quantified based on the strength of the luminescence. The bottleneck for this approach is that only one CRE can be measured at a time. A strategy that alleviates this problem are Massively Parallel Reporter Assays (MPRA). They enable measuring CRE activity for thousands of sequences in a single assay by switching from an optical to a sequencing-based readout.

### 1.3.2 Massively Parallel Reporter Assays measure CRE activity efficiently

Sequencing-based reporter assays exist in a variety of different forms, however, two approaches have been used the most and work in a similar fashion. Massively Parallel Reporter Assays (MPRA) use barcodes to identify the individual CREs, while the Self Transcribing Active Regulatory Region sequencing approach (STARR-seq) reads the CREs themselves as part of the transcript (Inoue and Ahituv 2015). In fact, both assays are referred to as MPRA in literature.

The first step in the workflow of an MPRA is the design of a library of sequences to be



**Figure 4. Workflow of Massively Parallel Reporter Assays.** First, regulatory regions to be tested have to be identified, split into overlapping tiles of 200-300 bp, adapter sequences for cloning have to be added and final constructs have to be synthesized on an oligo array. Next, the synthesized library has to be introduced into an expression plasmid containing the CRE tiles, a minimal promoter sequence, 10-50 unique barcode sequences per CRE and a reporter gene like GFP. These elements can be combined by classical cloning or using PCR and Gibson assembly. If a lentiviral system is to be used, an additional step is necessary to introduce the library into a suitable plasmid. Once the reporter library is prepared, its complexity is measured using high throughput sequencing. The two options for the introduction of the library are either lentiviral transduction with genomic integration or transfection of non-integrating episomal vectors. Depending on the experimental question, cell lines are transfected/transduced with the library and cultured. Additionally, stimuli can be applied to measure CRE activity under dynamic conditions. During this experiment the reporter gene will be transcribed more or less, according to how strongly the CRE drives expression. Finally, RNA is extracted and barcodes are counted by sequencing. Activity is then calculated per CRE as expressed barcode counts over total counts in the initial library. Created with BioRender.com

tested *in silico*. Due to length limitations in sequence synthesis, potential regulatory regions can not be assayed as a whole, but have to be broken up into pieces of around 200-300 bp (Klein et al. 2016). These sequences are usually tiled across the whole CRE with considerable overlap to increase resolution. The average length of a putative enhancer is ~420 bp (Mills et al. 2020), their length can frequently be over 1 kilobases (kb) (Li and Wunderlich 2017). Bound by the cost and capabilities of DNA synthesis, cloning and transfection efficiencies, the typical size of MPRA libraries is in the range of 10,000 to 100,000 different sequences. Notably, the library should also contain a number of scrambled control sequences that match the CRE tiles in regard to base composition (Ashuach et al. 2019). After DNA synthesis, the next step is to generate a highly complex plasmid library containing all CRE tiles in similar frequencies. To enable this high throughput library cloning, all sequences should be flanked by universal adapter sequences (Fig. 4, panel 1). One important step is to introduce DNA barcodes that uniquely identify the CRE tile. Those barcodes can either be synthesized with the CRE tiles or added by PCR afterwards. Early protocols designed specific barcodes as part of the synthesized oligo pool, however, this limited the assay in several regards (Melnikov et al. 2012). Mainly, it reduces the length of the sequence to be tested further due to the fixed total length for oligo synthesis. In addition, it was found that barcode sequences themselves can have impacts on activity measurements and thus bias the readout (Ulirsch et al. 2016; Lee et al. 2021). This limitation can be overcome by adding more barcodes for the same CRE and average the activity measurements over them. In fact, to properly account for this, each CRE should be represented by at least 10 barcodes (Ulirsch et al. 2016). If synthesized as fixed barcodes this effectively reduces the number of CRE tiles that can be tested 10-fold, as arrays are limited in the number of sequences that can be synthesized simultaneously. A more efficient solution to add many barcodes per CRE, is introducing random barcodes by PCR. As PCR is inherently biased and the process of adding the barcodes is stochastic, a complexity of at least 50 times the initial size of the CRE pool should be aimed for. This should result in CRE tiles having an average of 50 barcodes and at least 10 for the lowest ones.

Finally, the CREs have to be introduced into an expression plasmid containing a minimal promoter and a reporter gene e.g. GFP (Fig. 4, panel 2). Depending on the mode of delivery,

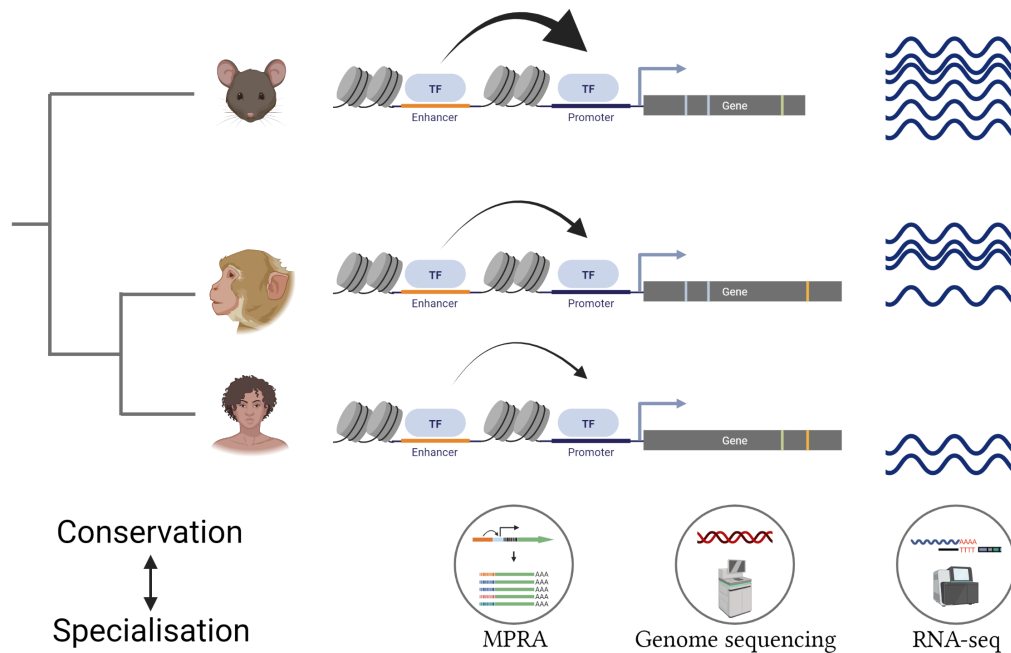
the final expression plasmid can be episomal or lentiviral. At this point, library complexity and uniformity are confirmed by high throughput sequencing to determine relative abundance of each CRE tile in the pool and potentially associate tiles to their barcodes if introduced by PCR. Different protocols have been published to make the cloning step straightforward and ensure high library complexity. They differ in the reporter gene used, the cloning strategy, the placement of the barcode and the CRE, as well as the sequencing library preparation (Klein et al. 2020).

For the actual reporter assay, the library is introduced into the cells of interest, i.e. cells in which the CRE is suspected to be active. Typically the cells are either transfected or lentivirally transduced for this purpose. Depending on the type of question, the cells are then stimulated or cultured for few days before they are harvested. In most cases, cells are subsequently lysed and RNA is extracted followed by sequencing and counting the transcribed barcodes (Fig. 4, panel 3). The activity is finally calculated as RNA counts divided by DNA counts and summarized over all barcodes per CRE tile (Fig. 4, panel 4). Another alternative that enables high throughput screening is sorting cells into bins based on their GFP positivity and counting representation of barcodes in each of these bins by sequencing. Based on the GFP signal and enrichment of a CRE in the expression bins, activity is inferred (Sharon et al. 2012).

## 1.4 Applying RNA-seq and MPRAs to measure gene expression evolution across primates

The first cross species comparisons of DNA date back to the discovery of different ratios of the four nucleotides between different species by Chargaff and colleagues (Vischer and Chargaff 1948). While their method just compared genome wide fractions, it was already enough to distinguish different species. For the next five decades, researchers compared sequences on a per gene level until the era of genomics (Felsenstein 1985). The power of

comparative approaches in genomics has been recognized early on and was one of the reasons for sequencing the mouse genome (Loots et al. 2000; Mouse Genome Sequencing Consortium et al. 2002). Using evolutionary information i.e. sequence conservation, regulatory elements could be identified and annotated (Mouse Genome Sequencing Consortium et al. 2002; Jegga et al. 2002).



**Figure 5. Comparative genomics on three levels.** Comparative genomics can be performed on multiple levels, coding sequences, regulatory sequences and gene expression. Generally, inferences about either conservation or derivedness/specialization of a feature can be made on the basis of these comparisons. Both evolution of coding sequences and expression evolution have been recognized as an informative predictor of function. Expression evolution can be measured on the level of the output i.e. gene expression or on the level of the regulatory landscape leading to a particular transcriptional output. Created with BioRender.com

Due to its role as the major animal model for humans, the mouse genome has been extensively characterized using comparative genomics (Breschi et al. 2017). Although the mouse is a very useful model, non-human primates can help to bridge the phylogenetic gap between human and mouse. This is particularly important for biomedical purposes (Enard 2012), but also studies of the evolution of gene expression. Already some of the

first comparative studies using primates, showed their huge potential to identify human specific innovations (Enard et al. 2002b). While comparative genomics was initially limited by the availability of genome sequences, soon after the human genome, sequencing of the chimpanzee genome followed (Chimpanzee Sequencing and Analysis Consortium 2005) and by 2009 several key primate species had been sequenced (Marques-Bonet et al. 2009). By now, genomes of 57 primates are available, nine of which are considered high quality assemblies (Kuderna et al. 2020).

### 1.4.1 Gene expression is encoded in the genome

While much of the effort in primate genomics was focused on sequence comparisons and annotation (Rogers and Gibbs 2014) the importance of gene regulation has historically been acknowledged (Britten and Davidson 1971; Britten and Davidson 1969) and argued for to explain the rather small differences in protein coding sequences and cross species DNA hybridization between humans and chimpanzee (King and Wilson 1975). The large phenotypic differences between species, but also between cell types within species, can be understood by comparing gene expression patterns across development, cell types and conditions. As outlined above gene expression can not readily be computed from the DNA sequence and is dependent on the activity of gene regulatory elements. Expression is an intermediate phenotype that can be used to compare stable, genetic, differences in gene regulation between species and importantly can be measured easily.

Early on, gene expression between primate species was compared using microarrays (Enard et al. 2002a), however, RNA-seq has been a transformative tool to study gene expression evolution, as it is species agnostic and enables studying more distantly related species (Brawand et al. 2011; Levin et al. 2016). In addition to RNA-seq, other high throughput assays, mainly ChIP-seq, ATAC-seq and RRBS but also MPRA have been employed in a comparative framework (Klughammer et al. 2022; García-Pérez et al. 2021; Uebbing et al. 2021; Whalen et al. 2022; Klein et al. 2018b). In summary, comparative gene expression evolution has been a very successful strategy to better understand human health and disease as reviewed extensively (Romero et al. 2012; Enard 2014; Khaitovich et al. 2006).



### 1.4.2 Induced pluripotent stem cells are an invaluable tool for comparative gene expression evolution

However, one common challenge for comparative studies, is the availability of comparable samples. While the origin of the DNA that is sequenced is largely irrelevant for genome comparisons, e.g. recently sequence comparisons of DNA extracted from feces revealed population dynamics in chimpanzees (Fontsero et al. 2022), gene expression is cell type and state specific. Obtaining comparable samples for several individuals and species is much more challenging, especially as access to primates for such experiments is restricted for ethical but also conservation reasons (Kuderna et al. 2020). Soon after their development, induced pluripotent stem cells (iPSCs) have been recognized as a potential solution, as they are comparable between species, can be derived from diverse primary cell sources, can be cultured indefinitely *in vitro* and can be differentiated into most cell types (Takahashi and Yamanaka 2006; Wunderlich et al. 2014). However, currently only few cell lines are available publicly and primary cells for reprogramming are still sparse.

#### A non-invasive primary cell source for the generation of primate iPSCs

Non-invasive sampling is a practical solution. Particularly, urine derived primary cells have been shown to reprogram efficiently for humans once isolated (Zhou et al. 2011). I contributed to the development of a protocol to derive primate iPSCs from urine, sampled in an unsterile environment (Geuder et al. 2021). Already as little as five milliliters of urine are sufficient to isolate primary cells and the addition of a broad spectrum antibiotic agent efficiently removed contamination. In this manner, iPSCs of human as well as orangutan and gorilla could be generated. I contributed to this project by using the bulk RNA-seq protocol prime-seq to characterize both the primary cells as well as the iPSCs. Transcriptome characterization using prime-seq is a low cost alternative to other costly methods to assess pluripotency.

### 1.4.3 Brain size and folding are among the most diverged traits in mammals

One promise of comparative genomics is to lead to a better understanding of human specific traits such as the increase in brain size and folding of the neocortex in the human lineage (Enard 2015; Enard 2014; Enard 2016). When comparing the brain between species, several measures can be compared, for example brain mass, the ratio of brain to body mass or gyrification i.e. foldedness of the cortex (Jerison 1961; Zilles et al. 1989). Gyrification is a mammalian innovation and is found in several branches of the mammalian phylogeny at varying degrees (Lewitus et al. 2014). It is thought to have been present in the most recent common ancestor of all mammals (O’Leary et al. 2013), but was lost subsequently in some species like the mouse that is completely lissencephalic (smooth brained) (Kelava et al. 2013). Gyrification is usually summarized in the Gyrification Index (GI) and measured as the overall cortical surface area, including folds, divided by the outer surface area, excluding folds. This metric ranges between GI=1, indicating complete lissencephaly and up to GI=5.55 at the most extreme (Sun and Hevner 2014), with values greater than one indicating increasing degrees of folding. Both, brain mass and GI vary considerably throughout the mammalian and primate phylogeny (Lewitus et al. 2014; Sun and Hevner 2014) and are thus an interesting trait for evolutionary analyses.

#### The case of *TRNP1*

Cell biologically, the increase in cortex size and folding is relatively well understood and can be well explained by differences in the number of cortical progenitor cells and their mode of division (Espinós et al. 2022). However, the genetic basis of the evolutionary switches that alter those parameters are not as well understood. Despite several human or primate specific genes that have been linked to the phenotype, human or primate specific neofunctionalization of paralogs can not explain gyrification in other mammalian branches, let alone the last common ancestor of mammals.

In 2009, researchers found that a knock down of the TMF-regulated nuclear protein 1

(*TRNP1*) in mice can induce folding of the mouse cortex. Furthermore, *TRNP1* leads to a switch between symmetric and asymmetric cell division of basal radial glia cells. *TRNP1* originally has been described as a highly conserved gene in mammals that drives proliferation (Volpe et al. 2006). Recent molecular characterization revealed that expression of *TRNP1* leads to shortening of the cell cycle by phase separation (Esgleas et al. 2020). Increased *TRNP1* expression in human basal progenitor cells is mediated by higher levels of histone H3 acetylation proximal to the promoter and leads to increased proliferation in humans relative to the mouse (Kerimoglu et al. 2021). Expansion of this population of progenitor cells is associated with increased cortical folding (Espinós et al. 2022). In addition, studies in ferret found *TRNP1* to be expressed in a precise temporal and spacial pattern during cortical development (Martínez-Martínez et al. 2016).

#### **1.4.4 Co-evolution of regulatory sequences of *TRNP1* with cortical folding**

While the aforementioned research links *TRNP1* expression to gyrification, the evolutionary perspective can help to shed light on which parts of the coding and regulatory sequences of *TRNP1* co-evolved with brain size and gyrification. To investigate this in a high throughput manner, we devised an approach to associate changes in the sequence to the different brain phenotypes in a phylogenetically aware manner. Towards this goal, I contributed by performing an MPRA on five putative regulatory regions of *TRNP1*. The putative CREs were defined by DNase-seq of fetal human and mouse brain. In addition to the human and mouse sequences, orthologous regulatory regions of 33 mammalian species were included. Transcription enhancing activity of these sequences was assayed in iPSC-derived human and cynomolgus macaque neural progenitor cells. Measured CRE tile activities were normalized and summarized per region and associated with brain phenotypes using a phylogeny aware regression model (phylogenetic generalized least squares, PGLS). We found that one putative CRE, located in the first intron of *TRNP1*, was significantly associated to gyrification, but not brain size or body mass, within old world monkeys and great apes. Combining MPRA

with our bulk RNA-seq method prime-seq, enabled us to not only measure CRE tile activity but also the trans environment, i.e. transcription factor expression within the assayed cells. Using motive enrichment analysis on the intron 1 CRE of all assayed catarrhine monkeys, revealed an association of the brain phenotypes with the presence of three transcription factor binding motives. In addition we found that the coding sequence of *TRNP1* co-evolved with brain size, and that *TRNP1* orthologs from species with bigger brains drive proliferation of E14 mouse neural stem cells more strongly than *TRNP1* orthologs from species with smaller, less gyrified brains.

### **1.4.5 Measuring primate gene expression evolution using high throughput transcriptomics and massively parallel reporter assays**

In summary, in this thesis I have elaborated on how the sequencing revolution has enabled modern molecular biology. Particularly, studying gene expression and regulation is now possible at unprecedented scale. Gene expression can be measured transcriptome wide using RNA-sequencing of populations of cells or even individual cells. Especially single cell RNA-sequencing has undergone continuous systematic optimizations, leading to the high throughput, high content technologies available today. Towards these methodological advances, I have contributed by co-developing the mcSCRB-seq protocol that for the first time introduced molecular crowding as a means to increase RT efficiency (Bagnoli et al. 2018). I have further participated in a single cell RNA-seq benchmarking effort that aimed to find the technologies most suited for the human cell atlas (Mereu et al. 2020).

While single cell RNA-seq technologies are one of the greatest recent technological innovations, bulk methods that assay 10,000s of cells per sample are a powerful and robust tool and often the better choice compared to single cell technologies. To make bulk RNA-seq available to any molecular biology lab, I, together with co-first author Aleksandar Janjic, have developed prime-seq, a bulk RNA-seq method that is cost-efficient and enables processing more samples at a similar resolution compared to standard methods (Janjic et al. 2022).

Next, I highlighted the challenges of linking gene regulation to the gene expression phenotype observed with RNA-seq and NGS-based massively parallel reporter assays as an emerging tool to study enhancer activity at scale. Applying these tools in a comparative evolution framework can help to better understand evolution of gene expression. In particular, I have used prime-seq to characterize urine derived induced pluripotent stem cells of different primate species, human, gorilla and orangutan. iPSCs are an important tool for comparative studies as they are a renewable, comparable and versatile source of pluripotent and differentiated cells (Geuder et al. 2021). Subsequently I used Neural Progenitor Cells (NPC) derived from the iPSCs described in Geuder et al. 2021 in a comparative study of brain size evolution. In this study we compared the coding and regulatory sequences of the gene *TRNP1*, implicated in cortical folding, across mammals and found co-evolution with brain phenotypes. I contributed by measuring activity of putative CRE sequences of *TRNP1* from 45 mammalian species using an MPRA combined with bulk RNA-seq (Kliesmete et al. 2021).



## 2 | Discussion

### 2.1 Developing and Improving RNA-sequencing methods

RNA-seq methods have been improved immensely over the last decade, leading to increased sensitivity, information content, robustness and reduced costs (Janjic et al. 2022). These improvements led to the adoption of this technology in almost all areas of biology. Especially single cell RNA-seq method developments have helped push the boundaries continuously at an incredible pace. Recently, we and others transferred those improvements to bulk RNA-seq, bringing it one step closer to being a standard technique that any molecular biology lab independent of budget and specialized expertise and equipment can use (Janjic et al. 2022). One aspect that should not be neglected in this regard is that data analysis has to become just as standard for the method to be accessible to every researcher. While there is no such thing as a standard analysis, much effort is put into standardizing parts of the analysis workflow that can then be used in a modular way to tailor the analysis to the scientific question (e.g. Zhang and Jonassen 2020; Spinozzi et al. 2020). In addition, as data analysis is being integrated into the curriculum of bachelor and master programs, the current generation of graduate students is much more proficient in processing this kind of data. So is RNA-seq method development at a point where further optimization is futile? While methods are suitable now to analyze large sample sizes with relatively limited bias, there is still plenty of room for improvement.

### 2.1.1 The issue of barcode swapping

Persistent limitations particularly for single cell RNA-seq are that methods are either optimal in terms of cost and throughput or data quality. High throughput methods like droplet-based methods or split-pool approaches suffer from background noise, multiplets and lower gene detection compared to the leading methods (Ding et al. 2020). High content methods like the SMART-seq host of methods with its most recent iteration SMART-seq3xpress (Hagemann-Jensen et al. 2022), on the other hand, are still relatively expensive and much less accessible, despite great improvements in regard to cost enabled by miniaturization. To reach a throughput, even remotely in the range of the high throughput methods, considerable expertise and non standard equipment like liquid handling robots are required.

One area that has been neglected in method development is the phenomenon of noise introduced by barcode swapping. This is particularly a problem of methods that use pooled amplification by PCR, where chimeric PCR products can arise. These chimeric PCR products lead to the misassignment of transcripts to cells in cases where the 3' barcode sequence of one transcript gets swapped (Dixit 2021). In addition, even for switching events within the same barcode group, this phenomenon will lead to inflated molecule counts. Methods usually test for this kind of noise implicitly using human/mouse mixture experiments, where the fraction of reads assigned to a species per cell gives information about potential doublets but also barcode swapping. Another way to look at it is by investigating how well a method can distinguish between different cell types, a property often visualized using dimensionality reduction and evaluated based on clustering, marker genes and silhouette scores. In a recent benchmark study of single cell RNA-seq methods for cell atlas building this "clusterability" was actually a decisive factor (Mereu et al. 2020). One reason for the poor performance of some methods was potentially cross contamination in the form of barcode swapping or background RNA. While background RNA is a likely explanation for cross contamination in stochastic isolation methods it is less so for methods that rely on deterministic isolation. One more direct way to measure barcode swapping is by using genetic information to distinguish transcripts coming from different individuals. Using methods like cellSNP, designed for deconvolution of donors in pooled single cell RNA-seq experiments, it is possible to more



accurately quantify this phenomenon, however, as mentioned before, this signal is usually convoluted with background RNA (Huang et al. 2019). One method that explicitly models barcode swapping and background RNA contamination independently is cellbender (Fleming et al. 2019). Adapting it to investigate barcode swapping could be a potential first step in quantifying the effect of this contamination across methods. While it is important to better understand this artifact, methods have been optimized for high clustering accuracy and marker gene expression and thus low levels of barcode swapping. In contrast, using gene detection and transcript counts as main optimization parameter might favour conditions with more barcode swapping as it leads to higher gene counts, more uniform gene detection and high UMI counts. Recently, researchers could show that one method optimization that claimed big improvements in gene detection over previous methods was purely artifactual. They were able to show this experimentally by using a new type of spike-in molecules that contain UMIs called "molecular spikes". (Ziegenhain et al. 2022). In the future, this approach can also help identify barcode swapping more readily and enable direct optimization towards lower barcode swapping frequencies.

### **2.1.2 Using statistical power analysis for method development**

A strategy that we have used extensively in the past to compare different methods or parameters are statistical power simulations (Ziegenhain et al. 2017; Bagnoli et al. 2018; Vieth et al. 2019; Janjic et al. 2022). Power simulations are very useful as they integrate different aspects of a method, i.e. gene detection, variability, mean expression, as well as dynamic range and translate them into a tangible outcome, the power to detect differential expression (Vieth et al. 2017). However, methods with high rates of barcode swapping will perform better in terms of power, as swapping might shrink technical variance and increase gene detection. One caveat of this approach is that it assumes independence of each single cell leading to pseudo replication in the DE testing. However, when doing relative comparisons between methods, power to detect DE should still be an informative criterion. To enable

more realistic power simulations different populations of cells and between-cell biological variability have to be modeled explicitly.

In summary, single cell RNA-seq methods are far from optimal and still under active development and while it might seem that 10x Genomics is currently the only reasonable choice for scRNA-seq, this can change very quickly. The tireless efforts of many researchers have the potential to further improve single cell RNA-seq methods. Already now, 10x Genomics' Chromium platform has many competitors, e.g. ParseBio and Fluent Biosciences (*Scalable Single Cell Sequencing* 2021; Clark et al. 2022), and will only be able to keep its spot as the top method if it keeps innovating and improving.

### 2.1.3 The perfect RNA-seq method does not exist

Benchmarking studies have contributed immensely to narrow down the list of methods most suitable to answer particular biological questions. However, the variability of questions that can be investigated with RNA-seq is vast and there is no "one size fits all" solution that performs well for all types of questions. Generally, RNA-seq methods have been shaped largely by the sequencing technology. Examples are the typical fragment length, the need for amplification and even the need to reverse transcribe RNA in the first place. Third generation sequencing enables real time long read sequencing, identification of mutations, mapping of transcriptional start sites, single molecule resolution and allelic information (Murphy and Skene 2022). Some methods even allow for direct sequencing of RNA (Garalde et al. 2018). Nanopore-based sequencing is getting more affordable and accurate at the same time and might soon be at a similar per base cost as Illumina sequencing. So will third generation sequencing methods take over and change the way we do RNA-seq in its wake? The answer is likely no. Even if long read native RNA-sequencing becomes feasible at the scale of current cDNA-based RNA-seq, it is still less efficient in terms of resources. Short read sequencing is so successful because it generates just enough information to very efficiently and accurately count. In most cases, the sequence is not relevant beyond being assignable to the genomic position it originated from. Jay Shendure and coauthors made this point very clear by differentiating between sequencing and molecule counting (Shendure

et al. 2017). Sequencing is not necessary for most current applications where the goal is to accurately quantify the number of molecules. For applications where accurate counting of genes and transcripts for quantitative comparisons is the goal, current technology is near optimal. Currently, efforts that aim to annotate transcript isoforms or genes in non-model species use long read sequencing, particularly PacBio Iso-seq, but are often complemented with short read RNA-seq for quantification (Ferrández-Peral et al. 2021). The diversity of methods designed for particular applications will likely increase further in the future with the rise of epitranscriptomics, i.e. detection of RNA modifications, and investigations of post-transcriptional processes. Recently a study compared transcription levels, measured by RNA-seq, and translation levels, measured by Ribosome profiling and sequencing (Ribo-seq), in an evolutionary context. They found more tight control of translation levels compared to transcription levels across species, hinting towards conserved mechanisms to buffer changes in expression levels towards stable translation levels and thus protein expression (Wang et al. 2020).

### **2.1.4 Beyond RNA-seq**

Most developments in recent years have not focused on (single cell) RNA-seq itself but on combining it with other modalities, like single cell ATAC-seq, CRISPR perturbation screens, protein expression using antibodies, B-/T-cell receptor sequencing, detection of expressed mutations and additional levels of multiplexing (Lee et al. 2020). Some protocols even combine scRNA-seq with more than one additional layer of information (Clark et al. 2018). In addition to these multiomics measurements, spatial transcriptomics has been a big topic in the field for years and is starting to become more standard. Spatial transcriptomics, where cells are profiled in their native tissue context, is subjected to massive method development, both commercial as well as academic. The technologies are still sub-optimal in one or the other regard, either in terms of resolution or the need for expensive specialized equipment (Stark et al. 2019). However, the potential advantages are clear, and in addition to the more obvious advantages of having a spatial dimension to gene expression, there are some additional strengths. One aspect is for example that they enable measuring the transcriptomes of many

cells at a time with large capture areas, often containing millions of cells. In addition, they are less biased as they do not suffer from dissociation induced artifacts and enable profiling cells of all shapes and sizes.

## 2.2 Measuring and linking enhancer activity at scale

Much effort has been put into cataloging enhancers and understanding the code that underlies gene regulation. Many different assays have been developed for this purpose most of which provide indirect evidence of a DNA element participating in regulating gene expression. Methods that fall into this class are assays that measure accessibility of chromatin like ATAC-seq or DNase-seq (Buenrostro et al. 2013; Crawford et al. 2006). While both openness and proximity of regulatory and regulated sequences is a prerequisite, it is by no means a prove for enhancing activity. Especially linking potentially active sequences to targets is not possible with these methods and thus always relies on arbitrary proximity-based thresholds. A Key to overcome this association problem have been chromatin confirmation capture-based assays like HiC or ChIA-PET, that map three dimensional genome confirmation (Gasperini et al. 2020). Finally, ChIP-seq and similar methods associate histone modifications and other kinds of biochemical marks to particular regulatory states (Stillman 2018).

It has been shown that histones and histone modifications are highly conserved in eukaryotes (Grau-Bové et al. 2022). However, while correlated with particular states, the presence or absence of certain histone marks does not provide direct evidence that a sequence acts as enhancer. In addition, biochemical marks can not give quantitative information on how strongly a particular CRE enhances gene expression. More direct evidence can come from MPRA or CRISPR/Cas9-based enhancer perturbation screens.

### 2.2.1 Limitations and strength of MPRA

MPRAs have been discussed extensively as a potential approach to overcome a lot of the aforementioned limitations as they are the only high throughput assay that directly measure activity of a potential enhancer sequence (Inoue and Ahituv 2015). They are undoubtedly useful but at the same time limited in many regards. Firstly, they are limited by DNA synthesis abilities, secondly, MPRA are highly artificial *in vitro* assays that have biases and fail to capture higher order interactions.

### DNA synthesis has revolutionized biology but longer sequences at lower costs are needed

While DNA synthesis is as important to molecular biology as sequencing, or potentially even more, there has been less progress compared to sequencing technologies. Synthetic stretches of DNA often called oligonucleotides are essential i.e. for PCR where they are used as primers, for CRISPR/Cas9 gene editing where they are used as guide RNA, but also for sequencing, gene therapy, mRNA vaccines and many more. These applications, however, usually require only short sequences which might be part of the reason for the lack of innovation because short sequences can be synthesized efficiently with existing technology. The current gold standard for DNA synthesis is phosphoramidite chemistry-based synthesis on columns or microarrays (Song et al. 2021; Kosuri and Church 2014). Synthetic biology laboratories that aim to generate whole chromosomes from scratch would profit most from these synthesis capabilities, but also DNA data storage technology is in dire need for improved DNA synthesis (Song et al. 2021). New approaches that are currently under active development to enable these endeavors are based on enzymatic synthesis (Eisenstein 2020). MPRA represent another occasion where efficient high throughput synthesis of hundreds of thousands of sequences with lengths over 1kb are desirable. The human genome contains around 1 million putative enhancer sequences and thus ~50 times the number of genes (Mills et al. 2020). Enhancer sequences amount to 426 million base pairs in total. Considering that current technologies are limited to 300 bps, tiling requires considerable overlap to avoid edge effects and adapter

sequences have to be synthesized as well, the effective length is reduced to ~100 bp per tile. Testing all enhancer sequences in one assay would thus require the synthesis of over 5 million 300 bp sequences and subsequently cloning them with on average 50 unique barcodes resulting in a total of at least 250 mio unique sequences. This is in principle possible but completely infeasible. Strategies for the assembly of short sequences into more physiological lengths have been devised but suffer from higher error rates (Klein et al. 2016). A systematic comparison of MPRAs assessed the impact of enhancer length and orientation on activity scores. They found only slight impact of the enhancer orientation as defined in the original description of enhancers and their properties (Banerji et al. 1981). Consistent with the notion that MPRAs are limited by DNA synthesis technology, they found considerable differences when comparing scores of short CRE tiles with their longer counterparts. They highlight that longer sequences capture more of the true biological signal (Klein et al. 2020).

Another way to overcome the limitations of DNA synthesis is by avoiding it all together. This is the approach taken by the STARR-seq MPRAs (Arnold et al. 2013), where instead of synthesizing enhancer sequences they are often derived from genomic DNA. While this allows for a much higher number of sequences to be tested, it is much less controlled and does not allow for a pre-selection of sequences to test. Enrichment approaches combine STARR-seq with other approaches that enrich for putative enhancer sequences by hybridization, multiplexed PCR, bacterial artificial chromosome (BAC) libraries or ChIP-seq (Vanhille et al. 2015; Vockley et al. 2015; Muerdter et al. 2018; Vockley et al. 2016). One particularly interesting approach enriched for open chromatin regions using ATAC-seq and subsequently cloned the transposase accessible DNA into a STARR-seq plasmid. This not only ensures to sample regions of open chromatin in the cell type of interest but does so in a quantitative manner where regions that are more open, i.e. open in a higher proportion of cells, are more prevalent in the pool (Wang et al. 2018).

### MPRAs can measure sequence activity but lack a physiological chromatin environment

A major shortcoming of all MPRAs including STARR-seq is the *in vitro* nature of the assay that is completely devoid of genomic context, interaction with other enhancers and particularly the promoter. Several studies have tried to quantify these biases and overcome them. Muerdter et al. found leaky activation of the reporter by the bacterial origin of replication and subsequently showed that the bacterial origin of replication (ORI) in complete absence of a core promoter provides a less biased readout. In addition, they found an interferon- $\gamma$  response after introducing the exogenous DNA (Muerdter et al. 2018). An approach to overcome the episomal nature of the assay using integrating lentiviral vectors for delivery of the reporter library was developed and termed lentiMPRA (Inoue et al. 2017). However, the integration into the genome is random and thus not representative of the native chromosomal environment of the CRE. To test the impact of integration on enhancer activity estimates, Inoue et al. compared the same set of CREs in the same cell line using either integrating or integration deficient lentiviral particles. They found substantial differences between the two assays and showed that the results from the integrating assay are both more in line with ENCODE annotations and more reproducible between replicates (Inoue et al. 2017). In a follow up study the authors took a similar approach but increased the set of MPRA variants to be tested. In particular they tested what effect the positioning of the barcode and the CRE tile itself at either the 3' or the 5' end of the reporter construct has. In addition, they again tested integrating versus non integrating versions for each of the conditions as well as two STARR-seq versions either with or without promoter (Klein et al. 2020). This very comprehensive study found that the positioning of the CRE tile relative to the minimal promoter explained most of the variance between assays, followed by the distinction of plasmid-based and lentiviral assays. When comparing the dynamic range of the assays they found the biggest range for plasmid-based assays with the original MPRA plasmid (Melnikov et al. 2012) showing the greatest dynamic range and separation between positive and negative control sequences. Notably, they trained a model to predict enhancer activity from the sequence, based on biochemical, evolutionary,

and sequence-derived features that achieved reasonable predictive power. Using a model based on the same features, they aimed to pinpoint the differences between the different assays and found RNA binding proteins and splicing factors to be more predictive for assays that integrate the enhancer at the 3' end of the transcript. On the other hand, they found promoter binding proteins to be more predictive of activity in 5' CRE assays. In summary, they concluded that assays with 5' CRE are biased for enhancers that show promoter-like activity and 3' CRE assays like STARR-seq are biased by mRNA stability and splicing factors.

### **Are those really weaknesses?**

Even though often discussed as a drawback of MPRAs, some of these aspects can and should be viewed as advantages as well. The synthetic nature of the sequences enables design of *denovo* sequences to derive rules of gene regulation (King et al. 2020) or targeted perturbations of binding sites (Noack et al. 2022; Kreimer et al. 2022). Similarly, the artificial nature of episomal reporter assays can be viewed as a feature as it only measures the intrinsic activity of a sequence to act as enhancer in a given *trans*-environment without confounding factors.

## **2.2.2 New tools that can help derive rules of gene regulation by enhancers**

In recent years, a number of tools have been developed that are able to link enhancers to the genes they regulate. Either by linking genetic variants to expression differences via expression Quantitative Trait Loci (eQTL) or using perturbation screens.

### **eQTLs provide functional links but no mechanistic explanation**

Large consortium efforts like the Genotype Tissue Expression (GTEx) project (GTEx Consortium 2015) have enabled the identification of genetic variants that alter expression in diverse tissues (GTEx Consortium et al. 2017). This approach is powerful as it uses existing



genetic variance and is able to link CREs to the genes they regulate. A mechanistic link is not possible with this approach and it is limited to existing variation that is present at a reasonable frequency in the population. However, with the amount of sequencing that is being performed and its likely increase in the near future, more data will be available for this type of inference.

### **CRISPR screens can associate enhancer activity with transcriptional changes**

A promising approach that has been proposed are CRISPR perturbation screens (Klein et al. 2018a). CRISPR screens coupled with single cell RNA-seq to knock down genes have been used widely (Dixit et al. 2016; Datlinger et al. 2017), recently on a genome wide scale (Replogle et al. 2022). These types of screens measure the transcriptome of single cells and capture the gRNA that it has been perturbed with. Applying the same strategy to regulatory regions by directly targeting them and then measuring the impact of this perturbation on gene expression is a promising strategy, particularly due to the high amount of multiplexing that is possible with single cell technologies and pooled screens. Gasperini and colleagues showed the power of this approach by testing the impact of perturbation of nearly 6,000 candidate enhancers on gene expression. Usually perturbation screens aim to infect cells with on average one gRNA, limiting the number of perturbations that can be done in one experiment. Their strategy involved infecting cells with the gRNA library at a high multiplicity of infection (MOI) which led to the detection of on average 28 gRNAs per cell. While this might lead to interaction effects, they showed that this approach is very useful as they needed much fewer cells to reach the same power (Gasperini et al. 2019). Recently MPRA have been used in a similar manner with single cell readout and perturbation of binding sites in the tested CRE tiles along a differentiation trajectory (Kreimer et al. 2022). While such assays do not perturb the binding of transcription factors in the genomic sequence, they can measure the effect of such perturbation in different cell types i.e. *trans*-environments. The single cell transcriptome is just a means to classify cell types i.e. different *trans*-environments here rather than the actual readout. This can be a valuable strategy in testing enhancer

activity in diverse backgrounds. A more direct way of associating TFs with the CREs they act on was recently developed and termed transMPRA. Here gRNAs targeting specific TFs are coupled to putative enhancer sequences in one construct. This enables knockdown of a transcription factor and direct measurement of the impact of this knockdown on activity as measured by MPRA (Calderon et al. 2020). However the drawback of this approach is that the combinatorics lead to large numbers of TF/CRE pairs to be tested.

### **New approaches can be used to derive general rules**

The ultimate goal of all these studies is to derive generalizable rules that make it possible to predict gene expression from enhancer sequences, chromatin state (both accessibility and domain structure) and biochemical marks with high confidence. A number of recent studies have used smart experimental setups to derive some of those general rules and test some of the core assumptions related to transcriptional activation by enhancers. For example, Zuin and colleagues tested in how far distance between enhancer and promoter altered expression of their target. To this end, they integrated a reporter construct containing a well established enhancer-promoter pair into a so called "gene desert", a chromosomal region that is devoid of genes. Next they mobilized the enhancer by a piggyBac transposase mediated system causing it to randomly integrate in neighboring regions at different distances upstream and downstream of the promoter. Using this system they found a non-linear relationship between transcription activation and distance. As a function of distance the enhancer-promoter contact probability declines and with it burst frequency. Depending on the strength of the enhancer, they found complete or partial insulation by chromatin domain boundaries marked by CCCTC-binding factor (CTCF) (Zuin et al. 2022). A complementary approach used CRISPR-based recruitment of individual transcription factors to a promoter and measured how different TFs altered burst probability and size. They found three classes of TFs that altered bursting in different ways, by altering either burst probability, burst size or both. Notably, these groups of TFs were not classifiable by their DNA binding domain but rather the cofactors they are known to recruit (Mamrak et al. 2022). A third study used a smart combination of molecular tools, namely the auxin-inducible degron (AID) system to deplete

specific core cofactors and measure the effect of this perturbation on particular enhancers using a genome wide STARR-seq MPRA. By integrating an AID-tag in the genomic sequence of eight cofactors they were able to deplete the tagged protein upon auxin induction. With this strategy they could identify different classes of enhancer sequences based on their dependency on different cofactors (Neumayr et al. 2022).

### 2.2.3 If we can't crack the regulatory code, we have to learn it from the data

These new approaches have already enabled solving some long standing questions in the field of gene regulation. However, their generalizability remains limited as shown in the CRISPR perturbation screen performed by Gasperini et al. where they found many exceptions to the current rules of gene regulation. Many enhancers did not fall within the same TAD as the gene they regulate. Others were not identified as contacts in HiC data sets, however, at least enriched for proximity in 3D space. And one third of enhancers did not regulate the most proximal gene but rather skipped one TSS and regulated the next gene (Gasperini et al. 2019). This clearly shows that these presumptions are not set in stone but are context dependent. As it is impossible to measure all different modalities that contribute to gene expression in every possible cell type and state, we have to come up with a set of rules that is able to predict gene expression with high accuracy based on limited data, ideally only DNA sequence. One potential way out is to use machine learning for this task. While mathematics was once referred to as "unreasonably effective" in describing the natural sciences (Wigner 1960), Google scientists have proclaimed the "unreasonable effectiveness of data" to solve complex problems such as natural language understanding. Their claim is that, while we can not solve such problems with a simple formula or rule, we can use machine learning (Halevy et al. 2009). That this is true for long standing biological problems has been shown by the superior performance of Google's Neural Network-based protein folding prediction tool AlphaFold (Jumper et al. 2021). Already many groups have applied deep learning, particularly convolutional neural networks, to derive the rules of gene regulation.

For example by using what they call a Gigantic Parallel Reporter Assay in yeast (Boer et al. 2020; Vaishnav et al. 2022) and learning *cis*-regulatory logic from tens of millions of random DNA sequences measured in a single experiment. They first used a biologically interpretable model to predict how TF binding to promoters alters expression and found that weak regulatory interactions explain most of the expression (Boer et al. 2020). Next they used deep neural networks to predict expression and fitness landscapes and test how evolutionary pressures act on regulatory sequence evolution (Vaishnav et al. 2022). Another study performed deep single cell RNA and ATAC sequencing of the *Drosophila* brain across 9 developmental time points. By identifying chromatin accessibility in diverse cell types along the developmental trajectory and associating these with transcription factors and their downstream targets they were able to construct enhancer gene regulatory networks using deep learning (Janssens et al. 2022).

This shows how modern molecular methods combined with massively parallel sequencing and machine learning can help to elucidate principles of gene regulation. However, it has to be kept in mind that also machine learning approaches as employed above need to be interpretable in the way they generate their predictions if we truly want to understand gene regulation. In addition, it has been argued that already the earliest such methods have benefited a lot from the biological knowledge of their developers and this is likely still true today (Bromberg 2022).

## 2.3 Comparing gene expression between species relies on gene orthology

Comparative genomics and particularly comparative transcriptomics and regulation is a powerful approach and has led to great insight. However, while we are able to measure gene expression transcriptome wide, comparative approaches limit themselves to the comparison of 1-to-1 orthologous genes potentially at the cost of power and resolution. For well-annotated genomes like human and mouse this still leaves a large fraction of genes (15,736) for analysis (Yue et al. 2014). As soon as more species are compared or species with less well annotated genomes, this number drops drastically. For example a recent paper investigating isoform diversity in five primate species restricted their analysis to 7,858 1-to-1 orthologs (Ferrández-Peral et al. 2021).

### 2.3.1 Orthologs, Paralogs and functional conservation

Generally, two types of homologous genes are distinguished, orthologs and paralogs. While orthologs are generated by a speciation event, paralogs are descendants of the same gene and are generated by a duplication event (Koonin 2005). To search for these types of relationships, many computational tools have been developed that use sequence similarity and build a hierarchical gene tree using all homologs between the species. This leads to the identification of 1-to-1 orthologs as well as 1-to-many orthologs and paralogs (Trachana et al. 2011). As most methods compare transcriptomes on a gene level, all information for genes that can not be matched unambiguously in a 1-to-1 manner is usually discarded. This biases the analysis towards the most conserved genes and in addition assumes that sequence conservation and homology relationships are a predictor of functional conservation. This notion has been termed the ortholog conjecture and has been discussed extensively (Gabaldón and Koonin 2013; Stambouliau et al. 2020). Particularly, the assumption that orthologs are functionally more similar than paralogs has been challenged by some (Stambouliau et al. 2020).

### 2.3.2 Potential ways out, beyond single gene comparisons

A solution would ideally not even have to rely on the inferred 1-to-1 orthology relationship but would associate homologs based on their function. Some groups have tried to solve this problem by summarizing the expression by ortholog-groups (Hu et al. 2017; Levin et al. 2016; Leong et al. 2021), however this masks potentially important changes. A different approach could be associating genes within their ortholog-group based on their expression or their expression change upon a given stimulus or perturbation. To this end, it might be helpful to first align homologous cell types as the higher order evolutionary unit. Cell types can be defined based on their gene regulatory network (Arendt 2008; Arendt et al. 2019) and comparing the topology of these networks between species beyond 1-to-1 orthologs can shed light on which homologous genes are functionally the most similar. One study using whole organism atlases of several early branching metazoa, aligned cell types using the whole homology tree in a recursive manner. When comparing gene expression patterns within the ortholog-groups they found several examples of paralogs that showed more similar expression patterns than the corresponding orthologs (Tarashansky et al. 2021). Perturbation of cell type programs might help establish functional links between genes of different species. Using CRISPRi perturbation screens coupled with single cell RNA-seq read out can be used to perturb gene regulatory networks across species and obtain functionally homologous genes or groups of genes.

## 3 | Conclusion and Outlook

In this work I have presented the state of the art on RNA-sequencing, both for bulk and single cell approaches. Since its invention over ten years ago, RNA-seq has developed into a key tool for molecular biology. Single cell RNA-seq has led to huge improvements in RNA-seq methodology that decreased costs and increased sensitivity, accuracy and library complexity at the same time. I contributed to these improvements by developing a sensitive single cell RNA-seq method, benchmarking it against other methods and finally transferring these improvements into bulk RNA-seq. The Human Cell Atlas (HCA) has been established with the goal to build a reference of all human cell types using mainly scRNA-seq (Regev et al. 2017). The first results of this international effort are being published (Domínguez Conde et al. 2022; Sikkema et al. 2022; Tabula Sapiens Consortium\* et al. 2022) and soon a comprehensive set of many tissues will be available. This will be particularly useful for bulk RNA-seq. Using the information from detailed single cell atlases to deconvolute the cell type composition of bulk samples can help to use cell atlases efficiently for biomedical purposes (Chu et al. 2022). The advantage of low cost bulk methods like prime-seq over single cell RNA-seq is that they enable processing thousands of biological replicates in a single experiment.

One of the biggest challenges in the field of genomics today is understanding the code that regulates gene expression and leads to different cell types. The main differences between the genetic code and the regulatory code lie in how they affect the cellular phenotype. A fixed change in the coding sequence leads to an organism-wide change in protein sequence and thus protein function. In contrast, a change in the regulatory sequence of a gene leads to quantitative spatio-temporal changes, leaving the protein function intact and

only affecting specific cell types or developmental times. While both types of sequences are under evolutionary constraint, the fundamental difference in the type of code is that regulatory evolution does not directly act on the exact sequence of bases but rather on a CREs quantitative effect on gene expression. Regulatory evolution thus leads to functional conservation but not necessarily sequence conservation. Genes that show little expression variation between species must have conserved regulatory architectures. Thus, looking at the regulation of genes with conserved expression patterns and identifying the commonalities of all active/open regulatory regions per gene is a promising approach to understand the basis of the regulatory code. Practically, this can be achieved by comparing gene expression, chromatin state and enhancer activity in parallel in a dynamic system. While of course challenging, this is technically possible by combining MPRA with single cell RNA-seq and potentially even single cell ATAC-seq of the same cell. Measuring these modalities in a differentiation experiment involving different species can be used to define conserved expression patterns and associate them with enhancer activity and regulatory architecture. In this context, STARR-seq combined with ATAC-seq enrichment would provide a feasible strategy for testing many species-specific enhancer sequences at once without the need for *a priori* sequence selection.

In conclusion, the combination of new high throughput single cell technologies in a single assay in a comparative evolutionary framework enables to measure gene regulation at unprecedented scale and will lead to a better understanding of the evolution of gene expression and the regulatory code in the coming years.



## 4 | References



# Bibliography

- Abdelaal, Tamim, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J T Reinders, and Ahmed Mahfouz (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* 20.1, 194.
- Agarwal, Saurabh, Todd S Macfarlan, Maureen A Sartor, and Shigeki Iwase (2015). Sequencing of first-strand cDNA library reveals full-length transcriptomes. *Nat. Commun.* 6, 6002.
- Alexa, A and J Rahnenfuhrer (2022). *topGO: Enrichment Analysis for Gene Ontology*.
- Alfaro, Javier Antonio, Peggy Bohländer, Mingjie Dai, Mike Filius, Cecil J Howard, Xander F van Kooten, Shilo Ohayon, Adam Pomorski, Sonja Schmid, Aleksei Aksimentiev, Eric V Anslyn, Georges Bedran, Chan Cao, Mauro Chinappi, Etienne Coyaud, Cees Dekker, Gunnar Dittmar, Nicholas Drachman, Rienk Eelkema, David Goodlett, et al. (2021). The emerging landscape of single-molecule protein sequencing technologies. *Nat. Methods* 18.6, 604–617.
- Almog, Gilad, Mark Pratt, Florian Oberstrass, Linda Lee, Dan Mazur, Nate Beckett, Omer Barad, Ilya Soifer, Eddie Perelman, Yoav Etzioni, Martin Sosa, April Jung, Tyson Clark, Eliane Trepagnier, Gila Lithwick-Yanai, Sarah Pollock, Gil Hornung, Maya Levy, Matthew Coole, Tom Howd, et al. (2022). Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform.
- Alpern, Daniel, Vincent Gardeux, Julie Russeil, Bastien Mangeat, Antonio C A Meireles-Filho, Romane Breyse, David Hacker, and Bart Deplancke (2019). BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol.* 20.1, 71.

- 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.
- Anders, Simon and Wolfgang Huber (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11.10, R106.
- Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31.2, 166–169.
- Anderson, S (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* 9.13, 3015–3027.
- Arendt, Detlev (2008). The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* 9.11, 868–882.
- Arendt, Detlev, Paola Yanina Bertucci, Kaia Achim, and Jacob M Musser (2019). Evolution of neuronal types and families. *Curr. Opin. Neurobiol.* 56, 144–152.
- Arnold, Cosmas D, Daniel Gerlach, Christoph Stelzer, Łukasz M Boryń, Martina Rath, and Alexander Stark (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339.6123, 1074–1077.
- Ashuach, Tal, David S Fischer, Anat Kreimer, Nadav Ahituv, Fabian J Theis, and Nir Yosef (2019). MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* 20.1, 183.
- Avery, O T, C M Macleod, and M McCarty (1944). Studies On The Chemical Nature Of The Substance Inducing Transformation Of Pneumococcal Types Induction Of Transformation By A Desoxyribonucleic Acid Fraction Isolated From Pneumococcus Type III. *J. Exp. Med.* 79.2, 137–158.
- 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.
- Baltimore, D (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* 226.5252, 1209–1211.
- Banerji, J, S Rusconi, and W Schaffner (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27.2 Pt 1, 299–308.

- Bentley, David R, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, Jonathan M Boutell, Jason Bryant, Richard J Carter, R Keira Cheetham, Anthony J Cox, Darren J Ellis, Michael R Flatbush, Niall A Gormley, Sean J Humphray, Leslie J Irving, et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456.7218, 53–59.
- Boer, Carl G de, Eeshit Dhaval Vaishnav, Ronen Sadeh, Esteban Luis Abeyta, Nir Friedman, and Aviv Regev (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* 38.1, 56–65.
- Booeshaghi, A Sina, Ingileif B Hallgrímsdóttir, Ángel Gálvez-Merchán, and Lior Pachter (2022). Depth normalization for single-cell genomics count data.
- Branton, Daniel, David W Deamer, Andre Marziali, Hagan Bayley, Steven A Benner, Thomas Butler, Massimiliano Di Ventra, Slaven Garaj, Andrew Hibbs, Xiaohua Huang, Stevan B Jovanovich, Predrag S Krstic, Stuart Lindsay, Xinsheng Sean Ling, Carlos H Mastrangelo, Amit Meller, John S Oliver, Yuriy V Pershin, J Michael Ramsey, Robert Riehn, et al. (2008). The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* 26.10, 1146–1153.
- Brawand, David, Magali Soumillon, Anamaria Necșulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann (2011). The evolution of gene expression levels in mammalian organs. *Nature* 478.7369, 343–348.
- Bray, Nicolas L, Harold Pimentel, Páll Melsted, and Lior Pachter (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34.5, 525–527.
- Breschi, Alessandra, Thomas R Gingeras, and Roderic Guigó (2017). Comparative transcriptomics in human and mouse. *Nat. Rev. Genet.* 18.7, 425–440.
- Britten, R J and E H Davidson (1969). Gene regulation for higher cells: a theory. *Science* 165.3891, 349–357.
- (1971). Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* 46.2, 111–138.

- Bromberg, Yana (2022). Tightening the (neural) net for protein structure prediction. *Nat. Rev. Genet.* 23.6, 322–323.
- Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10.12, 1213–1218.
- Bues, Johannes, Marjan Biočanin, Joern Pezoldt, Riccardo Dainese, Antonius Chrisnandy, Saba Rezakhani, Wouter Saelens, Vincent Gardeux, Revant Gupta, Rita Sarkis, Julie Russeil, Yvan Saeys, Esther Amstad, Manfred Claassen, Matthias P Lutolf, and Bart Deplancke (2022). Deterministic scRNA-seq captures variation in intestinal crypt and organoid composition. *Nat. Methods* 19.3, 323–330.
- Calderon, Diego, Andria Ellis, Riza M Daza, Beth Martin, Jacob M Tome, Wei Chen, Florence M Chardon, Anh Leith, Choli Lee, Cole Trapnell, and Jay Shendure (2020). TransMPRA: A framework for assaying the role of many trans-acting factors at many enhancers.
- Cao, Junyue, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, Andrew Adey, Robert H Waterston, Cole Trapnell, and Jay Shendure (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357.6352, 661–667.
- Chen, Yunshun, Aaron T L Lun, and Gordon K Smyth (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res.* 5, 1438.
- Chidgeavadze, Z G, R S Beabealashvili, A M Atrazhev, M K Kukhanova, A V Azhayev, and A A Krayevsky (1984). 2',3'-Dideoxy-3' aminonucleoside 5'-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. *Nucleic Acids Res.* 12.3, 1671–1686.
- Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437.7055, 69–87.

- Chu, Tinyi, Zhong Wang, Dana Pe'er, and Charles G Danko (2022). Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer* 3.4, 505–517.
- Clark, Iain C, Kristina M Fontanez, Robert H Meltzer, Yi Xue, Corey Hayford, Aaron May-Zhang, Chris D'Amato, Ahmad Osman, Jesse Q Zhang, Pabodha Hettige, Cyrille L Delley, Daniel W Weisgerber, Joseph M Replogle, Marco Jost, Kiet T Phong, Vanessa E Kennedy, Cheryl A C Peretz, Esther A Kim, Siyou Song, William Karlon, et al. (2022). Microfluidics-free single-cell genomics with templated emulsification.
- Clark, Stephen J, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, Oliver Stegle, and Wolf Reik (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* 9.1, 781.
- Coffin, John M (2021). 50th anniversary of the discovery of reverse transcriptase. *Mol. Biol. Cell* 32.2, 91–97.
- Coker, Jeffrey Scott and Eric Davies (2004). Identifying adaptor contamination when mining DNA sequence data. *Biotechniques* 37.2, 194, 196, 198.
- Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szczęśniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, and Ali Mortazavi (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17.1, 13.
- Crawford, Gregory E, Ingeborg E Holt, James Whittle, Bryn D Webb, Denise Tai, Sean Davis, Elliott H Margulies, Yidong Chen, John A Bernat, David Ginsburg, Daixing Zhou, Shujun Luo, Thomas J Vasicek, Mark J Daly, Tyra G Wolfsberg, and Francis S Collins (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* 16.1, 123–131.
- Crick, F (1970). Central dogma of molecular biology. *Nature* 227.5258, 561–563.
- Crick, F H, L Barnett, S Brenner, and R J Watts-Tobin (1961). General nature of the genetic code for proteins. *Nature* 192, 1227–1232.

- Cui, Jiawen, Nan Shen, Zhaogeng Lu, Guolu Xu, Yuyao Wang, and Biao Jin (2020). Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome. *Plant Methods* 16, 85.
- Dard-Dascot, Cloelia, Delphine Naquin, Yves d'Aubenton-Carafa, Karine Alix, Claude Thermes, and Erwin van Dijk (2018). Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics* 19.1, 118.
- Datlinger, Paul, André F Rendeiro, Thorina Boenke, Martin Senekowitsch, Thomas Krausgruber, Daniele Barreca, and Christoph Bock (2021). Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* 18.6, 635–642.
- 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.
- De Rop, Florian V, Joy N Ismail, Carmen Bravo González-Blas, Gert J Hulselmans, Christopher Campbell Flerin, Jasper Janssens, Koen Theunis, Valerie M Christiaens, Jasper Wouters, Gabriele Marcassa, Joris de Wit, Suresh Poovathingal, and Stein Aerts (2022). Hydrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads. *Elife* 11.
- DeAngelis, M M, D G Wang, and T L Hawkins (1995). Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* 23.22, 4742–4743.
- 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.
- Dijk, Erwin L van, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30.9, 418–426.
- Ding, Jiarui, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T



- Nguyen, John Y H Kwon, Boaz Barak, William Ge, Amanda J Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K Shalek, Alexandra-Chloé Villani, et al. (2020). Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* 38.6, 737–746.
- Dixit, Atray (2021). Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments.
- 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.
- Djebali, Sarah, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K Marinov, Jainab Khatun, Brian A Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, et al. (2012). Landscape of transcription in human cells. *Nature* 489.7414, 101–108.
- Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29.1, 15–21.
- Domínguez Conde, C, C Xu, L B Jarvis, D B Rainbow, S B Wells, T Gomes, S K Howlett, O Suchanek, K Polanski, H W King, L Mamanova, N Huang, P A Szabo, L Richardson, L Bolt, E S Fasouli, K T Mahbubani, M Prete, L Tuck, N Richoz, et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 376.6594, eabl5197.
- Eberwine, J, H Yeh, K Miyashiro, Y Cao, S Nair, R Finnell, M Zettel, and P Coleman (1992). Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. U. S. A.* 89.7, 3010–3014.
- Eisenstein, Michael (2020). Enzymatic DNA synthesis enters new phase. *Nat. Biotechnol.* 38.10, 1113–1115.
- Enard, Wolfgang (2012). Functional primate genomics—leveraging the medical potential. *J. Mol. Med.* 90.5, 471–480.

- Enard, Wolfgang (2014). Comparative genomics of brain size evolution. *Front. Hum. Neurosci.* 8, 345.
- (2015). Human evolution: enhancing the brain. *Curr. Biol.* 25.10, R421–3.
- (2016). The Molecular Basis of Human Brain Evolution. *Curr. Biol.* 26.20, R1109–R1117.
- Enard, Wolfgang, Philipp Khaitovich, Joachim Klose, Sebastian Zöllner, Florian Heissig, Patrick Giavalisco, Kay Nieselt-Struwe, Elaine Muchmore, Ajit Varki, Rivka Ravid, Gaby M Doxiadis, Ronald E Bontrop, and Svante Pääbo (2002a). Intra- and interspecific variation in primate gene expression patterns. *Science* 296.5566, 340–343.
- Enard, Wolfgang, Molly Przeworski, Simon E Fisher, Cecilia S L Lai, Victor Wiebe, Takashi Kitano, Anthony P Monaco, and Svante Pääbo (2002b). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 418.6900, 869–872.
- Esgleas, Miriam, Sven Falk, Ignasi Forné, Marc Thiry, Sonia Najas, Sirui Zhang, Aina Mas-Sanchez, Arie Geerlof, Dierk Niessing, Zefeng Wang, Axel Imhof, and Magdalena Götz (2020). Trnp1 organizes diverse nuclear membrane-less compartments in neural stem cells. *EMBO J.* 39.16, e103373.
- Espinós, Alexandre, Eduardo Fernández-Ortuño, Enrico Negri, and Víctor Borrell (2022). Evolution of genetic mechanisms regulating cortical neurogenesis. *Dev. Neurobiol.*
- FANTOM Consortium and the RIKEN PMI and CLST (DGT), Alistair R R Forrest, Hideya Kawaji, Michael Rehli, J Kenneth Baillie, Michiel J L de Hoon, Vanja Haberle, Timo Lassmann, Ivan V Kulakovskiy, Marina Lizio, Masayoshi Itoh, Robin Andersson, Christopher J Mungall, Terrence F Meehan, Sebastian Schmeier, Nicolas Bertin, Mette Jørgensen, Emmanuel Dimont, Erik Arner, Christian Schmidl, et al. (2014). A promoter-level mammalian expression atlas. *Nature* 507.7493, 462–470.
- Felsenstein, Joseph (1985). Phylogenies and the Comparative Method. *Am. Nat.* 125.1, 1–15.
- Ferrández-Peral, Luis, Xiaoyu Zhan, Marina Álvarez-Estapé, Cristina Chiva, Paula Esteller-Cucala, Raquel García-Pérez, Eva Julià, Esther Lizano, Òscar Fornas, Eduard Sabidó, Qiye Li, Tomàs Marquès-Bonet, David Juan, and Guojie Zhang (2021). Transcriptome innovations in primates revealed by single-molecule long-read sequencing.
- Fiers, W, R Contreras, F Duerinck, G Haegeman, D Iserentant, J Merregaert, W Min Jou, F Molemans, A Raeymaekers, A Van den Berghe, G Volckaert, and M Ysebaert

- (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* 260.5551, 500–507.
- Fleming, Stephen J, John C Marioni, and Mehrtash Babadi (2019). CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets.
- Fontsere, Claudia, Martin Kuhlwilm, Carlos Morcillo-Suarez, Marina Alvarez-Estape, Jack D Lester, Paolo Gratton, Joshua M Schmidt, Paula Dieguez, Thierry Aebischer, Paula Álvarez-Varona, Anthony Agbor, Samuel Angedakin, Alfred K Assumang, Emmanuel A Ayimisin, Emma Bailey, Donatienne Barubiyo, Mattia Bessone, Andrea Carretero-Alonso, Rebecca Chancellor, Heather Cohen, et al. (2022). Population dynamics and genetic connectivity in recent chimpanzee history. *Cell Genom* 2.6, None.
- Franklin, R E and R G Gosling (1953). Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature* 172.4369, 156–157.
- Fuller, Carl W, Pius S Padayatti, Hadi Abderrahim, Lisa Adamiak, Nolan Alagar, Nagaraj Ananthapadmanabhan, Jihye Baek, Sarat Chinni, Chulmin Choi, Kevin J Delaney, Rich Dubielzig, Julie Frkanec, Chris Garcia, Calvin Gardner, Daniel Gebhardt, Tim Geiser, Zachariah Gutierrez, Drew A Hall, Andrew P Hodges, Guangyuan Hou, et al. (2022). Molecular electronics sensors on a scalable semiconductor chip: A platform for single-molecule measurement of binding kinetics and enzyme activity. *Proc. Natl. Acad. Sci. U. S. A.* 119.5.
- Gabaldón, Toni and Eugene V Koonin (2013). Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* 14.5, 360–366.
- Garalde, Daniel R, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, Michael Jordan, Jonah Ciccone, Sabrina Serra, Jemma Keenan, Samuel Martin, Luke McNeill, E Jayne Wallace, Lakmal Jayasinghe, Chris Wright, Javier Blasco, et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15.3, 201–206.
- García-Pérez, Raquel, Paula Esteller-Cucala, Glòria Mas, Irene Lobón, Valerio Di Carlo, Meritxell Riera, Martin Kuhlwilm, Arcadi Navarro, Antoine Blancher, Luciano Di Croce, José Luis Gómez-Skarmeta, David Juan, and Tomàs Marquès-Bonet (2021).

- Epigenomic profiling of primate lymphoblastoid cell lines reveals the evolutionary patterns of epigenetic activities in gene regulatory architectures. *Nat. Commun.* 12.1, 1–17.
- Gasparini, Molly, Andrew J Hill, José L McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D Zhang, Dana Jackson, Anh Leith, Jacob Schreiber, William S Noble, Cole Trapnell, Nadav Ahituv, and Jay Shendure (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* 176.1-2, 377–390.e19.
- Gasparini, Molly, Jacob M Tome, and Jay Shendure (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* 21.5, 292–310.
- Gene Ontology Consortium (2021). The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Res.* 49.D1, D325–D334.
- Genome 10K Community of Scientists (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.* 100.6, 659–674.
- 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.
- Gierahn, Todd M, Marc H Wadsworth 2nd, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love, and Alex K Shalek (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* 14.4, 395–398.
- Grau-Bové, Xavier, Cristina Navarrete, Cristina Chiva, Thomas Pribasniig, Meritxell Antó, Guifré Torruella, Luis Javier Galindo, Bernd Franz Lang, David Moreira, Purificación López-García, Iñaki Ruiz-Trillo, Christa Schleper, Eduard Sabidó, and Arnau Sebé-Pedrós (2022). A phylogenetic and proteomic reconstruction of eukaryotic chromatin evolution. *Nat Ecol Evol.*
- Graur, Dan, Yichen Zheng, Nicholas Price, Ricardo B R Azevedo, Rebecca A Zufall, and Eran Elhaik (2013). On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5.3, 578–590.

- Greenleaf, William J and Arend Sidow (2014). The future of sequencing: convergence of intelligent design and market Darwinism. *Genome Biol.* 15.3, 303.
- GTEX Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEX) pilot analysis: multitissue gene regulation in humans. *Science* 348.6235, 648–660.
- GTEX Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEX (eGTEX) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration & Visualization—EBI, Genome Browser Data Integration & Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts: Laboratory, Data Analysis & Coordinating Center (LDACC): NIH program management: et al. (2017). Genetic effects on gene expression across human tissues. *Nature* 550.7675, 204–213.
- Gubler, U and B J Hoffman (1983). A simple and very efficient method for generating cDNA libraries. *Gene* 25.2-3, 263–269.
- Hagemann-Jensen, Michael, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, Anton J M Larsson, Omid R Faridani, and Rickard Sandberg (2020). Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* 38.6, 708–714.
- Hagemann-Jensen, Michael, Christoph Ziegenhain, and Rickard Sandberg (2022). Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress. *Nat. Biotechnol.*
- Halevy, Alon, Peter Norvig, and Fernando Pereira (2009). The unreasonable effectiveness of data. *IEEE Intell. Syst.* 24.2, 8–12.
- Hammarsten, Olof (1895). Zur Kenntniss der Nucleoproteide. *Biol. Chem.* 19.1, 19–37.
- Han, Xiaoping, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, Daosheng Huang, Yang Xu, Wentao Huang, Mengmeng Jiang, Xinyi Jiang, Jie Mao, Yao Chen, Chenyu Lu, Jin Xie, Qun Fang, et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* 172.5, 1091–1107.e17.

- Hashimshony, Tamar, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, Yuval Dor, Aviv Regev, and Itai Yanai (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17, 77.
- Hashimshony, Tamar, Florian Wagner, Noa Sher, and Itai Yanai (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* 2.3, 666–673.
- Heather, James M and Benjamin Chain (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* 107.1, 1–8.
- Heid, C A, J Stevens, K J Livak, and P M Williams (1996). Real time quantitative PCR. *Genome Res.* 6.10, 986–994.
- Higuchi, R, C Fockler, G Dollinger, and R Watson (1993). Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biotechnology* 11.9, 1026–1030.
- Hölzer, Martin and Manja Marz (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *Gigascience* 8.5.
- 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.
- Huang, Yuanhua, Davis J McCarthy, and Oliver Stegle (2019). Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* 20.1, 273.
- Hughes, Travis K, Marc H Wadsworth 2nd, Todd M Gierahn, Tran Do, David Weiss, Priscila R Andrade, Feiyang Ma, Bruno J de Andrade Silva, Shuai Shao, Lam C Tsoi, Jose Ordovas-Montanes, Johann E Gudjonsson, Robert L Modlin, J Christopher Love, and Alex K Shalek (2020). Second-Strand Synthesis-Based Massively Parallel scRNA-Seq Reveals Cellular States and Molecular Features of Human Inflammatory Skin Pathologies. *Immunity* 53.4, 878–894.e7.
- Hunkapiller, T, R J Kaiser, B F Koop, and L Hood (1991). Large-scale and automated DNA sequence determination. *Science* 254.5028, 59–67.

- Hutter, Carolyn and Jean Claude Zenklusen (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell* 173.2, 283–285.
- Illumina Inc. (2017). *Sequencing Platforms*. <https://emea.illumina.com/systems/sequencing-platforms.html>. Accessed: 2022-6-27.
- (2022). *Illumina Announces Next Generation Products and Data at AGBT General Meeting to Advance Innovative Customer Solutions*. <https://investor.illumina.com/news/press-release-details/2022/Illumina-Announces-Next-Generation-Products-and-Data-at-AGBT-General-Meeting-to-Advance-Innovative-Customer-Solutions/default.aspx>. Accessed: 2022-6-27.
- Inoue, Fumitaka and Nadav Ahituv (2015). Decoding enhancers using massively parallel reporter assays. *Genomics* 106.3, 159–164.
- Inoue, Fumitaka, Martin Kircher, Beth Martin, Gregory M Cooper, Daniela M Witten, Michael T McManus, Nadav Ahituv, and Jay Shendure (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* 27.1, 38–52.
- International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431.7011, 931–945.
- Islam, Saiful, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21.7, 1160–1167.
- Islam, Saiful, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11.2, 163–166.
- Janjic, Aleksandar, Lucas E Wange, Johannes W Bagnoli, Johanna Geuder, Phong Nguyen, Daniel Richter, Beate Vieth, Binje Vick, Irmela Jeremias, Christoph Ziegenhain, Ines Hellmann, and Wolfgang Enard (2022). Prime-seq, efficient and powerful bulk RNA sequencing. *Genome Biol.* 23.1, 88.
- Janssens, Jasper, Sara Aibar, Ibrahim Ihsan Taskiran, Joy N Ismail, Alicia Estacio Gomez, Gabriel Aughey, Katina I Spanier, Florian V De Rop, Carmen Bravo González-Blas, Marc Dionne, Krista Grimes, Xiao Jiang Quan, Dafni Papisokrati, Gert Hulselmans,

- Samira Makhzami, Maxime De Waegeneer, Valerie Christiaens, Tony Southall, and Stein Aerts (2022). Decoding gene regulation in the fly brain. *Nature* 601.7894, 630–636.
- Jegga, Anil G, Shawn P Sherwood, James W Carman, Andrew T Pinski, Jerry L Phillips, John P Pestian, and Bruce J Aronow (2002). Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.* 12.9, 1408–1417.
- Jerison, H J (1961). Quantitative analysis of evolution of the brain in mammals. *Science* 133.3457, 1012–1014.
- Jiang, Lichun, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* 21.9, 1543–1551.
- Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596.7873, 583–589.
- Kelava, Iva, Eric Lewitus, and Wieland B Huttner (2013). The secondary loss of gyrencephaly as an example of evolutionary phenotypical reversal. *Front. Neuroanat.* 7, 16.
- Kerimoglu, Cemil, Linh Pham, Anton B Tonchev, M Sadman Sakib, Yuanbin Xie, Godwin Sokpor, Pauline Antonie Ulmke, Lalit Kaurani, Eman Abbas, Huong Nguyen, Joachim Rosenbusch, Alexandra Michurina, Vincenzo Capece, Meglena Angelova, Nenad Maricic, Beate Brand-Saberi, Miriam Esgleas, Mareike Albert, Radoslav Minkov, Emil Kovachev, et al. (2021). H3 acetylation selectively promotes basal progenitor proliferation and neocortex expansion. *Sci Adv* 7.38, eabc6792.
- Khaitovich, Philipp, Wolfgang Enard, Michael Lachmann, and Svante Pääbo (2006). Evolution of primate gene expression. *Nat. Rev. Genet.* 7.9, 693–702.
- Khorana, H G, H Büchi, H Ghosh, N Gupta, T M Jacob, H Kössel, R Morgan, S A Narang, E Ohtsuka, and R D Wells (1966). Polynucleotide synthesis and the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* 31, 39–49.



- King, Dana M, Clarice Kit Yee Hong, James L Shepherdson, David M Granas, Brett B Maricque, and Barak A Cohen (2020). Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. *Elife* 9.
- King, M C and A C Wilson (1975). Evolution at two levels in humans and chimpanzees. *Science* 188.4184, 107–116.
- Kivioja, Teemu, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9.1, 72–74.
- 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.
- Klein, Jason C, Vikram Agarwal, Fumitaka Inoue, Aidan Keith, Beth Martin, Martin Kircher, Nadav Ahituv, and Jay Shendure (2020). A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* 17.11, 1083–1091.
- Klein, Jason C, Wei Chen, Molly Gasperini, and Jay Shendure (2018a). Identifying Novel Enhancer Elements with CRISPR-Based Screens. *ACS Chem. Biol.* 13.2, 326–332.
- Klein, Jason C, Aidan Keith, Vikram Agarwal, Timothy Durham, and Jay Shendure (2018b). Functional characterization of enhancer evolution in the primate lineage. *Genome Biol.* 19.1, 99.
- Klein, Jason C, Marc J Lajoie, Jerrod J Schwartz, Eva-Maria Strauch, Jorgen Nelson, David Baker, and Jay Shendure (2016). Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.* 44.5, e43.
- Kliesmete, Zane, Lucas E Wange, Beate Vieth, Miriam Esgleas, Jessica Radmer, Matthias Hülsmann, Johanna Geuder, Daniel Richter, Mari Ohnuki, Magdalena Götz, Ines Hellmann, and Wolfgang Enard (2021). TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals.
- Klughammer, Johanna, Daria Romanovskaia, Amelie Nemc, Annika Posautz, Charlotte Seid, Linda C Schuster, Melissa C Keinath, Juan Sebastian Lugo Ramos, Lindsay Kosack, Annie Evankow, Dieter Prinz, Stefanie Kirchberger, Bekir Ergüner, Paul Datlinger, Nikolaus

- Fortelny, Christian Schmidl, Matthias Farlik, Kaja Skjærven, Andreas Bergthaler, Miriam Liedvogel, et al. (2022). Comparative analysis of genome-scale, base-resolution DNA methylation profiles across 580 animal species.
- Koboldt, Daniel C, Karyn Meltz Steinberg, David E Larson, Richard K Wilson, and Elaine R Mardis (2013). The next-generation sequencing revolution and its impact on genomics. *Cell* 155.1, 27–38.
- Koonin, Eugene V (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39, 309–338.
- Kosuri, Sriram and George M Church (2014). Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* 11.5, 499–507.
- Kreimer, Anat, Tal Ashuach, Fumitaka Inoue, Alex Khodaverdian, Chengyu Deng, Nir Yosef, and Nadav Ahituv (2022). Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. *Nat. Commun.* 13.1, 1504.
- Kuderna, Lukas Fk, Paula Esteller-Cucala, and Tomas Marques-Bonet (2020). Branching out: what omics can tell us about primate evolution. *Curr. Opin. Genet. Dev.* 62, 65–71.
- Lambert, Samuel A, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch (2018). The Human Transcription Factors. *Cell* 172.4, 650–665.
- Lander, E S, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409.6822, 860–921.
- Le, Anh Viet-Phuong, Dexing Huang, Tony Blick, Erik W Thompson, and Alexander Dobrovic (2015). An optimised direct lysis method for gene expression studies on low cell numbers. *Sci. Rep.* 5, 12859.
- Lee, Dongwon, Ashish Kapoor, Changhee Lee, Michael Mudgett, Michael A Beer, and Aravinda Chakravarti (2021). Sequence-based correction of barcode bias in massively parallel reporter assays. *Genome Res.* 31.9, 1638–1645.

- Lee, Hayan, James Gurtowski, Shinjae Yoo, Maria Nattestad, Shoshana Marcus, Sara Goodwin, W Richard McCombie, and Michael C Schatz (2016). Third-generation sequencing and the future of genomics.
- Lee, Jeongwoo, Do Young Hyeon, and Daehee Hwang (2020). Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.* 52.9, 1428–1442.
- Leek, Jeffrey T (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* 42.21.
- Leong, Jason Cheok Kuan, Yongxin Li, Masahiro Uesaka, Yui Uchida, Akihito Omori, Meng Hao, Wenting Wan, Yang Dong, Yandong Ren, Si Zhang, Tao Zeng, Fayou Wang, Luonan Chen, Gary Wessel, Brian T Livingston, Cynthia Bradham, Wen Wang, and Naoki Irie (2021). Derivedness Index for Estimating Degree of Phenotypic Evolution of Embryos: A Study of Comparative Transcriptomic Analyses of Chordates and Echinoderms. *Front Cell Dev Biol* 9, 749963.
- Levene, P A and T Mori (1929). Ribodesose and xyloidesose and their bearing on the structure of thymine. *J. Biol. Chem.* 83.3, 803–816.
- Levene, P A and R S Tipson (1935). The Ring Structure of Thymidine. *Science* 81.2091, 98.
- Levin, Michal, Leon Anavy, Alison G Cole, Eitan Winter, Natalia Mostov, Sally Khair, Naf-talie Senderovich, Ekaterina Kovalev, David H Silver, Martin Feder, Selene L Fernandez-Valverde, Nagayasu Nakanishi, David Simmons, Oleg Simakov, Tomas Larsson, Shang-Yun Liu, Ayelet Jerafi-Vider, Karina Yaniv, Joseph F Ryan, Mark Q Martindale, et al. (2016). The mid-developmental transition and the evolution of animal body plans. *Nature* 531.7596, 637–641.
- Lewitus, Eric, Iva Kelava, Alex T Kalinka, Pavel Tomancak, and Wieland B Huttner (2014). An adaptive threshold in mammalian neocortical evolution. *PLoS Biol.* 12.11, e1002000.
- Li, Lily and Zeba Wunderlich (2017). An Enhancer’s Length and Composition Are Shaped by Its Regulatory Task. *Front. Genet.* 8, 63.
- Li, Xinmin and Cun-Yu Wang (2021). From bulk, single-cell to spatial RNA sequencing. *Int. J. Oral Sci.* 13.1, 36.

- Li, Yingshu, Hang Yang, Hujun Zhang, Yongjie Liu, Hanqiao Shang, Herong Zhao, Ting Zhang, and Qiang Tu (2020). Decode-seq: a practical approach to improve differential gene expression analysis. *Genome Biol.* 21.1, 66.
- Liao, Yang, Gordon K Smyth, and Wei Shi (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30.7, 923–930.
- Loots, G G, R M Locksley, C M Blankespoor, Z E Wang, W Miller, E M Rubin, and K A Frazer (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288.5463, 136–140.
- 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.
- Luckey, J A, H Drossman, A J Kostichka, D A Mead, J D’Cunha, T B Norris, and L M Smith (1990). High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res.* 18.15, 4417–4421.
- Luecken, Malte D, M Büttner, K Chaichoompu, A Danese, M Interlandi, M F Mueller, D C Strobl, L Zappia, M Dugas, M Colomé-Tatché, and Fabian J Theis (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* 19.1, 41–50.
- Macosko, Evan Z, Anindita Basu, Rahul Satija, James Nemes, 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.
- Mamrak, Nicholas E, Nader Alerasool, Daniel Griffith, Alex S Holehouse, Mikko Taipale, and Timothee Lionnet (2022). The kinetic landscape of human transcription factors.
- Mardis, Elaine R (2006). Anticipating the 1,000 dollar genome. *Genome Biol.* 7.7, 112.
- (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24.3, 133–141.
- Margulies, Marcel, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen,

- Chun Heen Ho, Gerard P Irzyk, Szilveszter C Jando, et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437.7057, 376–380.
- 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.
- Marques-Bonet, Tomas, Oliver A Ryder, and Evan E Eichler (2009). Sequencing primate genomes: what have we learned? *Annu. Rev. Genomics Hum. Genet.* 10, 355–386.
- Martínez-Martínez, María Ángeles, Camino De Juan Romero, Virginia Fernández, Adrián Cárdenas, Magdalena Götz, and Víctor Borrell (2016). A restricted period for formation of outer subventricular zone defined by *Cdh1* and *Trnp1* levels. *Nat. Commun.* 7, 11812.
- Massoni-Badosa, Ramon, Giovanni Iacono, Catia Moutinho, Marta Kulis, Núria Palau, Domenica Marchese, Javier Rodríguez-Ubreva, Esteban Ballestar, Gustavo Rodriguez-Esteban, Sara Marsal, Marta Aymerich, Dolors Colomer, Elias Campo, Antonio Julià, José Ignacio Martín-Subero, and Holger Heyn (2020). Sampling time-dependent artifacts in single-cell genomics studies. *Genome Biol.* 21.1, 112.
- Maxam, A M and W Gilbert (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* 74.2, 560–564.
- McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Großberger (2018). UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* 3.29, 861.
- Melnikov, Alexandre, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G Callan Jr, Justin B Kinney, Manolis Kellis, Eric S Lander, and Tarjei S Mikkelsen (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30.3, 271–277.
- Mendel, Gregor Johann (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines zu Brünn 4.* Vol. 4, 3–47.
- Mereu, Elisabetta, Atefeh Lafzi, Catia Moutinho, Christoph Ziegenhain, Davis J McCarthy, Adrián Álvarez-Varela, Eduard Batlle, Sagar, Dominic Grün, Julia K Lau, Stéphane C Boutet, Chad Sanada, Aik Ooi, Robert C Jones, Kelly Kaihara, Chris Brampton, Yasha Talaga, Yohei Sasagawa, Kaori Tanaka, Tetsutaro Hayashi, et al. (2020). Benchmarking

- single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* 38.6, 747–755.
- Mills, Caitlin, Anushya Muruganujan, Dustin Ebert, Crystal N Marconett, Juan Pablo Lewinger, Paul D Thomas, and Huaiyu Mi (2020). PEREGRINE: A genome-wide prediction of enhancer to gene relationships supported by experimental evidence. *PLoS One* 15.12, e0243791.
- Misteli, Tom (2020). The Self-Organizing Genome: Principles of Genome Architecture and Function. *Cell* 183.1, 28–45.
- Mootha, Vamsi K, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, Todd R Golub, Pablo Tamayo, Bruce Spiegelman, Eric S Lander, Joel N Hirschhorn, David Altshuler, et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34.3, 267–273.
- Mouse Genome Sequencing Consortium, Robert H Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F Abril, Pankaj Agarwal, Richa Agarwala, Rachel Ainscough, Marina Alexandersson, Peter An, Stylianos E Antonarakis, John Attwood, Robert Baertsch, Jonathon Bailey, Karen Barlow, Stephan Beck, Eric Berry, Bruce Birren, Toby Bloom, et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420.6915, 520–562.
- Muerdter, Felix, Łukasz M Boryń, Ashley R Woodfin, Christoph Neumayr, Martina Rath, Muhammad A Zabidi, Michaela Pagani, Vanja Haberle, Tomáš Kazmar, Rui R Catarino, Katharina Schernhuber, Cosmas D Arnold, and Alexander Stark (2018). Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* 15.2, 141–149.
- Murphy, Alan E and Nathan G Skene (2022). A balanced measure shows superior performance of pseudobulk methods over mixed models and pseudoreplication approaches in single-cell RNA-sequencing analysis.
- National Center for Human Genome Research (U.S.), United States., Department of Energy., Office of Health and Environmental Research., and Human Genome Program. (1990).

- Understanding our genetic inheritance : the U.S. Human Genome Project : the first five years, FY 1991-1995.* [Bethesda, Md.]; [Germantown, Md.]; Springfield, Va.: U.S. Dept. of Health, Human Services, Public Health Service, National Institutes of Health, National Center for Human Genome Research ; U.S. Dept. of Energy, Office of Energy Research, Office of Health, and Environmental Research, Human Genome Program ; National Technical Information Service [distributor].
- National Research Council, Division on Earth and Life Studies, Commission on Life Sciences, and Committee on Mapping and Sequencing the Human Genome (1988). *Mapping and Sequencing the Human Genome.* National Academies Press.
- National Research Council (US) Chemical Sciences Roundtable (1999). *The Role of Computational Biology in the Genomics Revolution.* National Academies Press (US).
- Neumayr, Christoph, Vanja Haberle, Leonid Serebreni, Katharina Karner, Oliver Hendy, Ann Boija, Jonathan E Henninger, Charles H Li, Karel Stejskal, Gen Lin, Katharina Bergauer, Michaela Pagani, Martina Rath, Karl Mechtler, Cosmas D Arnold, and Alexander Stark (2022). Differential cofactor dependencies define distinct types of human enhancers. *Nature.*
- Nirenberg, M and P Leder (1964). RNA Codewords And Protein Synthesis. The Effect Of Trinucleotide Upon The Binding Of sRNA To Ribosomes. *Science* 145.3639, 1399–1407.
- Nirenberg, M W and J H Matthaei (1961). The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* 47, 1588–1602.
- Noack, Florian, Silvia Vangelisti, Gerald Raffl, Madalena Carido, Jeisimhan Diwakar, Faye Chong, and Boyan Bonev (2022). Multimodal profiling of the transcriptional regulatory landscape of the developing mouse cortex identifies *Neurog2* as a key epigenome remodeler. *Nat. Neurosci.* 25.2, 154–167.
- Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J Hoyt, Mark Diekhans, Glennis A Logsdon, Michael Alonge, Stylianos E Antonarakis, Matthew Borchers, Gerard G Bouffard, Shelise Y Brooks, Gina

- V Caldas, et al. (2022). The complete sequence of a human genome. *Science* 376.6588, 44–53.
- O’Leary, Maureen A, Jonathan I Bloch, John J Flynn, Timothy J Gaudin, Andres Giallombardo, Norberto P Giannini, Suzann L Goldberg, Brian P Kraatz, Zhe-Xi Luo, Jin Meng, Xijun Ni, Michael J Novacek, Fernando A Perini, Zachary S Randall, Guillermo W Rougier, Eric J Sargis, Mary T Silcox, Nancy B Simmons, Michelle Spaulding, Paúl M Velazco, et al. (2013). The Placental Mammal Ancestor and the Post-K-Pg Radiation of Placentals. *Science* 339.6120, 662–667.
- O’Neil, Dominic, Heike Glowatz, and Martin Schlumpberger (2013). Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Curr. Protoc. Mol. Biol.* Chapter 4, Unit 4.19.
- 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.
- Oude Munnink, Bas B, Nathalie Worp, David F Nieuwenhuijse, Reina S Sikkema, Bart Haagmans, Ron A M Fouchier, and Marion Koopmans (2021). The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nat. Med.* 27.9, 1518–1524.
- Parekh, Swati, Beate Vieth, Christoph Ziegenhain, Wolfgang Enard, and Ines Hellmann (2018). Strategies for quantitative RNA-seq analyses among closely related species.
- Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14.4, 417–419.
- Phipson, Belinda, Luke Zappia, and Alicia Oshlack (2017). Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res.* 6, 595.
- Picelli, Simone, Asa K Björklund, Björn Reinius, Sven Sagasser, Gösta Winberg, and Rickard Sandberg (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* 24.12, 2033–2040.



- 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.
- Pool, Allan-Hermann, Helen Poldsam, Sisi Chen, Matt Thomson, and Yuki Oka (2022). Enhanced recovery of single-cell RNA-sequencing reads for missing gene expression data.
- Prakadan, Sanjay M, Alex K Shalek, and David A Weitz (2017). Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.* 18.6, 345–361.
- Pratapa, Aditya, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and T M Murali (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17.2, 147–154.
- Ramsköld, Daniel, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30.8, 777–782.
- Regev, Aviv, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, et al. (2017). The Human Cell Atlas. *Elife* 6.
- Repogle, Joseph M, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L Bonnar, Marco Jost, Thomas M Norman, and Jonathan S Weissman (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 2021.12.16.473013.
- Rhie, Arang, Shane A McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, William Chow, Arkarachai Fungtammasan, Juwan Kim, Chul Lee, Byung June Ko, Mark Chaisson, Gregory L Gedman, Lindsey J Cantin, Francoise Thibaud-Nissen, Leanne Haggerty, Iliana Bista, Michelle Smith, Bettina Haase, et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592.7856, 737–746.

- Risso, Davide, John Ngai, Terence P Speed, and Sandrine Dudoit (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32.9, 896–902.
- 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.
- Roberts, Richard J, Mauricio O Carneiro, and Michael C Schatz (2013). The advantages of SMRT sequencing. *Genome Biol.* 14.7, 405.
- Robinson, Mark D and Alicia Oshlack (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11.3, R25.
- Rogers, Jeffrey and Richard A Gibbs (2014). Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat. Rev. Genet.* 15.5, 347–359.
- 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.
- Rosenberg, Alexander B, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, Suzie H Pun, Drew L Sellers, Bosiljka Tasic, and Georg Seelig (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360.6385, 176–182.
- Saelens, Wouter, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 37.5, 547–554.
- Saiki, R K, D H Gelfand, S Stoffel, S J Scharf, R Higuchi, G T Horn, K B Mullis, and H A Erlich (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239.4839, 487–491.
- Sandberg, Rickard (2014). Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* 11.1, 22–24.
- Sanger, F and A R Coulson (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94.3, 441–448.

- Sanger, F, J E Donelson, A R Coulson, H Kössel, and D Fischer (1973). Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage f1 DNA. *Proc. Natl. Acad. Sci. U. S. A.* 70.4, 1209–1213.
- Sanger, F, S Nicklen, and A R Coulson (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74.12, 5463–5467.
- Sasagawa, Yohei, Hiroki Danno, Hitomi Takada, Masashi Ebisawa, Kaori Tanaka, Tetsutaro Hayashi, Akira Kurisaki, and Itoshi Nikaido (2018). Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* 19.1, 29.
- Scalable Single Cell Sequencing* (2021). <https://www.parsebiosciences.com/>. Accessed: 2022-5-26.
- 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.
- Sharon, Eilon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* 30.6, 521–530.
- Shendure, Jay, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston (2017). DNA sequencing at 40: past, present and future. *Nature* 550.7676, 345–353.
- Shiraki, Toshiyuki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, Shiro Fukuda, Daisuke Sasaki, Anna Podhajska, Matthias Harbers, Jun Kawai, Piero Carninci, and Yoshihide Hayashizaki (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* 100.26, 15776–15781.
- Sikkema, L, D Strobl, L Zappia, E Madisson, N S Markov, L Zaragosi, M Ansari, M Arguel, L Apperloo, C Bécavin, M Berg, E Chichelnitskiy, M Chung, A Collin, A C A Gay, B Hooshiar Kashani, M Jain, T Kapellos, T M Kole, C Mayr, et al. (2022). An integrated cell atlas of the human lung in health and disease.

- Soneson, Charlotte and Mark D Robinson (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* 15.4, 255–261.
- Song, Li-Fu, Zheng-Hua Deng, Zi-Yi Gong, Lu-Lu Li, and Bing-Zhi Li (2021). Large-Scale de novo Oligonucleotide Synthesis for Whole-Genome Synthesis and Data Storage: Challenges and Opportunities. *Front Bioeng Biotechnol* 9, 689797.
- 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.
- Spinozzi, Giulio, Valentina Tini, Alessia Adorni, Brunangelo Falini, and Maria Paola Martelli (2020). ARPIR: automatic RNA-Seq pipelines with interactive report. *BMC Bioinformatics* 21.Suppl 19, 574.
- Staden, R (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* 6.7, 2601–2610.
- Stamboulian, Moses, Rafael F Guerrero, Matthew W Hahn, and Predrag Radivojac (2020). The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. *Bioinformatics* 36.Suppl\_1, i219–i226.
- Stark, Rory, Marta Grzelak, and James Hadfield (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20.11, 631–656.
- Stefani, Giovanni and Frank J Slack (2008). Small non-coding RNAs in animal development. *Nat. Rev. Mol. Cell Biol.* 9.3, 219–230.
- Stillman, Bruce (2018). Histone Modifications: Insights into Their Influence on Gene Expression. *Cell* 175.1, 6–9.
- Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102.43, 15545–15550.
- Sun, Qinyu, Qinyu Hao, and Kannanganattu V Prasanth (2018). Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. *Trends Genet.* 34.2, 142–157.

- Sun, Tao and Robert F Hevner (2014). Growth and folding of the mammalian cerebral cortex: from molecules to malformations. *Nat. Rev. Neurosci.* 15.4, 217–232.
- Svec, David, Daniel Andersson, Milos Pekny, Robert Sjöback, Mikael Kubista, and Anders Ståhlberg (2013). Direct cell lysis for single-cell gene expression profiling. *Front. Oncol.* 3, 274.
- Svensson, Valentine, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14.4, 381–387.
- 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.
- Swerdlow, H and R Gesteland (1990). Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.* 18.6, 1415–1419.
- Tabula Sapiens Consortium\*, Robert C Jones, Jim Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaupt, Phillip Brown, William Harper, Marisa Hemenez, Ravikumar Ponnusamy, Ahmad Salehi, Bhavani A Sanagavarapu, Eileen Spallino, Ksenia A Aaron, Waldo Concepcion, James M Gardner, Burnett Kelly, et al. (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* 376.6594, eabl4896.
- 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.
- Tarashansky, Alexander J, Jacob M Musser, Margarita Khariton, Pengyang Li, Detlev Arendt, Stephen R Quake, and Bo Wang (2021). Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* 10.
- Temin, H M and S Mizutani (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* 226.5252, 1211–1213.

- Tian, Luyi, Xueyi Dong, Saskia Freytag, Kim-Anh Lê Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S Weber, Azadeh Seidi, Jafar S Jabbari, Shalin H Naik, and Matthew E Ritchie (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* 16.6, 479–487.
- Trachana, Kalliopi, Tomas A Larsson, Sean Powell, Wei-Hua Chen, Tobias Doerks, Jean Muller, and Peer Bork (2011). Orthology prediction methods: a quality assessment using curated protein families. *Bioessays* 33.10, 769–780.
- Uebbing, Severin, Jake Gockley, Steven K Reilly, Acadia A Kocher, Evan Geller, Neeru Gandotra, Curt Scharfe, Justin Cotney, and James P Noonan (2021). Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc. Natl. Acad. Sci. U. S. A.* 118.2.
- Ulirsch, Jacob C, Satish K Nandakumar, Li Wang, Felix C Giani, Xiaolan Zhang, Peter Rogov, Alexandre Melnikov, Patrick McDonel, Ron Do, Tarjei S Mikkelsen, and Vijay G Sankaran (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* 165.6, 1530–1545.
- Vaishnav, Eeshit Dhaval, Carl G de Boer, Jennifer Molinet, Moran Yassour, Lin Fan, Xian Adiconis, Dawn A Thompson, Joshua Z Levin, Francisco A Cubillos, and Aviv Regev (2022). The evolution, evolvability and engineering of gene regulatory DNA. *Nature* 603.7901, 455–463.
- Van Maaten and Hinton (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.*
- Vanhille, Laurent, Aurélien Griffon, Muhammad Ahmad Maqbool, Joaquin Zacarias-Cabeza, Lan T M Dao, Nicolas Fernandez, Benoit Ballester, Jean Christophe Andrau, and Salvatore Spicuglia (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat. Commun.* 6, 6905.
- Velculescu, V E, L Zhang, B Vogelstein, and K W Kinzler (1995). Serial analysis of gene expression. *Science* 270.5235, 484–487.
- Venter, J C, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, et al. (2001). The sequence of the human genome. *Science* 291.5507, 1304–1351.

- 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.
- Vieth, Beate, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann (2017). powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* 33.21, 3486–3488.
- Vischer, E and E Chargaff (1948). The separation and quantitative estimation of purines and pyrimidines in minute amounts. *J. Biol. Chem.* 176.2, 703–714.
- Vockley, Christopher M, Anthony M D'Ippolito, Ian C McDowell, William H Majoros, Alexias Safi, Lingyun Song, Gregory E Crawford, and Timothy E Reddy (2016). Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome. *Cell* 166.5, 1269–1281.e19.
- Vockley, Christopher M, Cong Guo, William H Majoros, Michael Nodzenski, Denise M Scholtens, M Geoffrey Hayes, William L Lowe Jr, and Timothy E Reddy (2015). Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* 25.8, 1206–1214.
- Volpe, Marina, Sally Shpungin, Chany Barbi, Galya Abrham, Hanna Malovani, Ron Wides, and Uri Nir (2006). trnp: A conserved mammalian gene encoding a nuclear protein that accelerates cell-cycle progression. *DNA Cell Biol.* 25.6, 331–339.
- Wang, Xinchun, Liang He, Sarah M Goggin, Alham Saadat, Li Wang, Nasa Sinnott-Armstrong, Melina Claussnitzer, and Manolis Kellis (2018). High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* 9.1, 5380.
- Wang, Zhong-Yi, Evgeny Leushkin, Angélica Liechti, Svetlana Ovchinnikova, Katharina Mößinger, Thoomke Brüning, Coralie Rummel, Frank Grützner, Margarida Cardoso-Moreira, Peggy Janich, David Gatfield, Boubou Diagouraga, Bernard de Massy, Mark E Gill, Antoine H F M Peters, Simon Anders, and Henrik Kaessmann (2020). Transcriptome and translome co-evolution in mammals. *Nature* 588.7839, 642–647.

- Wangsanuwat, Chatarin, Kellie A Heom, Estella Liu, Michelle A O'Malley, and Siddharth S Dey (2020). Efficient and cost-effective bacterial mRNA sequencing from low input samples through ribosomal RNA depletion. *BMC Genomics* 21.1, 717.
- Wasserman, Wyeth W and Albin Sandelin (2004). Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 5.4, 276–287.
- Watson, J D and F H Crick (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171.4356, 737–738.
- Whalen, Sean, Fumitaka Inoue, Hane Ryu, Tyler Fair, Eirene Markenscoff-Papadimitriou, Kathleen Keough, Martin Kircher, Beth Martin, Beatriz Alvarado, Orry Elor, Dianne Laboy Cintron, Alex Williams, Md Abul Hassan Samee, Sean Thomas, Robert Krenzik, Erik M Ullian, Arnold Kriegstein, Jay Shendure, Alex A Pollen, Nadav Ahituv, et al. (2022). Machine-learning dissection of Human Accelerated Regions in primate neurodevelopment.
- Wigner, Eugene P (1960). The unreasonable effectiveness of mathematics in the natural sciences. Richard Courant lecture in mathematical sciences delivered at New York University, May 11, 1959. *Commun. Pure Appl. Math.* 13.1, 1–14.
- Wodicka, L, H Dong, M Mittmann, M H Ho, and D J Lockhart (1997). Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 15.13, 1359–1367.
- Wong, Marisa L and Juan F Medrano (2005). Real-time PCR for mRNA quantitation. *Biotechniques* 39.1, 75–85.
- Wulf, Madalee G, Sean Maguire, Paul Humbert, Nan Dai, Yanxia Bei, Nicole M Nichols, Ivan R Corrêa Jr, and Shengxi Guan (2019). Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. *J. Biol. Chem.* 294.48, 18220–18231.
- 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.
- Yu, Guangchuang and Qing-Yu He (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* 12.2, 477–479.



- Yu, Lijia, Yue Cao, Jean Y H Yang, and Pengyi Yang (2022). Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biol.* 23.1, 49.
- Yue, Feng, Yong Cheng, Alessandra Breschi, Jeff Vierstra, Weisheng Wu, Tyrone Ryba, Richard Sandstrom, Zhihai Ma, Carrie Davis, Benjamin D Pope, Yin Shen, Dmitri D Pervouchine, Sarah Djebali, Robert E Thurman, Rajinder Kaul, Eric Rynes, Anthony Kirilusha, Georgi K Marinov, Brian A Williams, Diane Trout, et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515.7527, 355–364.
- Zajac, Pawel, Saiful Islam, Hannah Hochgerner, Peter Lönnerberg, and Sten Linnarsson (2013). Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PLoS One* 8.12, e85270.
- Zappia, Luke, Belinda Phipson, and Alicia Oshlack (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* 14.6, e1006245.
- Zhang, Xiaokang and Inge Jonassen (2020). RASflow: an RNA-Seq analysis workflow with Snakemake. *BMC Bioinformatics* 21.1, 110.
- 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 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.

- Zhu, Y Y, E M Machleder, A Chenchik, R Li, and P D Siebert (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30.4, 892–897.
- Ziegenhain, Christoph, Gert-Jan Hendriks, Michael Hagemann-Jensen, and Rickard Sandberg (2022). Molecular spikes: a gold standard for single-cell RNA counting. *Nat. Methods* 19.5, 560–566.
- 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.
- Zilles, K, E Armstrong, K H Moser, A Schleicher, and H Stephan (1989). Gyrification in the cerebral cortex of primates. *Brain Behav. Evol.* 34.3, 143–150.
- Zimmerman, Kip D, Mark A Espeland, and Carl D Langefeld (2021). A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* 12.1, 738.
- Zimmerman, S B and B H Pfeiffer (1983). Macromolecular crowding allows blunt-end ligation by DNA ligases from rat liver or *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 80.19, 5852–5856.
- Zuin, Jessica, Gregory Roth, Yinxiu Zhan, Julie Cramard, Josef Redolfi, Ewa Piskadlo, Pia Mach, Mariya Kryzhanovska, Gergely Tihanyi, Hubertus Kohler, Mathias Eder, Christ Leemans, Bas van Steensel, Peter Meister, Sebastien Smallwood, and Luca Giorgetti (2022). Nonlinear control of transcription through enhancer-promoter interactions. *Nature* 604.7906, 571–577.

## 5 | Appendices

## 5.1 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).

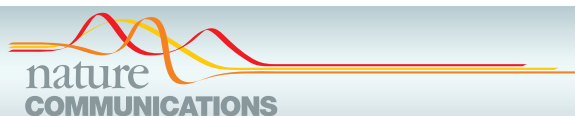
doi: <https://doi.org/10.1038/s41467-018-05347-6>

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.



ARTICLE

DOI: 10.1038/s41467-018-05347-6

OPEN

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

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

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.

<sup>1</sup>Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Großhaderner Straße 2, 82152 Martinsried, Germany.

<sup>2</sup>Present address: Department of Cell & Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden. <sup>3</sup>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](mailto:enard@bio.lmu.de))

## ARTICLE

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-05347-6

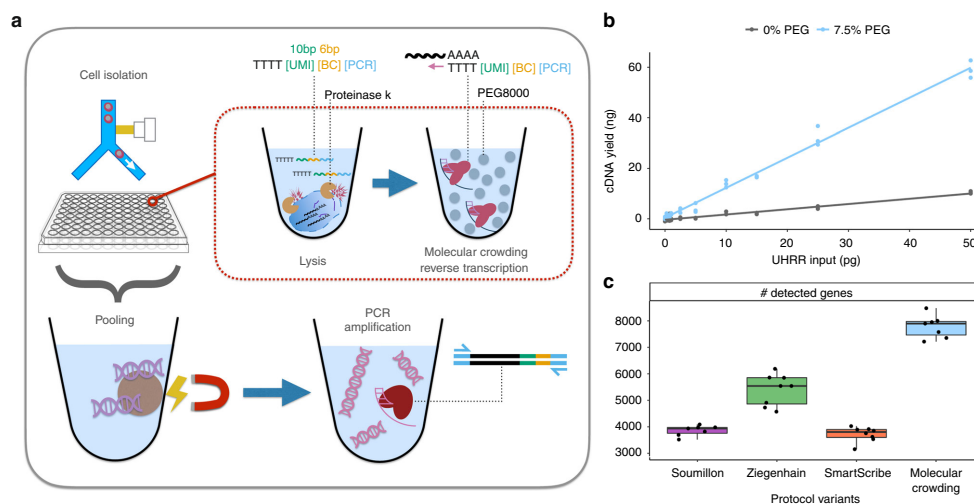
Whole transcriptome single-cell RNA sequencing (scRNA-seq) is a transformative tool with wide applicability to biological and biomedical questions<sup>1,2</sup>. Recently, many scRNA-seq protocols have been developed to overcome the challenge of isolating, reverse transcribing, and amplifying the small amounts of mRNA in single cells to generate high-throughput sequencing libraries<sup>3,4</sup>. However, as there is no optimal, one-size-fits all protocol, various inherent strengths and trade-offs exist<sup>5-7</sup>. Among flexible, plate-based methods, single-cell RNA barcoding and sequencing (SCRB-seq)<sup>8</sup> is one of the most powerful and cost-efficient<sup>6</sup>, 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)<sup>9</sup> 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<sub>2</sub>, betaine and trehalose, as these led to the increased sensitivity of the Smart-seq2 protocol<sup>10</sup>. 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<sup>11</sup> and is thought to increase enzymatic reaction rates by mimicking (macro)molecular crowding, i.e., by reducing the effective reaction volume<sup>12</sup>. As small reaction volumes can increase the sensitivity of scRNA-seq protocols<sup>5,13</sup>, 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 SCR-seq protocol variants (Supplementary Tables 1, 2): the original SCR-seq protocol<sup>8</sup> (“Soumillon”; with Maxima H- as RT and Advantage2 as PCR enzyme), the slightly adapted protocol benchmarked in Ziegenhain et al.<sup>6</sup> (“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<sup>14</sup> 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<sup>5,6</sup>. 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 ( $\geq 1$  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 SCR-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<sup>15,16</sup>. 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<sup>6,17</sup>. 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<sup>14</sup>, 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<sup>17</sup>. In addition to choosing Terra for cDNA amplification, we also reduced the number of cycles from 19 in the original SCR-seq protocol to 14, as fewer cycles are expected to decrease amplification bias further<sup>15</sup> and 14 cycles still generated sufficient amounts of cDNA (~1.6–2.4 ng/ $\mu$ l) 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.

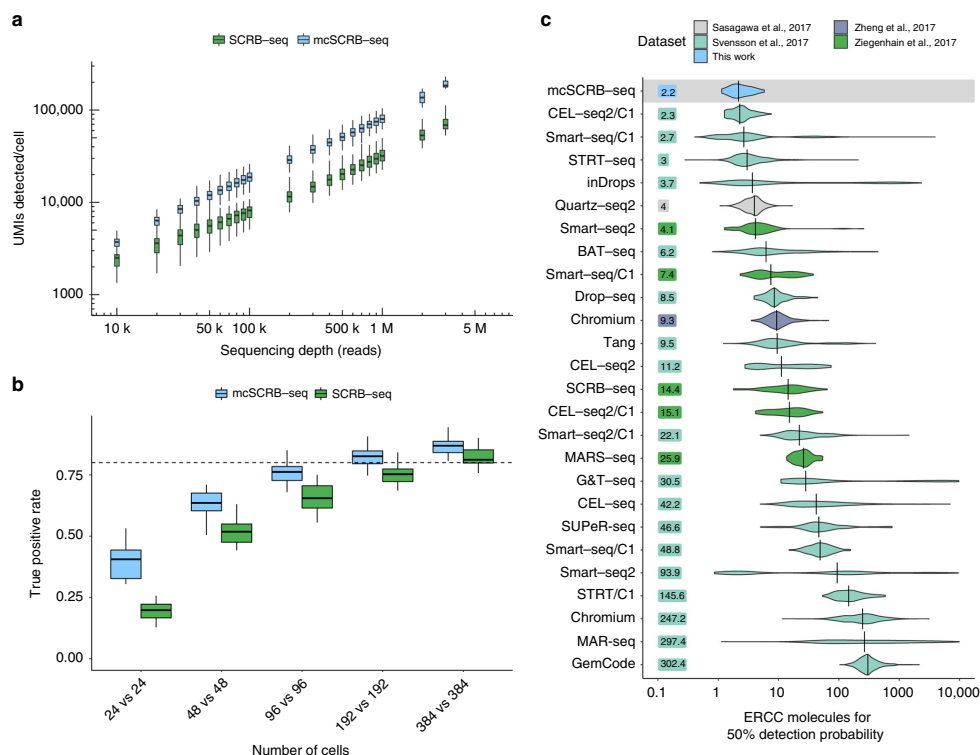
With the final improved version of the molecular crowding protocol (mcSCR-seq), we tested to what extent cross-contamination occurs. For example, chimeric PCR products may occur following the pooling of cDNA<sup>18</sup> and we assessed whether this might potentially be influenced by PEG that is present 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 mcSCR-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, mcSCR-seq, that has higher sensitivity, a less biased amplification and little crosstalk of reads across cells.

#### mcSCR-seq increases sensitivity 2.5-fold more than SCR-seq.

To directly compare the entire mcSCR-seq protocol to the previously benchmarked SCR-seq protocol used in Ziegenhain et al.<sup>6</sup> (Supplementary Table 2), we sorted for each method 48 and 96 single mESCs from one culture into plates, and added ERCC spike-ins<sup>19</sup>. 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, mcSCR-seq clearly outperforms SCR-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, mcSCR-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<sup>20</sup>. As expected, the higher number of UMIs in mcSCR-seq also results in a higher number of detected genes. For instance, at 500,000 reads, mcSCR-seq detected 50,969 UMIs that corresponded to

## ARTICLE

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-05347-6



**Fig. 2** Comparison of mcSCR-seq to SCR-seq and other protocols. **a** Number of UMIs detected in libraries generated from 249 single mESCs using SCR-seq or mcSCR-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 mcSCR-seq and SCR-seq estimated by power simulations using the powsimR package<sup>22</sup>. The empirical mean-variance distribution of the 10,904 genes that were detected in at least 10 cells in either mcSCR-seq or SCR-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 that 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 mcSCR-seq and other protocols, calculated as the number of ERCC molecules needed to reach a 50% detection probability as calculated in Svensson et al.<sup>5</sup>. Per-cell distributions are shown using violin plots with vertical lines and numbers indicating the median per protocol

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

Decisively, we find by using power simulations<sup>6,22</sup> that mcSCR-seq requires approximately half as many cells as SCR-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 mcSCR-seq compared to SCR-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 SCR-seq<sup>6</sup> to less than 0.54 € for mcSCR-seq (Supplementary Fig. 12a and Supplementary Table 3). Together, this makes mcSCR-seq sixfold more cost-efficient than SCR-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 SCR-seq was already one of the most cost-efficient protocols in our recent benchmarking study<sup>6</sup>, this likely makes mcSCR-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<sup>5</sup>. As in Svensson et al.<sup>5</sup>, 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).



mcSCR-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<sup>5,6</sup> as well as additional relevant protocols<sup>17,23</sup>, mcSCR-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.<sup>5</sup>, because the data were generated from many varying cell types sequenced in numerous labs. Similarly, mcSCR-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.

#### mcSCR-seq detects biological differences in complex tissues.

Finally, we applied mcSCR-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 mcSCR-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<sup>24</sup>, 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<sup>23</sup> (Fig. 3c, d). Overall, we show that mcSCR-seq is a powerful tool to highlight biological differences, already when a low number of cells are sequenced.

#### Discussion

In this work, we developed mcSCR-seq, a scRNA-seq protocol utilizing molecular crowding. Based on benchmarking data generated from mouse ES cells, we show that mcSCR-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<sup>23,25</sup>. Compared to the previous SCR-seq protocol, mcSCR-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 mcSCR-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 mcSCR-seq to deal with more difficult cell types or samples.

Taken together, mcSCR-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  $\mu$ l. 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/ $\mu$ l. 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), Enz-Script (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  $\times$  M-MuLV reaction buffer (NEB), 1 mM dNTPs (Thermo Fisher), 1  $\mu$ M E3V6NEXT barcoded oligo-dT primer (IDT), and 1  $\mu$ M 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<sub>2</sub>, betaine, trehalose, and polyethylene glycol (PEG 8000). The final reaction volume of 10  $\mu$ l was maintained by adjusting the volume of H<sub>2</sub>O.

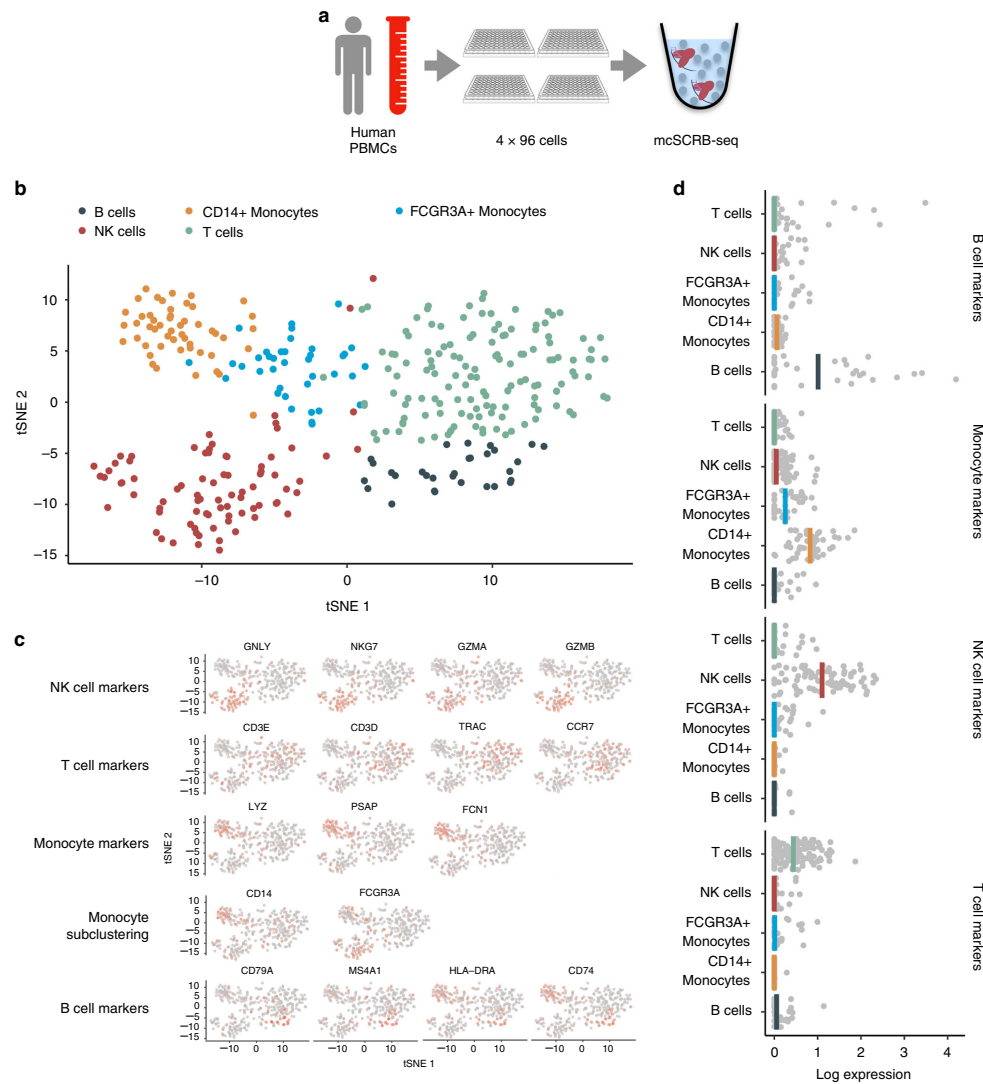
For this, we added increasing concentrations of MgCl<sub>2</sub> (3, 6, 9, and 12 mM; Sigma-Aldrich) in the RT buffer in the presence or absence of 1 M betaine (Sigma-Aldrich). Furthermore, the addition of 1 M betaine and 0.6 M trehalose (Sigma-Aldrich) was compared to the standard RT protocol. Lastly, 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), FidelityTaq (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  $\mu$ l 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 SCR-seq PCR cycling conditions<sup>5</sup>.

**Cell culture of mouse embryonic stem cells.** J1<sup>26</sup> and JM8<sup>27</sup> 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](http://www.eummcr.org)), respectively. They were used for the comparison of KAPA vs. Terra PCR amplification (Supplementary Fig. 5c) and the comparison of SCR-seq and mcSCR-seq, respectively. Both were cultured under feeder-free conditions on gelatin-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  $\mu$ g/ml streptomycin (Thermo Fisher), 2 mM L-glutamine (Thermo Fisher), 1  $\times$  MEM non-essential amino acids (NEAA, Thermo Fisher), 0.1 mM  $\beta$ -mercaptoethanol (Thermo Fisher), 1000 U/ml recombinant mouse LIF (Merck Millipore) and 2i (1  $\mu$ M PD032591 and 3  $\mu$ M CHIR99021 (Sigma-Aldrich)). mESCs were routinely passaged using 0.25% trypsin (Thermo Fisher).

## ARTICLE

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-05347-6



**Fig. 3** mSCR-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 mSCR-seq protocol, sequencing libraries were generated. **b** tSNE projection of PBMC cells ( $n = 349$ ) that were grouped into five clusters using the Seurat package<sup>24</sup>. 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<sup>28</sup>.

**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<sup>28</sup>.

**SCR-seq cDNA synthesis.** Cells were dissociated using trypsin and resuspended in 100  $\mu$ l 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  $\mu$ m 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  $-80^{\circ}\text{C}$ . Libraries were prepared as previously described<sup>6,9</sup>. 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  $\mu$ l reaction at  $42^{\circ}\text{C}$  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  $\mu$ l reaction volumes.

**mcSCR-seq cDNA synthesis.** A full step-by-step protocol for mcSCR-seq has been deposited in the protocols.io repository<sup>29</sup>. 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  $\mu$ l lysis buffer using a Sony SH800 sorter (Sony Biotechnology; 100  $\mu$ m chip). Lysis buffer consisted of a 1:500 dilution of Phusion HF buffer (New England Biolabs), 1.25  $\mu$ g/ $\mu$ l Proteinase K (Clontech), and 0.4  $\mu$ M barcoded oligo-dT primer (E3V6NEXT, IDT). After sorting, plates were immediately spun down and frozen at  $-80^{\circ}\text{C}$ . For libraries containing ERCCs, 0.1  $\mu$ l of 1:80,000 dilution of ERCC spike-in Mix 1 was used.

Before library preparation, proteins were digested by incubation at  $50^{\circ}\text{C}$  for 10 min. Proteinase K was then heat inactivated for 10 min at  $80^{\circ}\text{C}$ . Next, 5  $\mu$ l reverse transcription master mix consisting of 20 units Maxima H- enzyme (Thermo Fisher), 2  $\times$  Maxima H- Buffer (Thermo Fisher), 2 mM each dNTPs (Thermo Fisher), 4  $\mu$ M 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  $42^{\circ}\text{C}$ . Barcoded cDNA was then pooled in 2 ml DNA LoBind tubes (Eppendorf) and cleaned up using SPRI beads. Purified cDNA was eluted in 17  $\mu$ l and residual primers digested with Exonuclease I (Thermo Fisher) for 20 min at  $37^{\circ}\text{C}$ . After heat inactivation for 10 min at  $80^{\circ}\text{C}$ , 30  $\mu$ l PCR master mix consisting of 1.25 U Terra direct polymerase (Clontech) 1.66  $\times$  Terra direct buffer and 0.33  $\mu$ M SINGV6 primer (IDT) was added. PCR was cycled as given: 3 min at  $98^{\circ}\text{C}$  for initial denaturation followed by 15 cycles of 15 s at  $98^{\circ}\text{C}$ , 30 s at  $65^{\circ}\text{C}$ , 4 min at  $68^{\circ}\text{C}$ . Final elongation was performed for 10 min at  $72^{\circ}\text{C}$ .

**Library preparation.** Following preamplification, all samples were purified using SPRI beads at a ratio of 1:0.8 with a final elution in 10  $\mu$ l of  $\text{H}_2\text{O}$  (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 (P5NEXTPTS, 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<sup>14,30</sup>. 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 GRCh38.75). 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 mcSCR-seq using a Sony SH800 sorter (Sony Biotechnology; 100  $\mu$ m chip). cDNA was synthesized according to the mcSCR-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  $\mu$ l lysis buffer using a Sony SH800 sorter (Sony Biotechnology; 100  $\mu$ m 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  $\mu$ l reverse transcription master mix (see above) and 4  $\mu$ l  $\text{ddH}_2\text{O}$ . After the addition of 1  $\mu$ l 2  $\mu$ M barcoded oligo-dT primer (E3V6NEXT, IDT), cDNA was synthesized according to the mcSCR-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<sup>24</sup> 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.<sup>5</sup>, 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-seq<sup>17</sup>, we obtained expression values for ERCCs from Gene Expression Omnibus (GEO; GSE99866), sample GSM2656466; for Chromium<sup>23</sup> we obtained expression tables from the 10  $\times$  Genomics webpage (<https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/ercc>) and for SCR-seq, Smart-seq2, CEL-seq2/C1, MARS-seq and Smart-seq/C1<sup>6</sup>, 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<sup>22</sup>. Parameter estimation of the negative binomial distribution was done using scan normalized counts at 500,000 raw reads per cell<sup>31</sup>. 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<sup>32</sup> and scan normalization<sup>31</sup>.

**Batch effect analysis.** In order to detect genes differing between batches of one scRNA-seq protocol, data were normalized using scan<sup>31</sup>. Next, we tested for differentially expressed genes using limma-voom<sup>33,34</sup>. 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](https://github.com/cziegenhain/Bagnoli_2017).

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

## ARTICLE

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-05347-6

Further data including cDNA yield of optimization experiments is available on GitHub ([https://github.com/ziegenhain/Bagnoli\\_2017](https://github.com/ziegenhain/Bagnoli_2017)). A detailed step-by-step protocol for mcSCR-seq has been submitted to the protocols.io repository (mcSCR-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

- Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
- Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
- 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).
- Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I. & Enard, W. Quantitative single-cell transcriptomics. *Brief. Funct. Genomics* <https://doi.org/10.1093/bfpg/ely009> (2018).
- Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
- Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 (2017).
- 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/bfpg/ely001> (2018).
- 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).
- SEQ/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).
- Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
- Zimmerman, S. B. & Pfeiffer, 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).
- Rivas, G. & Minton, A. P. Macromolecular crowding in vitro, in vivo, and in between. *Trends Biochem. Sci.* **41**, 970–981 (2016).
- Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
- 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**, gty059 (2018).
- 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).
- Quail, M. A. et al. Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* **9**, 10–11 (2012).
- 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).
- Dixit, A. Correcting chimeric crosstalk in single cell RNA-seq experiments. Preprint at <https://doi.org/10.1101/093237> (2016).
- Baker, S. C. et al. The external RNA controls consortium: a progress report. *Nat. Methods* **2**, 731–734 (2005).
- Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
- Phipson, B., Zappia, L. & Oshlack, A. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res.* **6**, 595 (2017).
- 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).
- Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
- 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).
- 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).
- Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).
- Pettitt, S. J. et al. Agouti C57BL/6N embryonic stem cells for mouse genetic resources. *Nat. Methods* **6**, 493–495 (2009).
- Young, L., Sung, J., Stacey, G. & Masters, J. R. Detection of mycoplasma in cell cultures. *Nat. Protoc.* **5**, 929–934 (2010).
- Bagnoli, J., Ziegenhain, C., Janjic, A., Wange, L. E. & Vieth, B. mcSCR-seq protocol. <https://doi.org/10.17504/protocols.io.nrkdd4w> (2018).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- 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).
- 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).
- Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
- 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 SCR-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 iPSCs. 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., L.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 Information

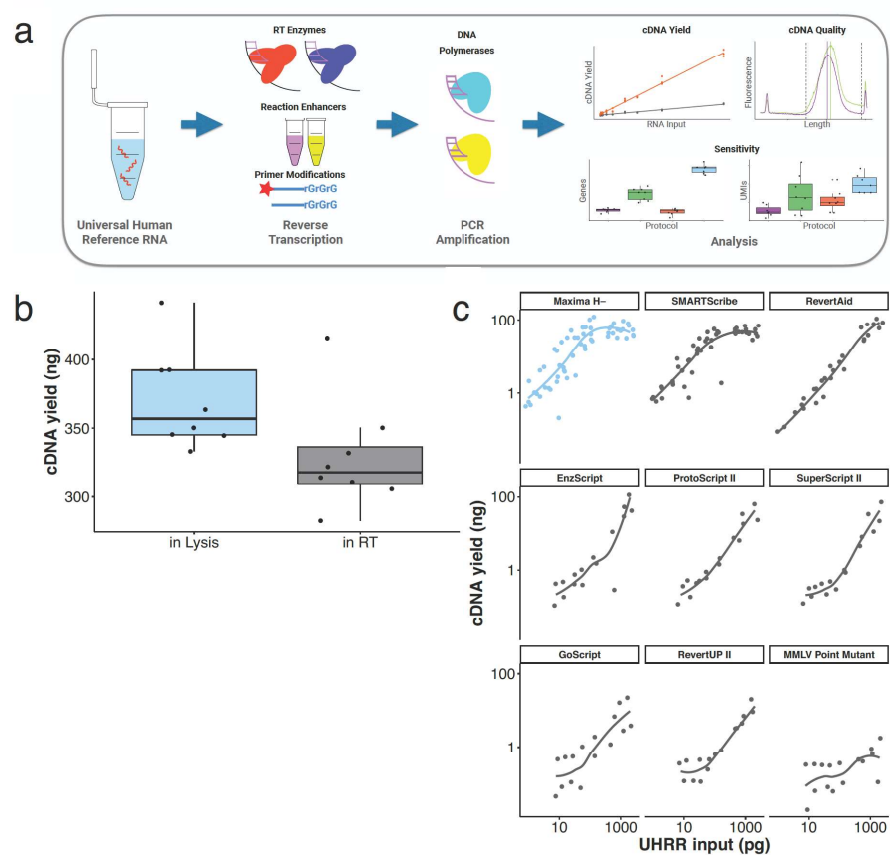
**Supplementary Information**

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

*Bagnoli et al.*



## Supplementary Figure 1



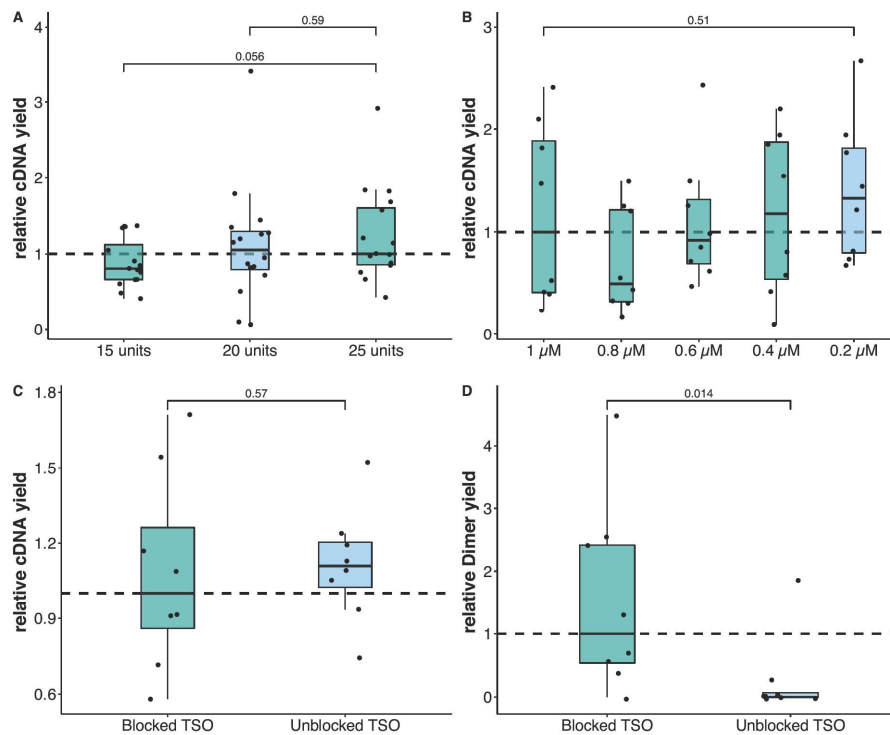
## Supplementary Figure 1: Schematic overview and optimization of reverse transcription

**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)

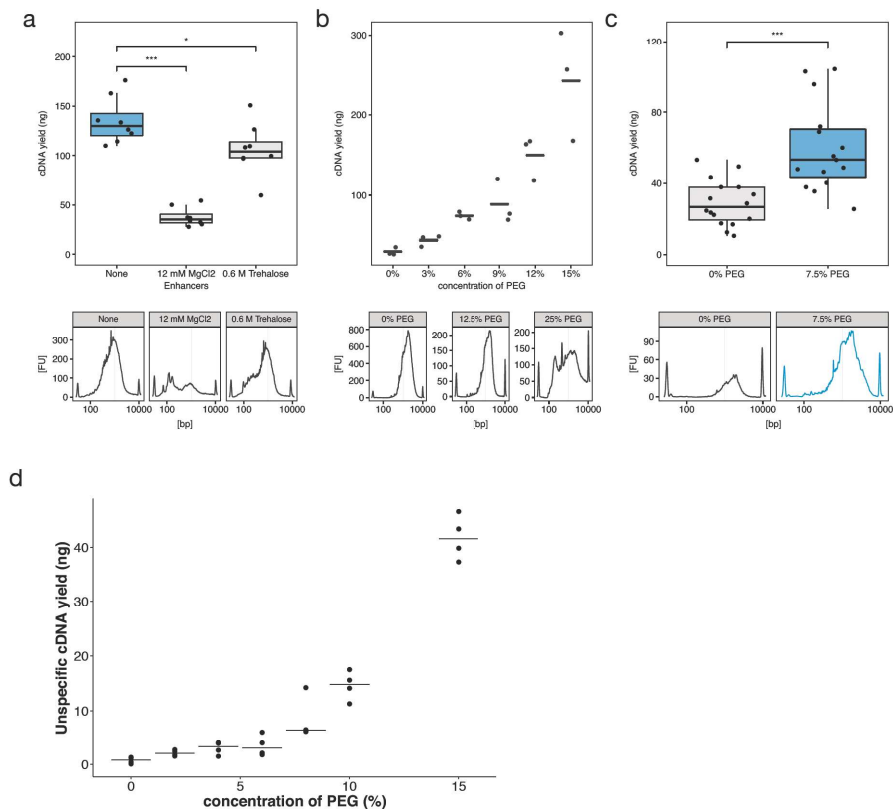
**d)** relative primer dimer yield using blocked or unblocked Template switching oligo (TSO, E5V6) estimated using no-input controls (see Methods).

All values are relative to the median of the condition used in the original SCR-seq protocol<sup>1</sup>, 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 mcSCR-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 MgCl<sub>2</sub> and Trehalose on cDNA synthesis (1 ng UHRR input per replicate; 21 PCR cycles).

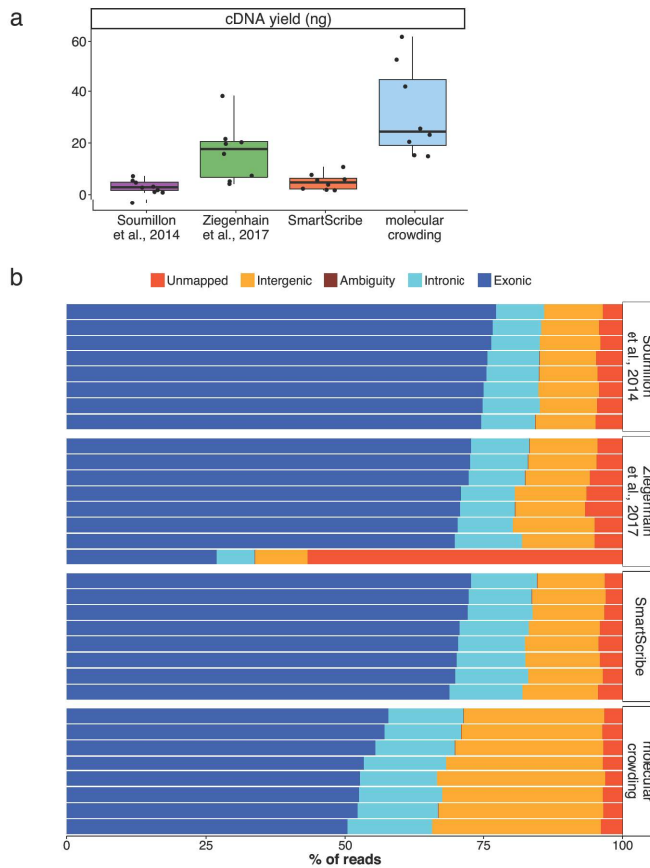
**b)** Concentration-dependent influence of PEG 8000 on cDNA yield (100 pg UHRR input per replicate; 23 PCR cycles).

**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.**

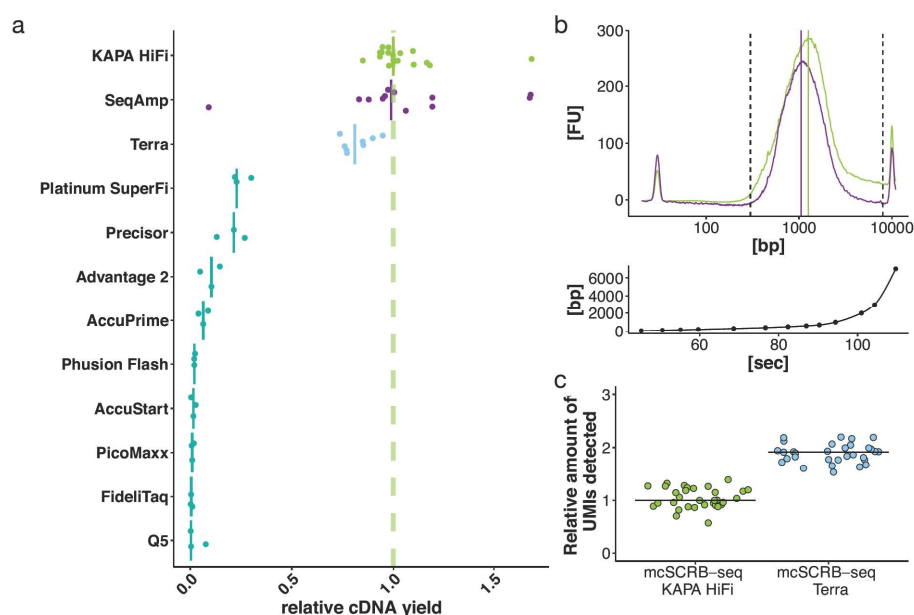
10 µg of UHRR were used as input for eight replicates for each of the four protocol variants (Supplementary Table 1).

**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 SCRBS-seq and mcSCRBS-seq protocols have the same fraction (~25%) of intergenic reads mapped when single mouse ES cells are used (Supplementary Figure 7c).

## Supplementary Figure 5

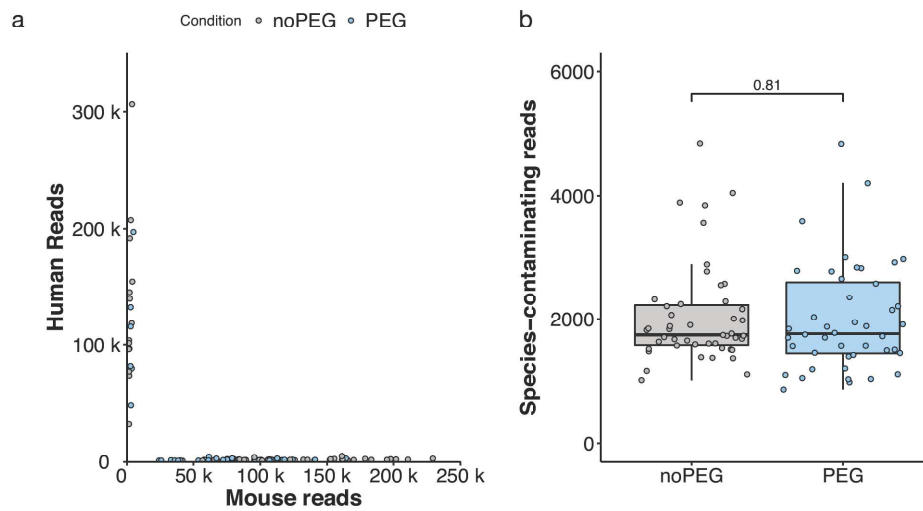
**Supplementary Figure 5: Optimization of PCR amplification.**

**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.<sup>2</sup>. 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

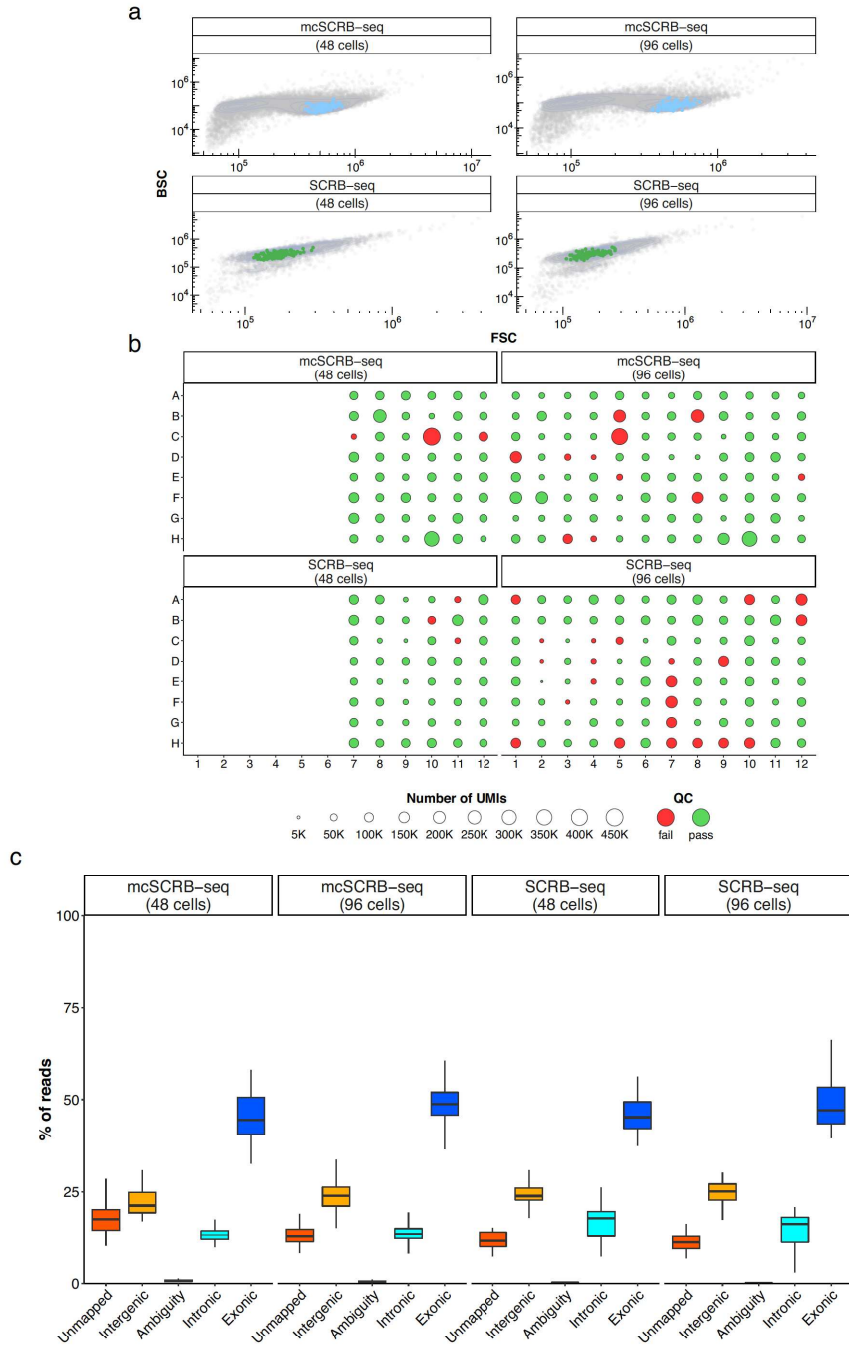
**Supplementary Figure 6: Species mixing experiment for mcSCR-seq**

Human induced pluripotent stem cells and Mouse embryonic stem cells were mixed and sorted in a 96-well plate. cDNA was synthesized using the mcSCR-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).

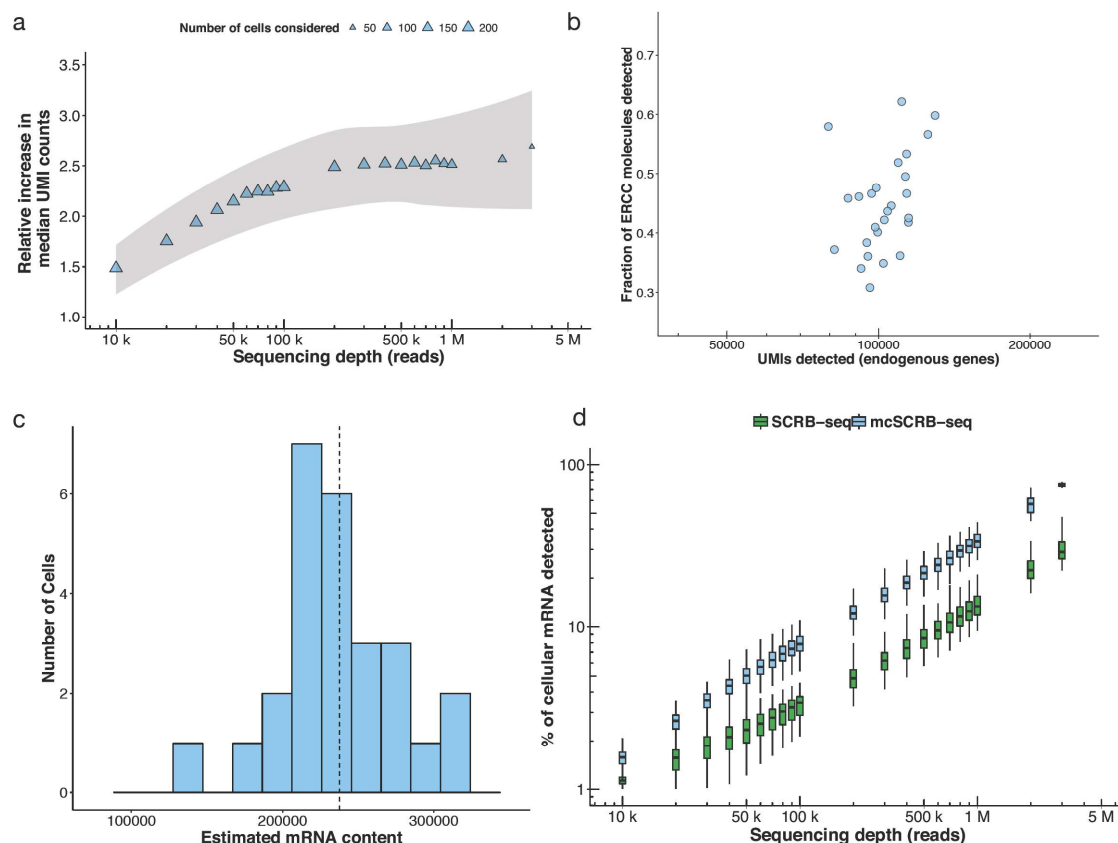
Supplementary Figure 7



**Supplementary Figure 7: Libraries from single mESCs generated with mcSCRB-seq and SCR-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 (SCR-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 (SCR-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

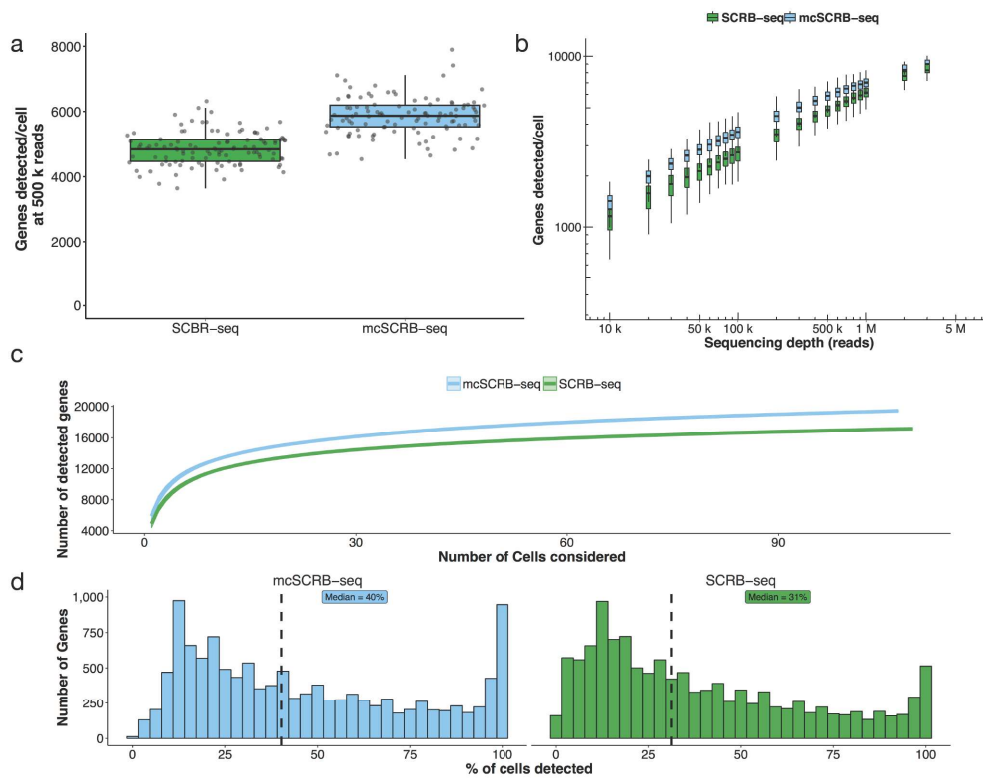
**Supplementary Figure 8: Sensitivity of SCRb-seq and mcSCRb-seq protocols.**

**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 were 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 SCR-seq and mcSCR-seq protocols by genes.**

**a)** Number of detected genes per cell and method (SCR-seq/mcSCR-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.

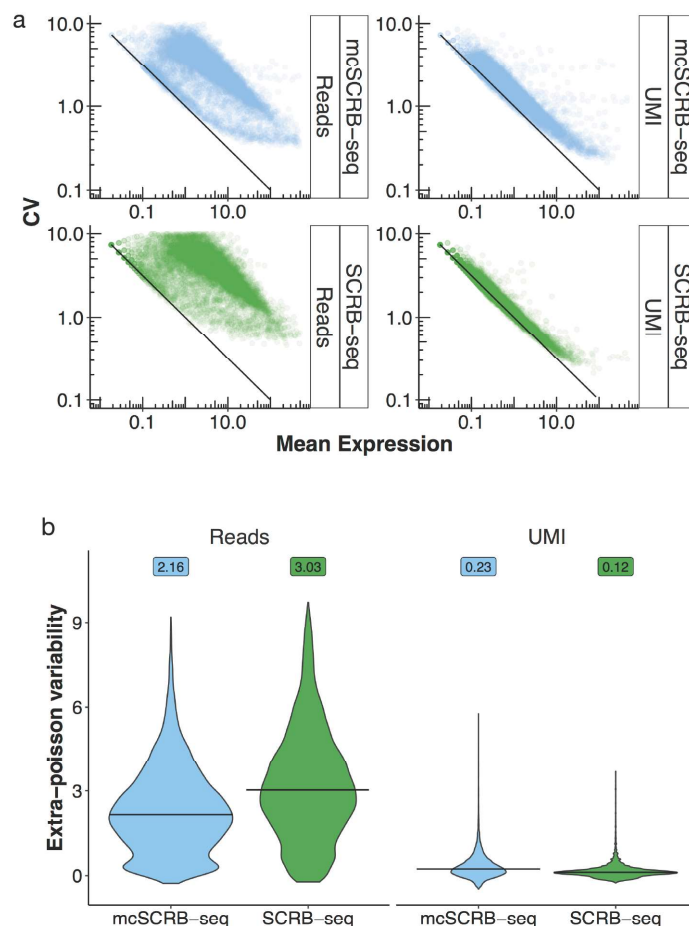
**b)** Number of detected genes per cell and method (SCR-seq/mcSCR-seq) dependent on sequencing depth (reads). Each box represents the median and first and third quartiles per sequencing depth and method. Sequencing depths and genes are plotted on a logarithmic axis (base 10).

**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



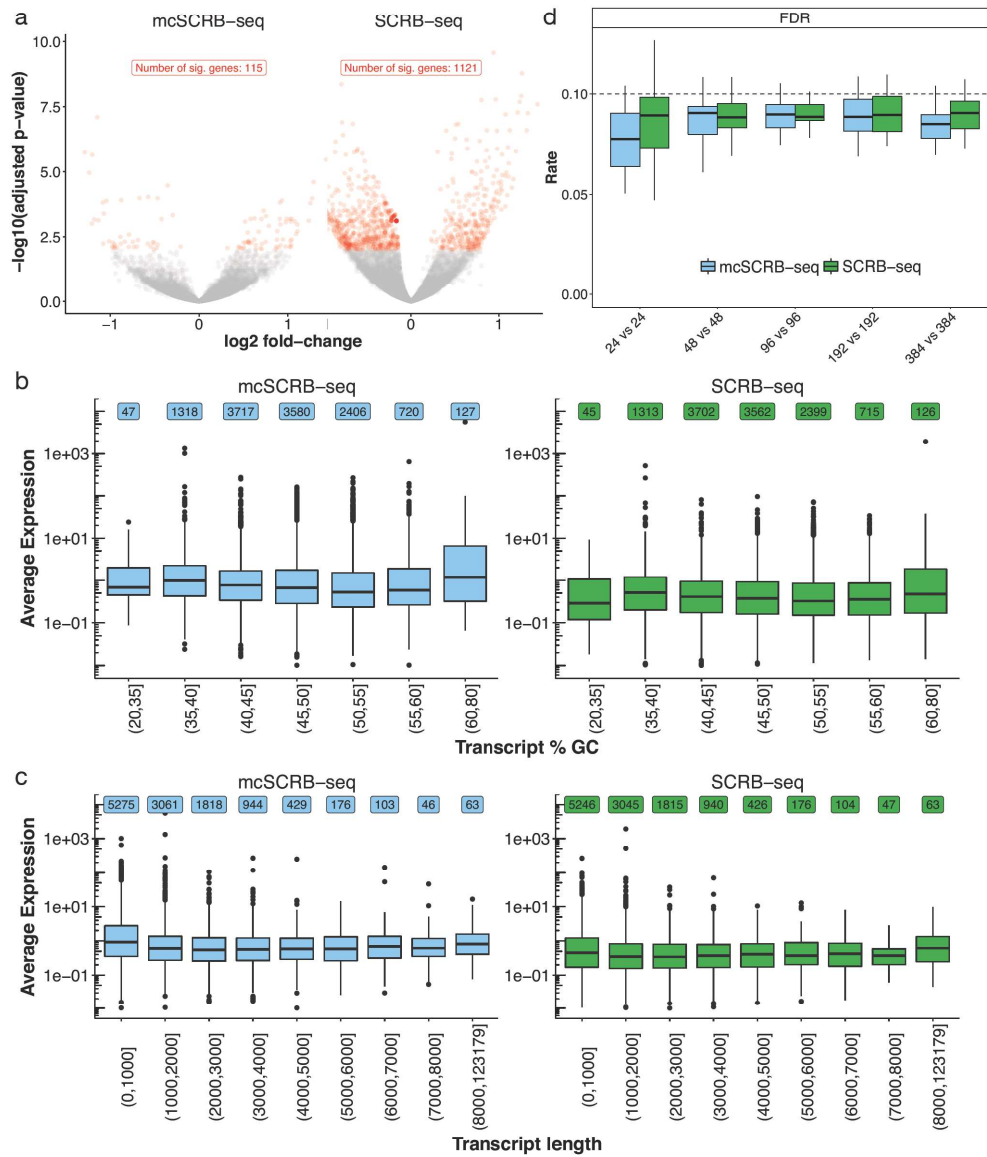
**Supplementary Figure 10: Variation parameters of SCRb-seq and mcSCRb-seq protocols by genes.**

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.

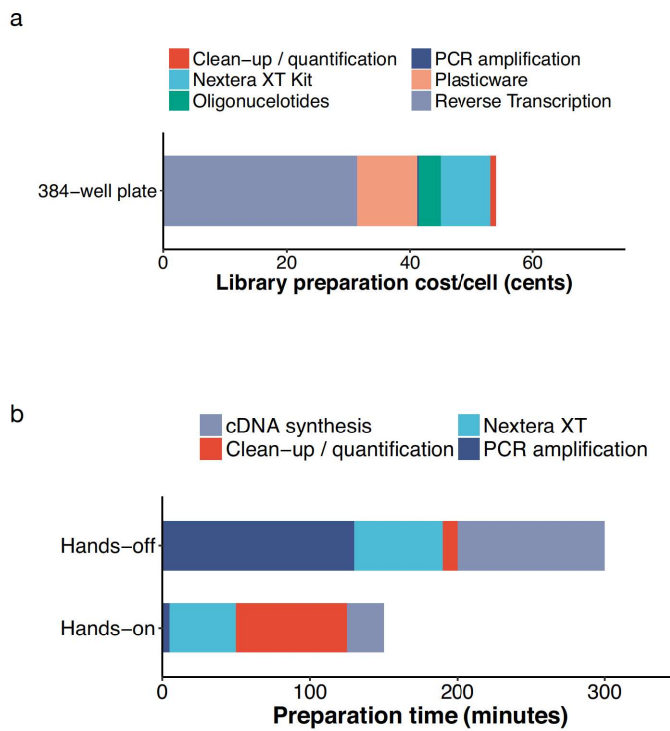
Supplementary Figure 11



**Supplementary Figure 11: Batch effects, biases and power analysis of SCR-seq and mcSCR-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<sup>3</sup> from empirical parameters estimated at 500,000 raw reads per cell. For SCR-seq and mcSCR-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.

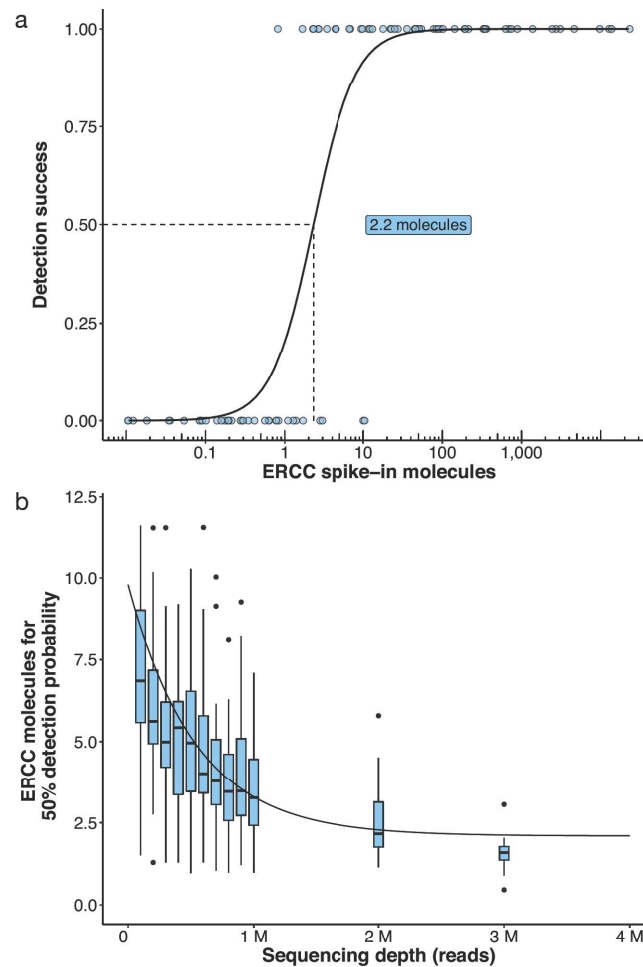
## Supplementary Figure 12

**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

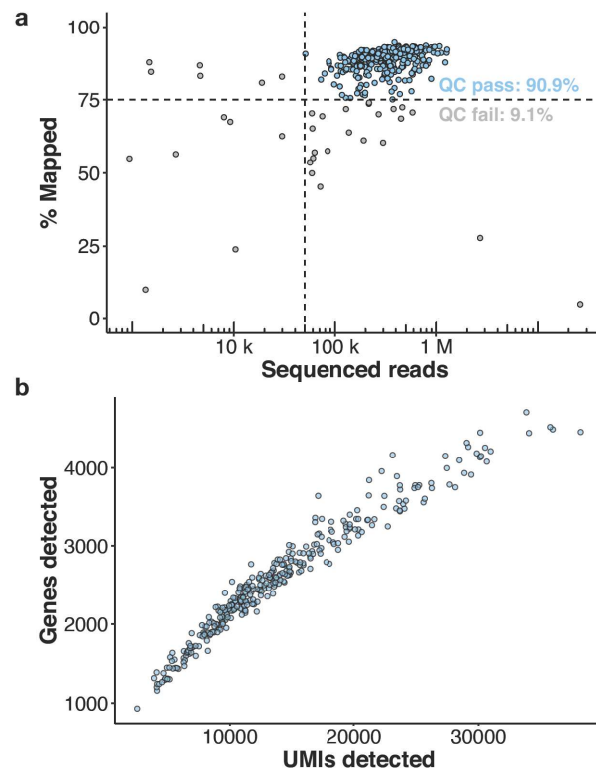


**Supplementary Figure 13 : Comparison of mcSCRb-seq to other scRNA-seq data based on ERCC spike-in detection probability**

**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 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 ( $\geq 1$  UMI) and detected UMIs are shown for all cells that passed quality control ( $n=349$ ).

Supplementary Table 1

<b>protocol variant</b>	<b>Soumillon</b>	<b>Ziegenhain</b>	<b>SmartScribe</b>	<b>molecular crowding</b>
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

	<b>SCRB-seq</b>	<b>mcSCRB-seq</b>
Lysis	Phusion HF	Phusion HF + Proteinase K + oligo-dT primers
Cell suspension	RNAprotect	PBS
Proteinase K	Ambion	Clontech
oligo-dT concentration	1 $\mu$ M	0.2 $\mu$ M
reverse transcription volume	2 $\mu$ l	10 $\mu$ l
RT amount	25 U	20 U
RT enhancer	none	7.5% PEG
TSO modification	5'-blocking	none
TSO concentration	1 $\mu$ M	2 $\mu$ M
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.<sup>2</sup> 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 €
<b>Total/cell</b>			<b>0,54 €</b>

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		
<b>total time</b>	<b>150</b>	<b>300</b>			

Supplementary Table 4. Detailed overview of hands-on and hands-off time necessary to create a sequenceable mcSCR-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



## 5.2 Benchmarking single-cell RNA-sequencing protocols for cell atlas projects

Mereu, Elisabetta; Lafzi, Atefeh; Moutinho, Catia; Ziegenhain, Christoph; McCarthy, Davis J; Álvarez-Varela, Adrián; Batlle, Eduard; Sagar, Grün, Dominic; Lau, Julia K; Boutet, Stéphane C; Sanada, Chad; Ooi, Aik; Jones, Robert C; Kaihara, Kelly; Brampton, Chris; Talaga, Yasha; Sasagawa, Yohei; Tanaka, Kaori; Hayashi, Tetsutaro; Braeuning, Caroline; Fischer, Cornelius; Sauer, Sascha; Trefzer, Timo; Conrad, Christian; Adiconis, Xian; Nguyen, Lan T; Regev, Aviv; Levin, Joshua Z; Parekh, Swati; Janjic, Aleksandar; **Wange, Lucas E**; Bagnoli, Johannes W; Enard, Wolfgang; Gut, Marta; Sandberg, Rickard; Nikaido, Itoshi; Gut, Ivo; Stegle, Oliver and Heyn, Holger

"Benchmarking single-cell RNA-sequencing protocols for cell atlas projects" (2020)

*Nature Biotechnology* 38, 747–755 (2020).

doi: <https://doi.org/10.1038/s41587-020-0469-4>

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

<https://www.nature.com/articles/s41587-020-0469-4#Sec35>

### Abstract

Single-cell RNA sequencing (scRNA-seq) is the leading technique for characterizing the transcriptomes of individual cells in a sample. The latest protocols are scalable to thousands of cells and are being used to compile cell atlases of tissues, organs and organisms. However, the protocols differ substantially with respect to their RNA capture efficiency, bias, scale and costs, and their relative advantages for different applications are unclear. In the present study, we generated benchmark datasets to systematically evaluate protocols in terms of their power to comprehensively describe cell types and states. We performed a multicenter study comparing 13 commonly used scRNA-seq and single-nucleus RNA-seq protocols applied to a heterogeneous reference sample resource. Comparative analysis revealed marked differences

in protocol performance. The protocols differed in library complexity and their ability to detect cell-type markers, impacting their predictive value and suitability for integration into reference cell atlases. These results provide guidance both for individual researchers and for consortium projects such as the Human Cell Atlas.



## Benchmarking single-cell RNA-sequencing protocols for cell atlas projects

Elisabetta Mereu<sup>1,26</sup>, Atefeh Lafzi<sup>1,26</sup>, Catia Moutinho<sup>1</sup>, Christoph Ziegenhain<sup>2</sup>, Davis J. McCarthy<sup>3,4,5</sup>, Adrián Álvarez-Varela<sup>6</sup>, Eduard Batlle<sup>6,7,8</sup>, Sagar<sup>9</sup>, Dominic Grün<sup>9</sup>, Julia K. Lau<sup>10</sup>, Stéphane C. Boutet<sup>10</sup>, Chad Sanada<sup>11</sup>, Aik Ooi<sup>11</sup>, Robert C. Jones<sup>12</sup>, Kelly Kaihara<sup>13</sup>, Chris Brampton<sup>13</sup>, Yasha Talaga<sup>13</sup>, Yohei Sasagawa<sup>14</sup>, Kaori Tanaka<sup>14</sup>, Tetsutaro Hayashi<sup>14</sup>, Caroline Braeuning<sup>15</sup>, Cornelius Fischer<sup>15</sup>, Sascha Sauer<sup>15</sup>, Timo Trefzer<sup>16</sup>, Christian Conrad<sup>16</sup>, Xian Adiconis<sup>17,18</sup>, Lan T. Nguyen<sup>17</sup>, Aviv Regev<sup>17,19,20</sup>, Joshua Z. Levin<sup>17,18</sup>, Swati Parekh<sup>21</sup>, Aleksandar Janjic<sup>22</sup>, Lucas E. Wange<sup>22</sup>, Johannes W. Bagnoli<sup>22</sup>, Wolfgang Enard<sup>22</sup>, Marta Gut<sup>1</sup>, Rickard Sandberg<sup>2</sup>, Itoshi Nikaido<sup>14,23</sup>, Ivo Gut<sup>1,24</sup>, Oliver Stegle<sup>3,4,25</sup> and Holger Heyn<sup>1,24</sup>✉

**Single-cell RNA sequencing (scRNA-seq) is the leading technique for characterizing the transcriptomes of individual cells in a sample. The latest protocols are scalable to thousands of cells and are being used to compile cell atlases of tissues, organs and organisms. However, the protocols differ substantially with respect to their RNA capture efficiency, bias, scale and costs, and their relative advantages for different applications are unclear. In the present study, we generated benchmark datasets to systematically evaluate protocols in terms of their power to comprehensively describe cell types and states. We performed a multicenter study comparing 13 commonly used scRNA-seq and single-nucleus RNA-seq protocols applied to a heterogeneous reference sample resource. Comparative analysis revealed marked differences in protocol performance. The protocols differed in library complexity and their ability to detect cell-type markers, impacting their predictive value and suitability for integration into reference cell atlases. These results provide guidance both for individual researchers and for consortium projects such as the Human Cell Atlas.**

Single-cell genomics provides an unprecedented view of the cellular makeup of complex and dynamic systems. Single-cell transcriptomic approaches in particular have led the technological advances that allow unbiased charting of cell phenotypes<sup>1</sup>. The latest improvements in scRNA-seq allow these technologies to scale to thousands of cells per experiment, providing comprehensive profiling of tissue composition<sup>2,3</sup>. This has led to the identification of new cell types<sup>4–6</sup> and the fine-grained description of cell plasticity in dynamic systems, such as development<sup>7,8</sup>. Recent large-scale efforts, such as the Human Cell Atlas (HCA) project<sup>9</sup>, are attempting to produce cellular maps of entire cell lineages, organs and organisms<sup>10,11</sup> by conducting phenotyping at the single-cell level. The HCA project aims to advance our understanding of tissue function and to serve as a reference for defining variation in

human health and disease. In addition to methods that capture the spatial organization of tissues<sup>12,13</sup>, the main approach being used is scRNA-seq analysis of dissociated cells. Therefore, tissues are disaggregated and individual cells captured either by cell sorting or using microfluidic systems<sup>1</sup>. In sequential processing steps, cells are lysed, the RNA is reverse transcribed to complementary DNA, amplified and processed to sequencing-ready libraries.

Continuous technological development has improved the scale, accuracy and sensitivity of scRNA-seq methods, and now allows us to create tailored experimental designs by selecting from a plethora of different scRNA-seq protocols. However, there are marked differences across these methods, and it is not clear which protocols are best for different applications. For large-scale consortium projects, experience has shown that neglecting benchmarking, standardization

<sup>1</sup>CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>2</sup>Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. <sup>3</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK. <sup>4</sup>European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. <sup>5</sup>St Vincent's Institute of Medical Research, Fitzroy, Victoria, Australia. <sup>6</sup>Institute for Research in Biomedicine, Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>7</sup>Catalan Institution for Research and Advanced Studies, Barcelona, Spain. <sup>8</sup>Centro de Investigación Biomédica en Red de Cáncer, Barcelona, Spain. <sup>9</sup>Max-Planck-Institute of Immunobiology and Epigenetics, Freiburg, Germany. <sup>10</sup>10x Genomics, Pleasanton, CA, USA. <sup>11</sup>Fluidigm Corporation, South San Francisco, CA, USA. <sup>12</sup>Department of Bioengineering, Stanford University, Stanford, CA, USA. <sup>13</sup>Bio-Rad, Hercules, CA, USA. <sup>14</sup>Laboratory for Bioinformatics Research, RIKEN Center for Biosystems, Dynamics Research, Saitama, Japan. <sup>15</sup>Max Delbrück Center for Molecular Medicine/Berlin Institute of Health, Berlin, Germany. <sup>16</sup>Digital Health Center, Berlin Institute of Health, Charité-Universitätsmedizin Berlin, Berlin, Germany. <sup>17</sup>Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>18</sup>Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>19</sup>Koch Institute of Integrative Cancer Research, MIT, Cambridge, MA, USA. <sup>20</sup>Howard Hughes Medical Institute, Department of Biology, MIT, Cambridge, MA, USA. <sup>21</sup>Max-Planck-Institute for Biology of Ageing, Cologne, Germany. <sup>22</sup>Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Martinsried, Germany. <sup>23</sup>School of Integrative and Global Majors, University of Tsukuba, Wako, Saitama, Japan. <sup>24</sup>Universitat Pompeu Fabra, Barcelona, Spain. <sup>25</sup>Division of Computational Genomics and Systems Genetics, German Cancer Research Center, Heidelberg, Germany. <sup>26</sup>These authors contributed equally: Elisabetta Mereu, Atefeh Lafzi. ✉e-mail: [holger.heyne@cnag.crg.eu](mailto:holger.heyne@cnag.crg.eu)

## ANALYSIS

## NATURE BIOTECHNOLOGY

and quality control at the start can lead to major problems later on in the analysis of the results<sup>14</sup>. Thus, success depends critically on implementing a high common standard. A comprehensive comparison of available scRNA-seq protocols will benefit both large- and small-scale applications of scRNA-seq.

The available scRNA-seq protocols vary in the efficiency of RNA-molecule capture, which results in differences in sequencing library complexity and the sensitivity of the method to identify transcripts and genes<sup>15–17</sup>. There has been no systematic testing of how their performance varies between cell types, and how this affects the resolution of cell phenotyping in complex samples. In the present study, we extend previous efforts to compare the molecule-capture efficiency of scRNA-seq protocols<sup>15,16</sup> by systematically evaluating the capability of these techniques to describe tissue complexity and their suitability for creating a cell atlas. We performed a multicenter benchmarking study to compare scRNA-seq protocols using a unified reference sample resource. Our reference sample contained: (1) a high degree of cell-type heterogeneity with various frequencies, (2) closely related subpopulations with subtle differences in gene expression, (3) a defined cell composition with trackable markers and (4) cells from different species. By analyzing human peripheral blood and mouse colon tissue, we have covered a broad range of cell types and states from cells in suspension and solid tissues, to represent common scenarios in cell atlas projects. We have also added spike-in cell lines to allow us to assess batch effects, and have combined different species to pool samples into a single reference. We performed a comprehensive comparative analysis of 13 different scRNA-seq protocols, representing the most commonly used methods. We applied a wide range of different quality control metrics to evaluate datasets from different perspectives, and to test their suitability for producing a reproducible, integrative and predictive reference cell atlas.

We observed striking differences among protocols in converting RNA molecules into sequencing libraries. Varying library complexities affected the protocol's power to quantify gene expression levels and to identify cell-type markers, a trend consistently observed across cell and tissue types. This critically impacted on the resolution of tissue profiles and the predictive value of the datasets. Protocols further differed in their capability to be integrated into reference tissue atlases and, thus, their suitability for consortium-driven projects with flexible production designs.

### Results

**Reference sample and experimental design.** We benchmarked current scRNA-seq protocols to inform the methodological selection process of cell atlas projects. Ideally, methods should: (1) be accurate and free of technical biases, (2) be applicable across distinct cell properties, (3) fully disclose tissue heterogeneity, including subtle differences in cell states, (4) produce reproducible expression profiles, (5) comprehensively detect population markers, (6) be integratable with other methods and (7) have predictive value with cells mapping confidently to a reference atlas.

For a systematic comparison of protocols, we designed a reference sample containing human peripheral blood mononuclear cells (PBMCs) and mouse colon, which are tissue types with highly heterogeneous cell populations, as determined by previous single-cell sequencing studies<sup>18,19</sup>. In addition to the well-defined cell types, the tissues contain cells in transition states (for example, colon transit-amplifying (TA) or enterocyte progenitor cells) that show transcriptional differences during their differentiation trajectory<sup>20</sup>. The reference sample also included a wide range of cell sizes (for example, B cells: ~7  $\mu\text{m}$ ; HEK293 cells: ~15  $\mu\text{m}$ ) and RNA content, which are key parameters that affect performance in cell capture and library preparation. Interrogation of tissues from different species allowed us to pool a large variety of cell types in a single reference sample to maximize complexity while minimizing variability

introduced during sample preparation. In addition to the intra-tissue complexity, the fluorescence-labeled, spiked-in cell lines allowed us to monitor cell-type composition during sample processing, and to identify batch effects and biases introduced during cell capture and library preparation.

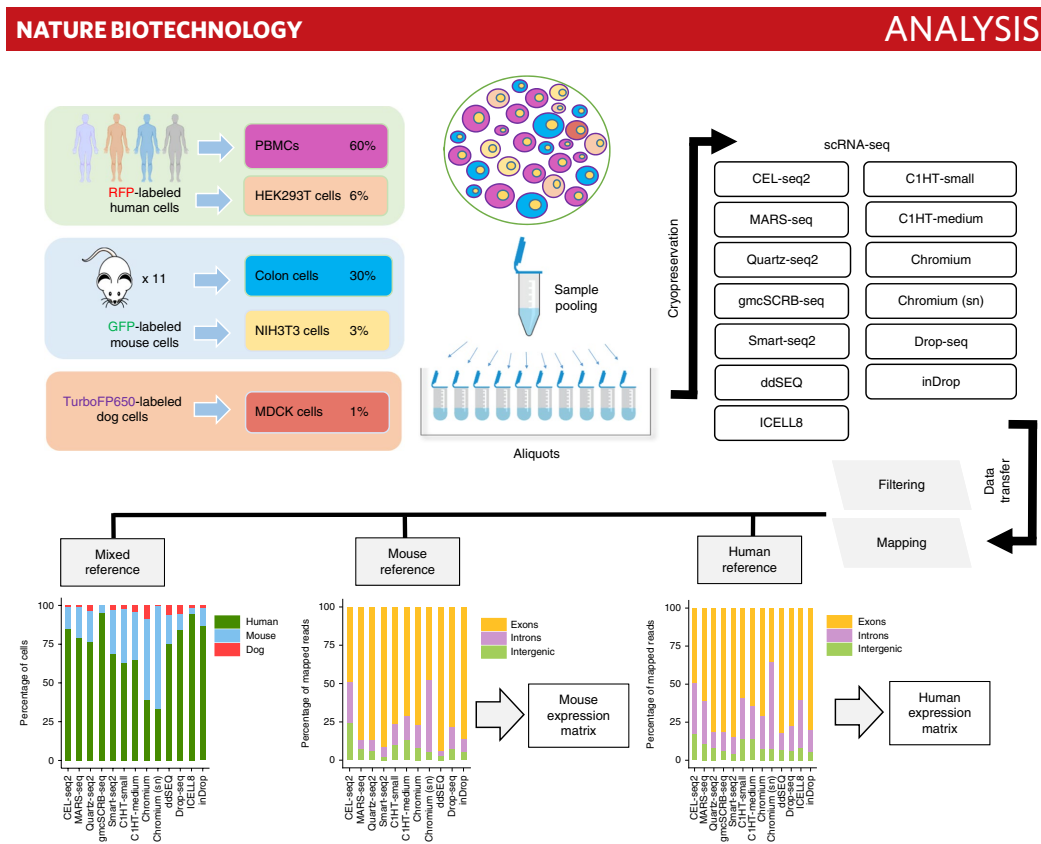
Specifically, the reference sample contained (estimated percentage viable cells): PBMCs (60%, human), colon cells (30%, mouse), HEK293T cells (6%, red fluorescent protein (RFP)-labeled human cell line), NIH3T3 cells (3%, green fluorescent protein (GFP)-labeled mouse cells) and MDCK cells (1%, TurboFP650-labeled dog cells) (Fig. 1). To reduce variability due to technical effects during library preparation, the reference sample was prepared in a single batch, distributed into aliquots of 250,000 cells and cryopreserved. We have previously shown that cryopreservation is suitable for single-cell transcriptomic studies of these tissue types<sup>21</sup>. For cell capture and library preparation, the thawed samples underwent FACS to remove damaged cells and physical doublets (see the next section for detailed analysis of cell viability sorting).

**A reference dataset for benchmarking experimental and computational protocols.** To obtain sufficient sensitivity to capture low-frequency cell types and subtle differences in the cell state, we profiled ~3,000 cells with each scRNA-seq protocol. In total, we produced datasets for five microtiter plate-based methods and seven microfluidic systems, including cell-capture technologies based on droplets (four), nanowells (one) and integrated fluidic circuits, to capture small (one) and medium (one)-sized cells (Fig. 1 and see Supplementary Table 1). We also included experiments to produce single-nucleus RNA-sequencing (snRNA-seq) libraries (one), and an experimental variant that profiled >50,000 cells to produce a reference of our complex sample. The unified sample resource and standardized sample preparation (see Methods) were designed largely to eliminate sampling effects and allow the systematic comparison of scRNA-seq protocol performance.

To compare the different protocols, and to create a resource for the benchmarking and development of computational tools (for example, batch effect correction, data integration and annotation), all datasets were processed in a uniform manner. Therefore, we designed a streamlined, primary data-processing pipeline tailored to the peculiarities of the reference sample (see Methods). Briefly, raw sequencing reads were mapped to a joint human, mouse and canine reference genome, and separately to their respective references to produce gene count matrices for subsequent analysis (accession no. GSE133549). Overall, we detected human, mouse and canine cell numbers consistent with the composition design of the reference sample (Fig. 1). However, some protocols varied markedly from the expected frequencies in human (34–95%), mouse (4–66%) and canine (0–9%) cells. Although the reference sample was prepared in a standardized way, we cannot entirely exclude the introduction of composition variability during sample handling. Thus, the subsequent evaluation of protocol performance was performed on cell types and states common to all protocols.

Notably, we observed a higher fraction of mouse colon cells in unsorted (Chromium) and the snRNA-seq datasets (Chromium (sn)). This probably results from damaging the more fragile colon cells during sample preparation, resulting in proportionally fewer colon cells when selecting for cell viability. To test whether this composition bias in scRNA-seq can be avoided by skipping viability selection, we generated matched datasets either selecting or not selecting for intact cells. After quality control the detection of mouse colon cells increased proportionally without viability selection (51% versus 19%), with good-quality cells showing comparable library complexity in both libraries (for example, numbers of detected genes; see Supplementary Figs. 1 and 2). However, considerably more cells were removed during quality filtering (44% versus 15%), and this is a source of unwanted sequencing costs that





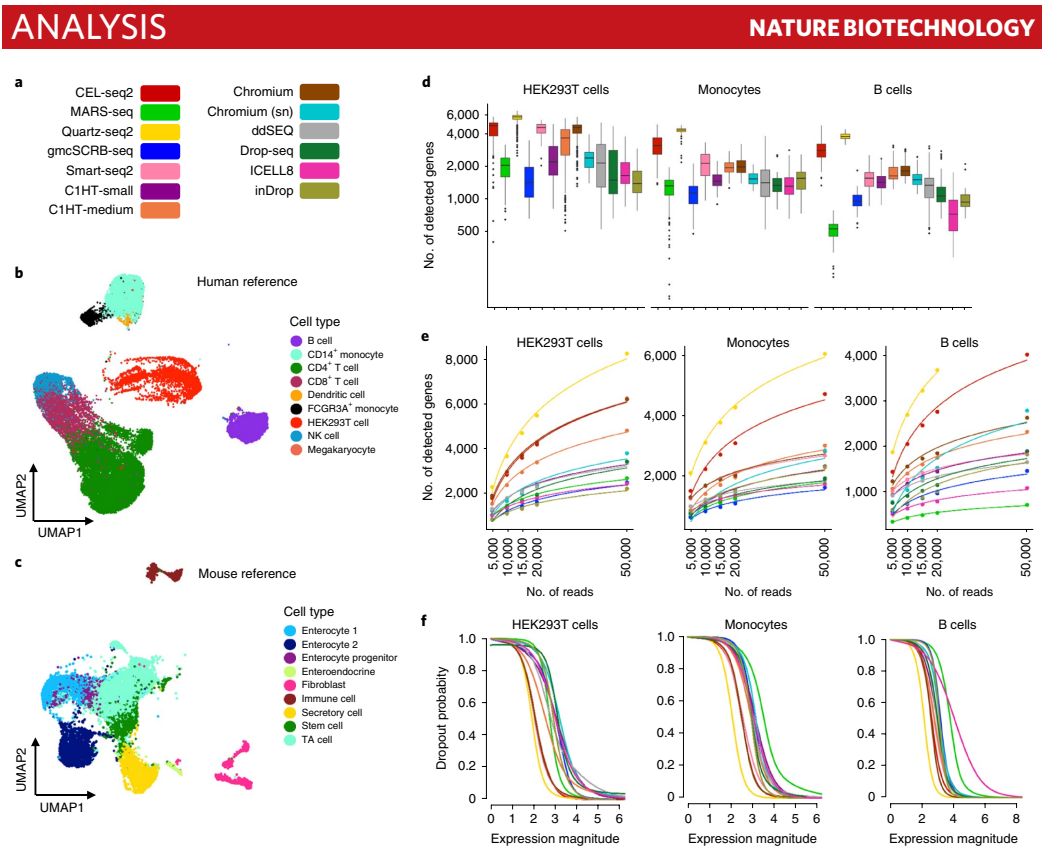
**Fig. 1 | Overview of the experimental design and data processing.** The reference sample consists of human PBMCs (60%), and HEK293T (6%), mouse colon (30%), NIH3T3 (3%) and dog MDCK cells (1%). The sample was prepared in one single batch, cryopreserved and sequenced by 13 different sc/snRNA-seq methods. Sequences were uniformly mapped to a joint human, mouse and canine reference, and then separately to produce gene expression counts for each sequencing method.

must be taken into account, especially for tissues with high cell damage. Consequently, replacing viability staining with thorough in silico quality filtering in cell atlas experiments might better conserve the composition of the original tissue, but result in higher sequencing costs.

The canine cells, spiked-in at a low concentration, were detected by all protocols (1–9%) except gmcSCR-seq. Furthermore, the different methods showed notable differences in mapping statistics between different genomic locations (Fig. 1). As expected, due to the presence of unprocessed RNA in the nucleus, the snRNA-seq experiment detected the highest proportion of introns, although scRNA-seq protocols also showed high frequencies of intronic and intergenic mappings. The increased detection of unprocessed transcripts in CEL-seq2 may be due to a freezing step (–80 °C) after cell isolation and subsequent denaturation at high temperatures (95 °C), which could favor the accessibility of nuclear and chromatin-bound RNA molecules.

**Molecule-capture efficiency and library complexity.** We produced reference datasets by analyzing 30,807 human and 19,749 mouse cells (Chromium v.2; Fig. 2a–c). The higher cell number allowed us to annotate the major cell types in our reference sample, and to extract population-specific markers (see Supplementary Table 2).

It was noteworthy that the reference samples solely provided the basis to assign cell identities and gene marker sets, and were not used to quantify the method’s performance. This strategy ensured that the choice of technology for deriving the reference does not influence downstream analyses. Cell clustering and reference-based cell annotation showed high agreement (average 83%; see Supplementary Table 3), and only cells with consistent annotations were used subsequently for comparative analysis at the cell-type level. The PBMCs (human) and colon cells (mouse) represented two largely different scenarios. Although the differentiated PBMCs clearly separated into subpopulations (for example, T/B cells, monocytes; Fig. 2b, and see Supplementary Figs. 3a and 4a–d), colon cells were ordered as a continuum of cell states that differentiate from intestinal stem cells into the main functional units of the colon (that is, absorptive enterocytes and secretory cells; Fig. 2c, and see Supplementary Figs. 3b and 5a–d). Notably, the subpopulation structure of our references was largely consistent with that of published datasets for human PBMCs<sup>18</sup> and mouse colon cells<sup>22</sup> (see Supplementary Figs. 6 and 7). After identifying major subpopulations and their respective markers in our reference sample, we clustered the cells of each sc/snRNA-seq protocol and annotated cell types using matchScore2 (see Methods). This algorithm allows a gene marker-based projection of single cells (cell by cell) on to a



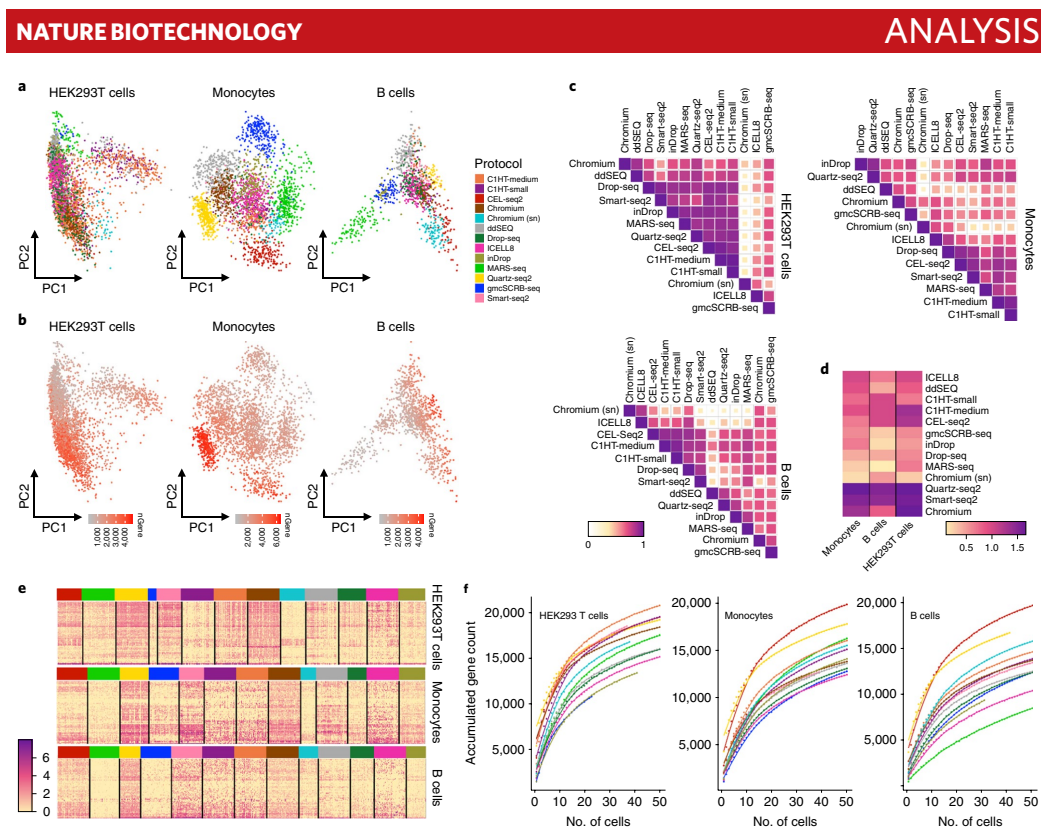
**Fig. 2 | Comparison of 13 sc/snRNA-seq methods.** **a**, Color legend of sc/snRNA-seq protocols. **b**, UMAP of 30,807 cells from the human reference sample (Chromium) colored by cell-type annotation. **c**, UMAP of 19,749 cells from the mouse reference (Chromium) colored by cell-type annotation. **d**, Boxplots displaying the minimum, the first, second and third quantiles, and the maximum number of genes detected across the protocols, in down-sampled (20,000) HEK293T cells, monocytes and B cells. Cell identities were defined by combining the clustering of each dataset and cell projection on to the reference. **e**, Number of detected genes at stepwise, down-sampled, sequencing depths. Points represent the average number of detected genes as a fraction of all cells of the corresponding cell type at the corresponding sequencing depth. **f**, Dropout probabilities as a function of expression magnitude, for each protocol and cell type, calculated on down-sampled data (20,000) for 50 randomly selected cells.

reference sample and, thus, the identification of cell types in our datasets (see Supplementary Figs. 8 and 9).

To compare the efficiency of messenger RNA capture between protocols, we down-sampled the sequencing reads per cell to a common depth and stepwise-reduced fractions. Stochasticity introduced during down-sampling did not affect the reproducibility of the results (see Supplementary Fig. 10). Library complexity was determined separately for largely homogeneous cell types with markedly different cell properties and function, namely human HEK293T cells, monocytes and B cells (Fig. 2d,e), and mouse colon secretory and TA cells (see Supplementary Fig. 11a,b). We observed large differences in the number of detected genes and molecules across the protocols, with consistent trends across cell types and gene quantification strategies (see Supplementary Fig. 11c,d). Notably, some protocols, such as Smart-seq2 and Chromium v2, performed better with higher RNA quantities (HEK293T cells) compared with lower starting amounts (monocytes and B cells), suggesting an input-sensitive optimum. Considering the different assay versions and application types of the Chromium system, a dedicated analysis showed

increased detection of molecules and genes from nuclei to intact cells and toward the latest protocol versions (see Supplementary Fig. 12). Consistent with the variable library complexity, the protocols presented large differences in dropout probabilities (Fig. 2f), with Quartz-seq2, Chromium v2 and CEL-seq2 showing consistently lower probability. Note that, despite the considerable differences between protocols, we observed a generally high technical reproducibility within the methods (see Supplementary Fig. 13).

**Technical effects and information content.** We further assessed the magnitude of technical biases, and the protocol's ability to describe cell populations. To quantify the technical variation within and across protocols, we selected highly variable genes (HVGs) across all datasets, and plotted the variation in the main principal components (PCs; Fig. 3a). Using the down-sampled data for HEK293T cells, monocytes and B cells, we observed strong protocol-specific profiles, with the main source of variability being the number of genes detected per cell (Fig. 3b). Data from snRNA-seq did not show notable outliers, indicating conserved representation of the



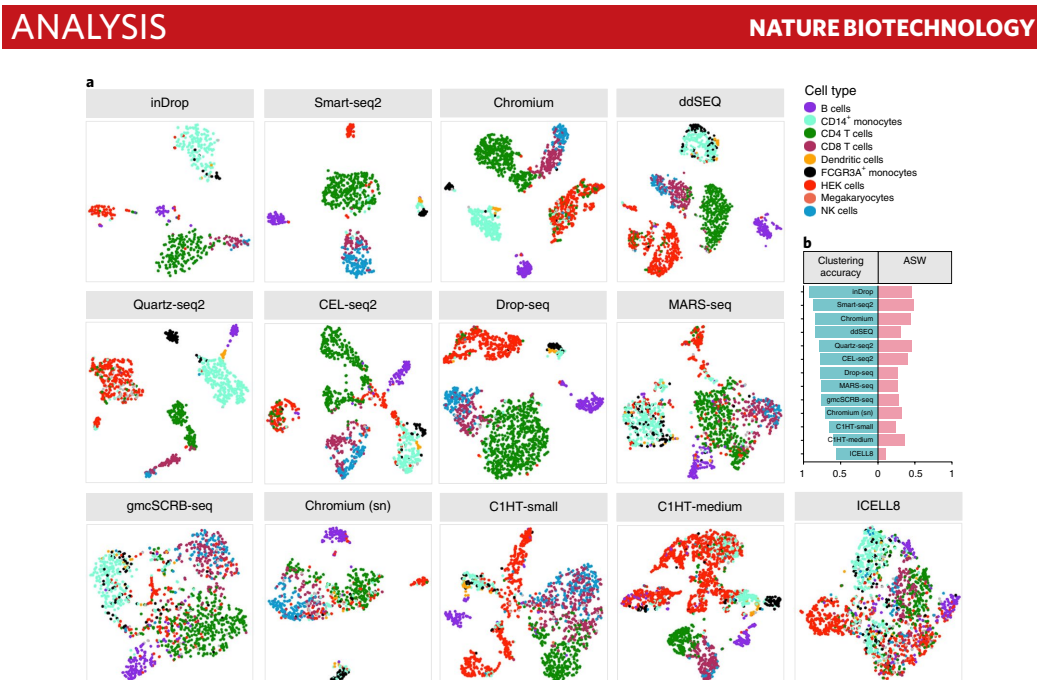
**Fig. 3 | Similarity measures of sc/snRNA-seq methods.** **a, b**, Principal component analysis on down-sampled data (20,000) using highly variable genes between protocols, separated into HEK293T cells, monocytes and B cells, and color coded by protocol (**a**) and number of detected genes per cell (**b**). **c**, Pearson's correlation plots across protocols using expression of common genes. For a fair comparison, cells were down-sampled to the same number for each method (B cells,  $n=32$ ; monocytes,  $n=57$ ; HEK293T cells,  $n=55$ ). Protocols are ordered by agglomerative hierarchical clustering. **d**, Average  $\log(\text{expression})$  values of cell-type-specific reference markers for down-sampled (20,000) HEK293T cells, monocytes and B cells. **e**,  $\log(\text{expression})$  values of reference markers on down-sampled data (20,000) for HEK293T cells, monocytes and B cells (maximum of 50 random cells per technique). **f**, Cumulative gene counts per protocol as the average of 100 randomly sampled HEK293T cells, monocytes and B cells, separately on down-sampled data (20,000).

transcriptome between the cytoplasm and the nucleus. To quantify the protocol-related variance, we identified the PCs that correlated with the protocol's covariates in a linear model<sup>23</sup>. Indeed, the variance in the data was mainly explained by the protocols (HEK293T cells = 37.3%, monocytes = 52.8% and B cells = 36.2%), a value that was reduced in HEK293T cells and monocytes when considering snRNA-seq as a specific covariate (HEK293T cells = 9.7%, monocytes = 22.2% and B cells = 48.3%; see Methods). The technical effects were also visible when using *t*-distributed stochastic neighbor embedding (tSNE) as a nonlinear, dimensionality reduction method (see Supplementary Fig. 14). By contrast, the methods largely mixed when the analysis was restricted to cell-type-specific marker genes, suggesting a conserved cell identity profile across techniques (see Supplementary Fig. 15).

Next, we quantified the similarities in information content of the protocols. Again, we used the down-sampled datasets and commonly expressed genes and calculated the correlation between methods in average transcript counts across multiple cells, thus compensating for the sparseness of single-cell transcriptome data.

For the three human cell types, we observed a broad spectrum of correlation across technologies, with generally lower correlation for smaller cell types (Fig. 3c). Although the transcriptome representation was generally conserved (Fig. 3a), the snRNA-seq protocol resulted in a notable outlier when correlating the expression levels of common genes across protocols, possibly driven by decreased correlation of immature transcripts. Restricting the correlation analysis to population-specific marker genes, we observed less variation between protocols (Pearson's  $r=0.5-0.7$ ), which underlines that the expression of these markers is largely conserved across the methods (see Supplementary Fig. 16).

To further test the suitability of protocols for describing cell types, we determined their sensitivity to detect population-specific expression signatures, and found that they had remarkably variable power to detect marker genes. Specifically, population markers were detected with different accuracies (see Supplementary Figs. 17 and 18), and the detection level varied substantially (Fig. 3d,e and see Supplementary Table 4). Quartz-seq2 and Smart-seq2 showed high expression levels for all cell-type signatures, indicating that they



**Fig. 4 | Clustering analysis of 13 sc/snRNA-seq methods on down-sampled datasets (20,000).** **a**, The tSNE visualizations of unsupervised clustering in human samples from 13 different methods. Each dataset was analyzed separately after down-sampling to 20,000 reads per cell. Cells are colored by cell type inferred by matchScore2 before down-sampling. Cells that did not achieve a probability score of 0.5 for any cell type were considered unclassified. **b**, Clustering accuracy and ASW for clusters in each protocol.

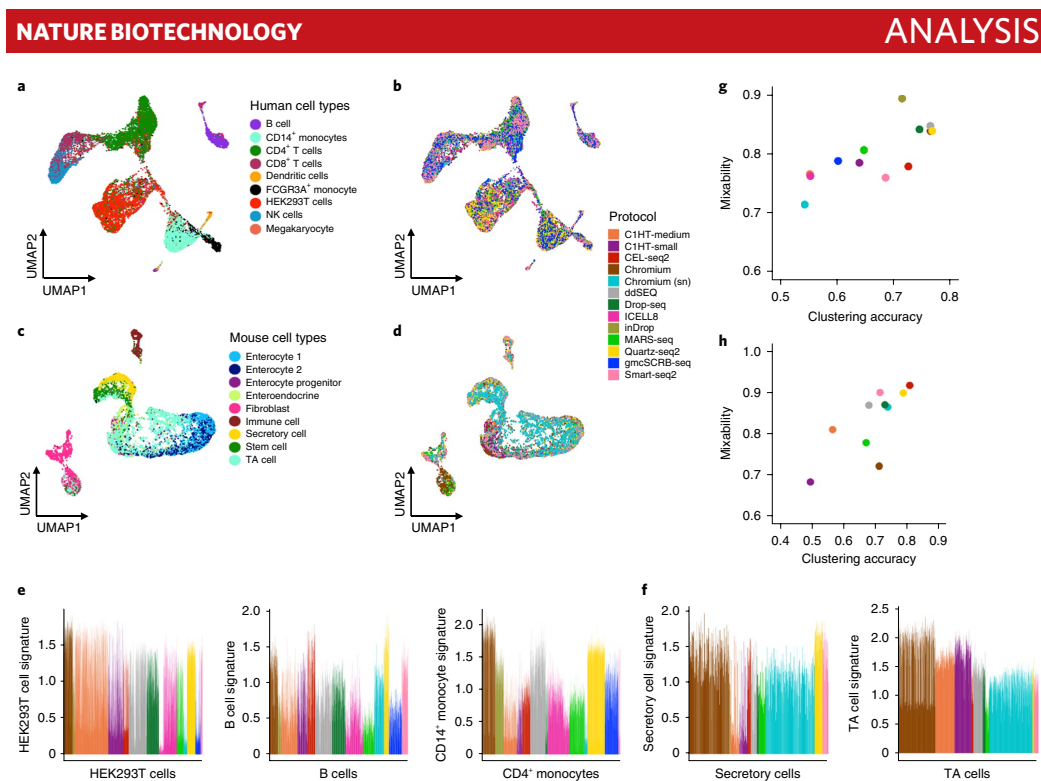
have higher power for cell-type identification. As marker genes are particularly important for data interpretation (for example, annotation), low marker detection levels could severely limit the interpretation of poorly explored tissues, or when trying to identify subtle differences across subpopulations. SnRNA-seq showed generally lower marker detection levels. However, gene markers were selected from intact cell experiments, which could lead to an underestimation of the performance of snRNA-seq to identify cell-type-specific signatures in this analysis approach.

The protocols also detected vastly different total numbers of genes when accumulating transcript information over multiple cells, with strong positive outliers observed for the smaller cell types (Fig. 3f). In particular, CEL-seq2 and Quartz-seq2 identified many more genes than other methods. Intriguingly, CEL-seq2 outperformed all other methods by detecting many weakly expressed genes; genes detected specifically by CEL-seq2 had significantly lower expression than the common genes detected by Quartz-seq2 ( $P < 2.2 \times 10^{-16}$ ). The greater sensitivity to weakly expressed genes makes this protocol particularly suitable for describing cell populations in detail, an important prerequisite for creating a comprehensive cell atlas and functional interpretation.

Surprisingly, considering the increased library complexity of scRNA-seq compared with snRNA-seq, the latter protocol identified a similar number of genes when combining information across multiple cells and suggesting overall similar transcriptome complexity of the two compartments (see Supplementary Fig. 12). ScRNA-seq detected additional genes enriched in biological processes such as organelle function, including many mitochondrial genes that were largely absent in the snRNA-seq datasets (see Supplementary Table 5).

To further illustrate the power of the different protocols to chart the heterogeneity of complex samples, we clustered and plotted down-sampled datasets in two-dimensional space (Fig. 4a) and then calculated the cluster accuracy and average silhouette width (ASW<sup>25</sup>, Fig. 4b), a commonly used measure for assessing the quality of data partitioning into communities. Consistent with the assumption that library complexity and sensitive marker detection provide greater power to describe complexity, methods that performed well for these two attributes showed better separation of subpopulations, and greater ASW and cluster accuracy. This is illustrated in the monocytes, for which accurate clustering protocols separated the major subpopulations (CD14<sup>+</sup> and FCGR3A<sup>+</sup>), whereas methods with low ASW did not distinguish between them. Similarly, several methods were able to distinguish between CD8<sup>+</sup> and natural killer (NK) cells, whereas others were not.

**Joint analysis across datasets.** A common scenario for cell atlas projects is that data are produced at different sites using different scRNA-seq protocols. However, the final atlas is created from a combination of datasets, which requires that the technologies used be compatible. To assess how suitable it is to combine the results from our protocols into a joint analysis, we used down-sampled human and mouse datasets to produce a joint quantification matrix for all techniques<sup>25</sup>. Importantly, single cells grouped themselves by cell type, suggesting that cell phenotypes are the main driver of heterogeneity in the joint datasets (Fig. 5a–d, and see Supplementary Figs. 19a,b and 20). Indeed, the combined data showed a clear separation of cell states (for example, T cell and enterocyte subpopulations) and rarer cell types, such as dendritic cells. However, within these populations, differences between the protocols pointed to the



**Fig. 5 | Integration of sc/snRNA-seq methods. a-d**, UMAP visualization of cells after integrating technologies for 18,034 human (a,b) and 7,902 mouse (c,d) cells. Cells are colored by cell type (a,c) and sc/snRNA-seq protocol (b,d). e,f, Barplots showing normalized and method-corrected (integrated) expression scores of cell-type-specific signatures for human HEK293T cells, monocytes, B cells (e), and mouse secretory and TA cells (f). Bars represent cells and colors methods. g,h, Evaluation of method integratability in human (g) and mouse (h) cells. Protocols are compared according to their ability to group cell types into clusters (after integration) and mix with other technologies within the same clusters. Points are colored by sequencing method.

presence of technical effects that could not be entirely removed with down-sampling to equal read depth and different merging tools (Fig. 5e,f, and see Supplementary Figs. 19c,d, 21a,b and 22a,b). To formally assess the capacity of the methods to be combined, we calculated the degree to which technologies mix in the merged datasets (Fig. 5g,h, and see Supplementary Figs. 21c,d and 22c,d). The suitability of protocols to be combined (mixability) was directly correlated with their power to discriminate between cell types (clustering accuracy). Thus, well-performing protocols result in high-resolution cellular maps and are suitable for consortium-driven projects that include different data sources. When integrating further down-sampled datasets, we observed a drop in mixing ability (see Supplementary Fig. 19e). Consequently, quality standard guidelines for consortia might define minimum coverage thresholds to ensure the subsequent option of data integration. A separate analysis of the single-nucleus and single-cell Chromium datasets resulted in well-integrated profiles, further supporting the potential to integrate cell atlases from cells and nuclei (see Supplementary Figs. 23 and 24).

Cell atlas datasets will serve as a reference for annotating cell types and states in future experiments. Therefore, we assessed cells' ability to be projected on to our reference sample (Fig. 2b,c). We used the population signature model defined by matchSCore2 and evaluated the protocols based on their cell-by-cell mapping probability, which reflects the confidence of cell annotation (see Supplementary Fig. 25a-c). Although there were some differences

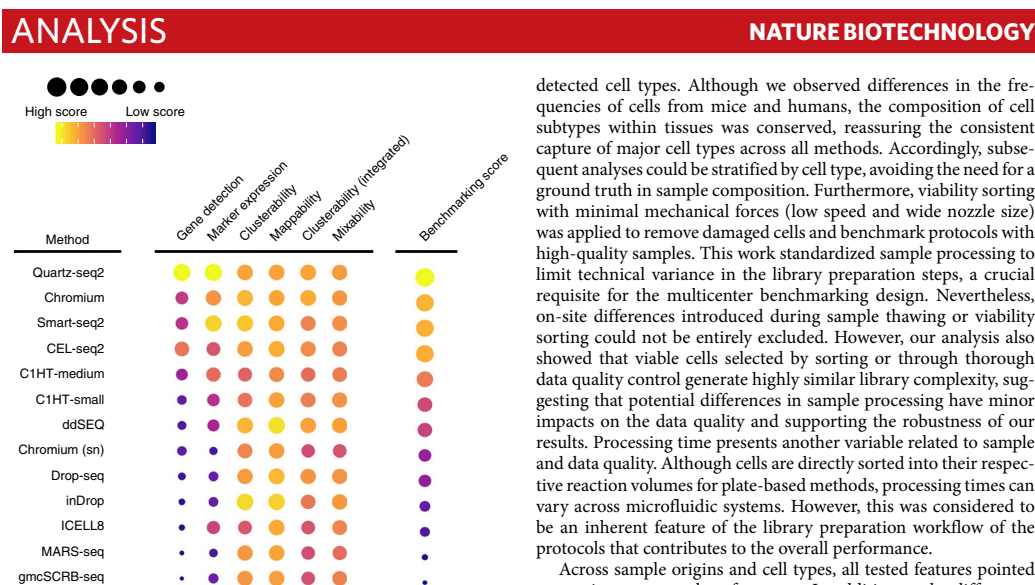
in the projection probabilities of the protocols, and a potential bias due to the selection of the reference protocol, a confident annotation was observed for most cells with inDrop and ddSEQ reporting the highest probabilities. Notably, high probability scores were also observed in further down-sampled datasets (see Supplementary Fig. 25b). This has practical consequences, because data derived from less well-performing methods (from a cell atlas perspective), or from poorly sequenced experiments, could be identifiable and thus suitable for specific analysis types, such as tissue composition profiling.

### Discussion

Systematic benchmarking of available technologies is a crucial prerequisite for large-scale projects. In the present study, we evaluated scRNA-seq protocols for their power to produce a cellular map of complex tissues. Our reference sample simulated common scenarios in cell atlas projects, including differentiated cell types and dynamic cell states. We defined the strengths and weaknesses of key features that are relevant for cell atlas studies, such as comprehensiveness, integratability and predictive value. The methods revealed a broad spectrum of performance, which should be considered when defining guidelines and standards for international consortia (Fig. 6).

We expect that our results will guide informed decision-making processes for designing sc/snRNA-seq studies. There are several features to consider when selecting protocols to produce a





**Fig. 6 | Benchmarking summary of 13 sc/snRNA-seq methods.** Methods are scored by key analytical metrics, characterizing protocols according to their ability to recapitulate the original structure of complex tissues, and their suitability for cell atlas projects. The methods are ordered by their overall benchmarking score, which is computed by averaging the scores across metrics assessed from the human datasets.

reproducible, integrative and predictive reference cell atlas. At a given sequencing depth, the number and complexity of detected RNA molecules define the power to describe cell phenotypes and infer their function. There are also additional essential features for cell atlas projects and their interpretation, such as population marker identification. Improved versions of plate-based methods, including Quartz-seq2, CEL-seq2 and Smart-seq2, generate such high-resolution transcriptome profiles. Also, microfluidic systems showed excellent performance in our comparison, particularly the Chromium system. Although the scale of plate-based experiments is limited by the lower throughput of their individual processing units, microfluidic systems, especially droplet-based methods, can be easily applied to thousands of cells simultaneously. Protocol modification scales up throughput even further, and allows more cost-effective experiments<sup>26–29</sup>. Generally, late multiplexing methods, such as Smart-seq2, are more costly, but costs can be reduced by miniaturization<sup>30</sup> and use of noncommercial enzymes<sup>31</sup>. Custom droplet-based protocols have lower costs than their commercialized counterparts, but the optimized chemistry in commercial systems resulted in improved performance in this comparison. Nevertheless, existing platforms are undergoing continued development in both the private (see Supplementary Fig. 12) and the academic sectors, so updated protocol versions promise to improve performance further. For consortium-driven projects, it is important to consider the integrability of data. We have shown that several protocols, including those with reduced library complexity and snRNA-seq, were readily integratable with other methods.

The use of PBMCs is ideal for multicenter benchmarking efforts; blood cells are easy to isolate and show a high recovery rate after freezing. We also included mouse colon, a solid tissue requiring dissociation before scRNA-seq. Tissue digestion and cryopreservation of colon cells present additional challenges (for example, increased rate of damaged cells), which we addressed by focusing on commonly

detected cell types. Although we observed differences in the frequencies of cells from mice and humans, the composition of cell subtypes within tissues was conserved, reassuring the consistent capture of major cell types across all methods. Accordingly, subsequent analyses could be stratified by cell type, avoiding the need for a ground truth in sample composition. Furthermore, viability sorting with minimal mechanical forces (low speed and wide nozzle size) was applied to remove damaged cells and benchmark protocols with high-quality samples. This work standardized sample processing to limit technical variance in the library preparation steps, a crucial requisite for the multicenter benchmarking design. Nevertheless, on-site differences introduced during sample thawing or viability sorting could not be entirely excluded. However, our analysis also showed that viable cells selected by sorting or through thorough data quality control generate highly similar library complexity, suggesting that potential differences in sample processing have minor impacts on the data quality and supporting the robustness of our results. Processing time presents another variable related to sample and data quality. Although cells are directly sorted into their respective reaction volumes for plate-based methods, processing times can vary across microfluidic systems. However, this was considered to be an inherent feature of the library preparation workflow of the protocols that contributes to the overall performance.

Across sample origins and cell types, all tested features pointed to consistent protocol performance. In addition to the differences in protocol performance, it was the cells' RNA content and complexity that dominated the molecule and gene detection rates, which we have seen through the stratified analysis of vastly different cell types. As such, we expect the conclusions to be valid beyond the human and mouse tissues tested in the present study.

Several additional steps are crucial for the success of single-cell projects, especially sample preparation. Optimization of sample procurement and tissue-processing conditions is of crucial importance to avoid composition biases and gene expression artifacts<sup>32–35</sup> that could limit the value of a cell atlas. Therefore, dedicated studies are required to define optimal conditions for tissue and organ preparation in healthy and disease contexts.

From a technical perspective, multiple steps of a protocol are critical for generating complex sequencing libraries. All sc/snRNA-seq methods require multi-step, whole-transcriptome amplification, including reverse transcription, conversion to amplifiable cDNA and amplification<sup>1</sup>. Theoretically, the multiplicative reaction efficiency of respective steps determines a method's power to detect RNA molecules, and in this sense Quartz-Seq2 was particularly efficient. We specifically tested for potential advantages of the Quartz-seq2 column-based over bead-based purification, but did not detect differences in cDNA yield (see Supplementary Fig. 26). However, we observed that bead concentration critically affected the yield of amplified cDNA. Moreover, performance was more stable for purification with columns compared with beads, which should be taken into account when implementing existing or developing new sc/snRNA-seq methods.

A further essential step toward complex libraries is the conversion of first-strand cDNA to amplifiable cDNA. Three main strategies are used for this conversion: (1) template switching, (2) RNaseH/DNA polymerase I-mediated, second-strand synthesis for in vitro transcription and (3) poly(A) tagging<sup>1</sup>. Improvement of the three strategies led to better quantitative performance of scRNA-seq<sup>36–39</sup>. For Quartz-Seq2 (ref. <sup>37</sup>), improved poly(A) tagging was most important to increase the amplified cDNA yield compared with Quartz-Seq<sup>40</sup>, and probably explains the excellent result in this benchmarking exercise. However, optimization of the cDNA conversion still has the potential to improve scRNA-seq methods.

Within the cDNA amplification step, increased PCR cycle numbers lead to PCR biases within the sequencing libraries. Early pooling increases the number of cDNA molecules in the amplification

step and reduces PCR bias. This especially favors early pooling methods at low sequencing depth (as performed in the present study), as previously shown for bulk RNA-seq<sup>41</sup>. Similarly, in vitro transcription linearly amplifies cDNA with fewer biases than PCR-based methods, and partly explains the good performance of CEL-seq2. Furthermore, early multiplexing of different cell numbers leads to different PCR cycle requirements (Quartz-Seq2 with 768 cells and 10 cycles versus gmcSCR-seq with 96 cells and 19 cycles, using the same DNA polymerase for amplification). The number of cells per amplification pool depends on the amount of amplifiable cDNA, implying that the good performance of Quartz-Seq2 was mainly due to efficient conversion of amplifiable cDNA from RNA with poly(A) tagging.

It is equally important to benchmark computational pipelines for data analysis and interpretation<sup>23,42–44</sup>. We envision the datasets provided by our study serving as a valuable resource for the single-cell community to develop and evaluate new strategies for an informative and interpretable cell atlas. Moreover, the multicenter benchmarking framework presented in the present study can readily be transferred to other organs where common tissue/cell types are analyzed using different scRNA-seq protocols (for example, brain atlas projects).

#### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-020-0469-4>.

Received: 7 May 2019; Accepted: 26 February 2020;

Published online: 6 April 2020

#### References

- Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* **13**, 2742–2757 (2018).
- Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.* **18**, 345–361 (2017).
- Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
- Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
- Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
- Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).
- Karaïskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
- Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
- Regev, A. et al. Science forum: the human cell atlas. *eLife* **6**, e27041 (2017).
- Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
- Plass, M. et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaq1723 (2018).
- Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl Acad. Sci. USA* **113**, 11046–11051 (2016).
- Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
- Alioto, T. S. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).
- Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 (2017).
- Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
- Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
- Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
- Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
- Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
- Guillaumet-Adkins, A. et al. Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol.* **18**, 45 (2017).
- Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* **562**, 367–372 (2018).
- Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
- Azuaje, F. A cluster validity framework for genome expression data. *Bioinforma* **18**, 319–320 (2002).
- Lin, Y. et al. scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl Acad. Sci. USA* **116**, 9775–9784 (2019).
- Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
- Stoeckius, M. et al. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
- McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
- Gaublomme, J. T. et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat. Commun.* **10**, 1–8 (2019).
- Mora-Castilla, S. et al. Miniaturization technologies for efficient single-cell library preparation for next-generation sequencing. *J. Lab. Autom.* **21**, 557–567 (2016).
- Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
- Brink, S. C. Vanden et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
- Wohnhaas, C. T. et al. DMSO cryopreservation is the method of choice to preserve cells for droplet-based single-cell RNA sequencing. *Sci. Rep.* **9**, 1–14 (2019).
- Tosti, L. et al. Single nucleus RNA sequencing maps acinar cell states in a human pancreas cell atlas. Preprint at *bioRxiv* <https://doi.org/10.1101/733964> (2019).
- Massoni-Badosa, R. et al. Sampling artifacts in single-cell genomics cohort studies. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.01.15.897066> (2020).
- Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
- 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).
- Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
- Bagnoli, J. W. et al. Sensitive and powerful single-cell RNA sequencing using mcSCR-seq. *Nat. Commun.* **9**, 2937 (2018).
- Sasagawa, Y. et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method. reveals non-genetic gene-expression heterogeneity. *Genome Biol.* **14**, 3097 (2013).
- 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).
- Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
- Saelens, W. et al. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
- Holland, C. H. et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* **21**, 36 (2020).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020

## ANALYSIS

## NATURE BIOTECHNOLOGY

## Methods

**Ethical statement.** The present study was approved by the Parc de Salut MAR Research Ethics Committee (reference no. 2017/7585/I) to H.H. We adhered to ethical and legal protection guidelines for human participants, including informed consent.

**Reference sample. Cell lines.** NIH3T3-GFP, MDCK-TurboFP650 and HEK293-RFP cells were cultured at 37 °C in an atmosphere of 5% (v/v) carbon dioxide in Dulbecco's modified Eagle's medium, supplemented with 10% (w/v) fetal bovine serum (FBS), 100 U penicillin, and 100 µg l<sup>-1</sup> of streptomycin (Invitrogen). On the reference sample preparation day, the culture medium was removed and the cells were washed with 1× phosphate-buffered saline (PBS). Afterwards, cells were trypsinized (trypsin 100×), pelleted at 800g for 5 min, washed in 1× PBS, resuspended in PBS + ethylenediaminetetraacetic acid (EDTA) (2 mM) and stored on ice.

**Mouse colon tissue.** The colons from 11 mice (7 *LGR5/GFP* and 4 wild-type) were dissected and removed. For single-cell separation the colons were treated separately. The colon was sliced, opened and washed twice in cold 1× Hank's balanced salt solution (HBSS). It was then placed on a Petri dish on ice and minced with razor blades until disintegration. The minced tissue was transferred to a 15-ml tube containing 5 ml of 1× HBSS and 83 µl of collagenase IV (final concentration 166 U ml<sup>-1</sup>). The solution was incubated for 15 min at 37 °C (vortexed for 10 s every 5 min). To inactivate the collagenase IV, 1 ml of FBS was added and it was vortexed for 10 s. The solution was filtered through a 70-µm nylon mesh (changed when clogged). Finally, all samples were combined, and the cells pelleted for 5 min at 400g and 4 °C. The supernatant was removed and the cells resuspended in 20 ml of 1× HBSS and stored on ice.

**Isolation of PBMCs.** Whole blood was obtained from four donors (two female, two male). The extracted blood was collected in heparin tubes (GP Supplies) and processed immediately. For each donor, PBMCs were isolated according to the manufacturer's instructions for Ficoll extraction (pluriSelect). Briefly, blood from two heparin tubes (approximately 8 ml) was combined, diluted in 1× PBS and carefully added to a 50-ml tube containing 15 ml of Ficoll. The tubes were centrifuged for 30 min at 500g (minimum acceleration and deceleration). The interphase was carefully collected and diluted with 1× PBS + 2 mM EDTA. After a second centrifugation, the supernatant was discarded and the pellet resuspended in 2 ml of 1× PBS + 2 mM EDTA and stored on ice.

**Preparation of the reference sample.** Cell counting was performed using an automated cell counter (TC20 Automated Cell Counter, Bio-Rad Laboratories). The reference sample was calculated to include human PBMCs (60%), mouse colon cells (30%), and HEK293T (6%, RFP-labeled human cell line), NIH3T3 (3%, GFP-labeled mouse cells) and MDCK (1%, TurboFP650-labeled dog cells) cells. To adjust for cell integrity loss during sample processing, we measured the viability during cell counting and accounted for an expected viability loss after cryopreservation (10% for cell lines and PBMCs; 50% for colon cells<sup>31</sup>). All single-cell solutions were combined in the proportions mentioned above and diluted to 250,000 viable cells per 0.5 ml. For cryopreservation, 0.5 ml of cell suspension was aliquoted into cryotubes and gently mixed with a freezing solution (final concentration 10% dimethylsulfoxide; 10% heat-inactivated FBS). Cells were then frozen by gradually decreasing the temperature (1 °C min<sup>-1</sup>) to -80 °C (cryopreserved), and stored in liquid nitrogen. MARS-Seq and Smart-Seq2 experiments were performed to validate sample quality and composition before distributing aliquots to the partners.

**Sample processing.** Samples were stored at -80 °C on arrival. Before processing, samples were de-frozen in a water bath (37 °C) with continuous agitation until the material was almost thawed. The entire volume was transferred to a 15-ml Falcon tube using a 1,000-µl tip (wide-bored or cut tip) without mixing by pipetting; 1,000 µl of prewarmed (37 °C) Hibernate-A was added drop-wise while gently swirling the sample. The sample was then rested for 1 min. An additional 2,000 µl of prewarmed (37 °C) Hibernate-A was added drop-wise while gently swirling the sample. The sample was again rested for 1 min. Another 2,000 µl of prewarmed (37 °C) Hibernate-A was added drop-wise while gently swirling the sample and the sample was rested for 1 min. Then, 3,000 µl of prewarmed (37 °C) Hibernate-A was added drop-wise and the Falcon tube inverted six times. The sample was rested for 1 min. An additional 5,000 µl of prewarmed (37 °C) Hibernate-A was added drop-wise and the Falcon tube inverted six times. The sample was rested for 1 min. It was then centrifuged at 400g for 5 min at 4 °C (pellet clearly visible). The supernatant was removed until 500 µl remained in the tube. The pellet was resuspended by gentle pipetting. Then 3,500 µl of 1× PBS + 2 mM EDTA was added and the sample stored on ice until processing. Before FACS isolation, cells were filtered through a nylon mesh and 3 µl DAPI was added before gentle mixing. During FACS isolation, DAPI-positive cells were excluded to remove dead and damaged cells. Furthermore, the exclusion of GFP-positive cells simulated the removal of a cell type from a complex sample. Supplementary Fig. 27 shows representative FACS plots and gating strategies.

**ScRNA-seq library preparation.** For a detailed sample processing description, see Supplementary Notes.

**Data analysis.** For primary data preprocessing, clustering, sample deconvolution and annotation, and reference datasets, see Supplementary Notes.

**MatchScore2.** To systematically assign cell identities to unannotated cells coming from different protocols, we used matchScore2, a mathematical framework for classifying cell types based on reference data (<https://github.com/elimereu/matchScore2>). The reference data consist of a matrix of gene expression counts in individual cells, the identity of which is known. The main steps of the matchScore2 annotation are the following:

- (1) Normalization of the reference data. Gene expression counts are log(normalized) for each cell using the natural logarithm of 1 + counts per 10,000. Genes are then scaled and centered using the ScaleData function in the Seurat package.
- (2) Definition of signatures and their relative scores. For each of the cell types in the reference data, positive markers were computed using Wilcoxon's rank-sum test. The top 100 ranked markers in each cell type were used as the signature for that type. To each cell, we assigned a vector  $\mathbf{x} = (x_1, \dots, x_n)$  of signature scores, where  $n$  is the number of cell types in the reference data. The  $i$ th signature score for the  $k$ th cell is computed as follows:

$$\text{Score}_k = \sum_{j \in J} z_{jk}$$

where  $J$  is the set of genes in signature  $i$ , and  $z_{jk}$  represents the  $z$ -score of gene  $j$  in the  $k$ th cell.

- (3) Training of the probabilistic model on the reference data.

We proposed a supervised multinomial logistic regression model, which uses enrichment of the signature of each reference cell type in each cell to assign identity to that cell. In other words, for each cell  $k$  and signature  $i$ , we calculate the  $i$ th cell-type signature score  $x_i$  in the  $k$ th cell as described in point 2. The distribution of the signature scores is preserved, independent of which protocol is used (see Supplementary Figs. 28 and 29). More specifically, we defined the variables  $x_1, \dots, x_n$ , where  $\mathbf{x}_i$  is the vector in which the scores for signature  $i$  of all cells are contained. Then we used  $\mathbf{x}_i$  as the predictor of a multinomial logistic regression.

The model assumes that the number of cells from each type in the training reference data  $T_1, T_2, \dots, T_n$  are random variables and that the variable  $T = (T_1, T_2, \dots, T_n)$  follows a multinomial distribution  $M(N, \boldsymbol{\pi} = (\pi_1, \dots, \pi_n))$ , where  $\pi_i$  is the proportion of the  $i$ th cell type and  $N$  is the total number of cells.

To test the performance of the model, training and test sets were created by subsampling the reference into two datasets, maintaining the original proportions of cell types in both sets. The model was trained by using the multinomial function from the `nnet` R package (`decay = 1 \times 10^{-4}`, `maxit = 500`). To improve the convergence of the model function,  $\mathbf{x}_i$  variables were scaled to the interval [0,1].

**Cell classification.** For each cell, model predictions consisted of a set of probability values per identity class, and the highest probability was used to annotate the cell if it was >0.5; otherwise the cell remained unclassified.

**Model accuracy.** To evaluate the fitted model using our reference datasets, we assessed the prediction accuracy in the test set, which was around 0.9 for human and 0.85 for mouse reference. We further assessed matchScore2 classifications in datasets from other sequencing methods by looking at the agreement between clusters and classification. Notably, the resulting average agreement was 80% (range: from 58% in `gmcSCRb-seq` to 92% in `Quartz-Seq2`), whereas the rate for unclassified cells was <2%.

**Down-sampling.** To decide on a common down-sampling threshold for sequencing depth per cell, we inspected the distribution of the total number of reads per cell for each technique, and chose the lowest first quartile (fixed to 20,000 reads per cell). We then performed stepwise down-sampling (25%, 50% and 75%) using the `zUMIs` down-sampling function. We omitted cells that did not achieve the required minimum depth (see Supplementary Table 6). Notably, stochasticity introduced during down-sampling did not affect the results of the present study, as exemplified by the consistent numbers of detected molecules across different down-sampling iterations (see Supplementary Fig. 10).

**Estimation of dropout probabilities.** We investigated the impact of dropout events in HEK293T cells, monocytes and B cells extracted for each technique on down-sampled data (20,000 reads per cell). For datasets with >50 cells from the selected populations, we randomly sampled 50 cells to eliminate the effect of differing cell number. The dropout probability was computed using the `SCDE` R package<sup>32</sup>. `SCDE` models the measurements of each cell as a mixture of a negative binomial process to account for the correlation between amplification and detection of a transcript and its abundance, and a Poisson process to account for the background signal. We then used estimated individual error models for each cell as a function of expression magnitude to compute dropout probabilities using



## NATURE BIOTECHNOLOGY

## ANALYSIS

SCDE's  $s_{cde}$  failure probability function. Next, we calculated the average estimated dropout probability for each cell type and technique. To integrate dropout measures into the final benchmarking score, we calculated the area under the curve of the expression prior and failure probabilities (see Fig. 2f and also Supplementary Table 7). We expected that protocols resulting in fewer dropouts would have smaller areas under the curve.

**Quantification of variance introduced by batches.** To quantify the amount of variance that is introduced by batches (protocols, processing units or experiments), we used the top 20 PCs and the s.d. of each PC, previously calculated on HVGs. Next, using the  $pcRegression$  function of kBET R package<sup>23</sup>, we regressed the batch covariate (protocols/processing units/experiments as categories defined in the kBET model) and each PC to obtain the coefficient of determination as an approximation of the variance explained by batches, and the proportions of explained variance in each PC. We either reported the percentage of the variance that correlates significantly with the batch in the first 20 PCs, or R-squared measures of the model for each PC.

**Cumulative number of genes.** The cumulative number of detected genes in the down-sampled data was calculated separately for each cell type. For cell types with >50 cells annotated, we randomly selected 50 cells and calculated the average number of detected genes per cell after 50 permutations over  $n$  sampled cells, where  $n$  is an increasing sequence of integers from 1 to 50.

**GO enrichment analysis.** To compare functional gene sets between single-cell and single-nucleus datasets, we performed Gene Ontology (GO) enrichment analysis on the set of protocol-specific genes using simpleGO (<https://github.com/iaconogi/simpleGO>). For each cell type (HEK293T cells, monocytes and B cells), we selected two gene sets extracted from the cumulated genes and using the maximum number of detected cells common to all three Chromium versions: (1) genes that were uniquely detected in the intersection of Chromium (v.2) and (v.3), but not in Chromium (sn), and (2) genes that were uniquely identified with Chromium (sn). For each of the gene sets, we identified the union over cell types before applying simpleGO.

**Correlation analysis.** Pearson's correlations across protocols were computed independently for B cells, monocytes and HEK293T cells. For each cell type, cells were down-sampled to the maximum common number of cells across all protocols. Gene counts of commonly expressed genes (from datasets down-sampled to 20,000 reads) were averaged across cells before computing their Pearson's correlations. The corplot library was then used to plot the resulting correlations. Protocols were ordered by agglomerative hierarchical clustering.

**Silhouette scores.** To measure the strength of the clusters, we calculated the ASW<sup>24</sup>. The down-sampled data (20,000 reads per cell) were clustered by Seurat<sup>46</sup>, using graph-based clustering with the first eight PCs and a resolution of 0.6. We then computed an ASW for the clusters using a Euclidean distance matrix (based on PCs 1–8). We reported the ASW for each technique separately.

**Dataset merging.** Dataset integration across protocols is challenging and we applied different tools to assess the integrability of the sc/snRNA-seq methods, while conserving biological variability. To integrate datasets, we used Seurat<sup>46</sup>, harmony<sup>27</sup> and scMerge<sup>25</sup>, evaluated the results separately and averaged the integration capacity of the protocols into a joint score. We combined down-sampled count matrices using the  $sce\_cbind$  function in scMerge, which includes the union of genes from different batches. Although both harmony and Seurat integration apply similar preprocessing steps (log(normalization), scaling and HVG identification), as implemented in the Seurat tool, scMerge uses a set of genes with stable expression levels across different cell types, and then creates pseudo-replicates across datasets, allowing the estimation and correction for undesired sources of variability. However, for all three alignment methods, Seurat was applied to perform clustering and Uniform Manifold Approximation and Projection (UMAP) after the protocol correction, to minimize the variability related to the downstream analysis. The clustering accuracy metric was used together with the mixability score to quantify the success of the integration. Omitting the cell integration step before visualizing the datasets together in a single tSNE/UMAP resulted in a protocol-specific distribution with cell types scattered to multiple clusters (see Supplementary Fig. 30).

**Clustering accuracy.** To determine the clusterability of methods to identify cell types, we measured the probability of cells being clustered with cells of the same type. Let  $C_k$ ,  $k \in \{1, \dots, N\}$  represent the cluster of cells corresponding to a unique cell type (based on the highest agreement between clusters and cell types), and  $T_j$ ,  $j \in \{1, \dots, S\}$  represent the set of different cell types, where  $C_k \subseteq T_j$ . For each cell type  $T_j$ , we compute the proportion  $p_{jk}$  of  $T_j$  cells that cluster in their correct cluster  $C_k$ . We define the cell-type separation accuracy as the average of these proportions.

**Mixability.** To account for the level of mixing of each technology, we used kBET<sup>23</sup> to quantify batch effects by measuring the rejection rate of Pearson's  $\chi^2$  test for random neighborhoods. To make a fair comparison, kBET was applied to the

common cell types separately by subsampling batches to the minimum number of cells in each cell type. Due to the reduced number of cells, the option heuristic was set to 'False', and the testSize was increased to ensure a minimum number of cells. Mixability was calculated by averaging cell-type-specific rejection rates.

**Benchmarking score.** To create an overall benchmarking score against which to compare technologies, we considered six key metrics: gene detection, overall level of expression in transcriptional signatures, cluster accuracy, classification probability, cluster accuracy after integration and mixability. Each metric was scaled to the interval [0,1], then, to equalize the weight of each metric score, the harmonic mean across these metrics was calculated to obtain the final benchmarking scores. Gene detection, overall expression in cell-type signatures and classification probabilities were computed separately for B cells, HEK293T cells and monocytes, and then aggregated by the arithmetic mean across cell types. Notably, the choice of protocol to create the reference dataset (Chromium) for initial cell annotation had no impact on the outcome of the present study (see Supplementary Fig. 31).

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

### Data availability

All raw sequencing data and processed gene expression files are freely available through the Gene Expression Omnibus (accession no. GSE133549).

### Code availability

All code for the analysis is provided as supplementary material. All code is also available under [https://github.com/ati-lz/HCA\\_Benchmarking](https://github.com/ati-lz/HCA_Benchmarking) and <https://github.com/elimerreu/matchScore2>.

### References

- Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
- Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
- Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

### Acknowledgements

This project has been made possible in part by grant no. 2018-182827 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. H.H. is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III (ISCIII). C.M. is supported by an AECC postdoctoral fellowship. This work has received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement no. H2020-MSCA-ITN-2015-675752 (Singek), and the Ministerio de Ciencia, Innovación y Universidades (SAF2017-89109-P; AEI/FEDER, UE). S. was supported by the German Research Foundation's (DFG's) (GR4980) Behrens-Weise-Foundation. D.G. and S. are supported by the Max Planck Society. C.Z. was supported by the European Molecular Biology Organization through the long-term fellowship ALTF 673-2017. The snRNA-seq data were generated with support from the National Institute of Allergy and Infectious Diseases (grant no. U24AI118672), the Mantion Foundation and the Klarman Cell Observatory (to A.R.). I.N. was supported by JST CREST (grant no. JPMJCR16G3), Japan, and the Projects for Technological Development, Research Center Network for Realization of Regenerative Medicine by Japan, the Japan Agency for Medical Research and Development. A.J., L.E.W., J.W.B. and W.E. were supported by funding from the DFG (EN 1093/2-1 and SFB1243 TP A14). We thank ThePaperMill for critical reading and scientific editing services and the Eukaryotic Single Cell Genomics Facility at SciLifeLab (Stockholm, Sweden) for support. This publication is part of a project (BCLLATLAS) that received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 810287). Core funding was from the ISCIII and the Generalitat de Catalunya.

### Author contributions

H.H. designed the study. E.M. and A.L. performed all data analyses. C.M., A.A.V. and E.B. prepared the reference sample. C.Z., D.J.M., S.P. and O.S. supported the data analysis. M.G. and I.G. provided technical and sequencing support. S., D.G., J.K.L., S.C.B., C.S., A.O., R.C.J., K.K., C.B., Y.T., Y.S., K.T., T.H., C.B., C.F., S.S., T.T., C.C., X.A., L.T.N., A.R., J.Z.L., A.J., L.E.W., J.W.B., W.E., R.S. and I.N. provided sequencing-ready single-cell libraries or sequencing raw data. H.H., E.M. and A.L. wrote the manuscript with contributions from the co-authors. All authors read and approved the final manuscript.

### Competing interests

A.R. is a co-founder and equity holder of Celsius Therapeutics, and an SAB member of Thermo Fisher Scientific and Syros Pharmaceuticals. He is also a co-inventor on patent applications to numerous advances in single-cell genomics, including droplet-based

## ANALYSIS

## NATURE BIOTECHNOLOGY

sequencing technologies, as in PCT/US2015/0949178, and methods for expression and analysis, as in PCT/US2016/059233 and PCT/US2016/059239. K.K., C.B. and Y.T. are employed by Bio-Rad Laboratories. J.K.L. and S.C.B. are employees and shareholders at 10x Genomics, Inc. S.C.B. is a former employee and shareholder of Fluidigm Corporation. C.S. and A.O. are employed by Fluidigm. All other authors declare no conflicts of interest associated with this manuscript.

**Additional information**

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41587-020-0469-4>.

**Correspondence and requests for materials** should be addressed to H.H.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

## Supplementary Information

In the format provided by the authors and unedited.

## Benchmarking single-cell RNA-sequencing protocols for cell atlas projects

Elisabetta Mereu<sup>1,26</sup>, Atefeh Lafzi<sup>1,26</sup>, Catia Moutinho<sup>1</sup>, Christoph Ziegenhain<sup>2</sup>, Davis J. McCarthy<sup>3,4,5</sup>, Adrián Álvarez-Varela<sup>6</sup>, Eduard Batlle<sup>6,7,8</sup>, Sagar<sup>9</sup>, Dominic Grün<sup>9</sup>, Julia K. Lau<sup>10</sup>, Stéphane C. Boutet<sup>10</sup>, Chad Sanada<sup>11</sup>, Aik Ooi<sup>11</sup>, Robert C. Jones<sup>12</sup>, Kelly Kaihara<sup>13</sup>, Chris Brampton<sup>13</sup>, Yasha Talaga<sup>13</sup>, Yohei Sasagawa<sup>14</sup>, Kaori Tanaka<sup>14</sup>, Tetsutaro Hayashi<sup>14</sup>, Caroline Braeuning<sup>15</sup>, Cornelius Fischer<sup>15</sup>, Sascha Sauer<sup>15</sup>, Timo Trefzer<sup>16</sup>, Christian Conrad<sup>16</sup>, Xian Adiconis<sup>17,18</sup>, Lan T. Nguyen<sup>17</sup>, Aviv Regev<sup>17,19,20</sup>, Joshua Z. Levin<sup>17,18</sup>, Swati Parekh<sup>21</sup>, Aleksandar Janjic<sup>22</sup>, Lucas E. Wang<sup>22</sup>, Johannes W. Bagnoli<sup>22</sup>, Wolfgang Enard<sup>22</sup>, Marta Gut<sup>1</sup>, Rickard Sandberg<sup>2</sup>, Itoshi Nikaido<sup>14,23</sup>, Ivo Gut<sup>1,24</sup>, Oliver Stegle<sup>3,4,25</sup> and Holger Heyn<sup>1,24</sup> ✉

<sup>1</sup>CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>2</sup>Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. <sup>3</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK. <sup>4</sup>European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. <sup>5</sup>St Vincent's Institute of Medical Research, Fitzroy, Victoria, Australia. <sup>6</sup>Institute for Research in Biomedicine, Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>7</sup>Catalan Institution for Research and Advanced Studies, Barcelona, Spain. <sup>8</sup>Centro de Investigación Biomédica en Red de Cáncer, Barcelona, Spain. <sup>9</sup>Max-Planck-Institute of Immunobiology and Epigenetics, Freiburg, Germany. <sup>10</sup>10x Genomics, Pleasanton, CA, USA. <sup>11</sup>Fluidigm Corporation, South San Francisco, CA, USA. <sup>12</sup>Department of Bioengineering, Stanford University, Stanford, CA, USA. <sup>13</sup>Bio-Rad, Hercules, CA, USA. <sup>14</sup>Laboratory for Bioinformatics Research, RIKEN Center for Biosystems, Dynamics Research, Saitama, Japan. <sup>15</sup>Max Delbrück Center for Molecular Medicine/Berlin Institute of Health, Berlin, Germany. <sup>16</sup>Digital Health Center, Berlin Institute of Health, Charité-Universitätsmedizin Berlin, Berlin, Germany. <sup>17</sup>Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>18</sup>Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>19</sup>Koch Institute of Integrative Cancer Research, MIT, Cambridge, MA, USA. <sup>20</sup>Howard Hughes Medical Institute, Department of Biology, MIT, Cambridge, MA, USA. <sup>21</sup>Max-Planck-Institute for Biology of Ageing, Cologne, Germany. <sup>22</sup>Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Martinsried, Germany. <sup>23</sup>School of Integrative and Global Majors, University of Tsukuba, Wako, Saitama, Japan. <sup>24</sup>Universitat Pompeu Fabra, Barcelona, Spain. <sup>25</sup>Division of Computational Genomics and Systems Genetics, German Cancer Research Center, Heidelberg, Germany. <sup>26</sup>These authors contributed equally: Elisabetta Mereu, Atefeh Lafzi. ✉e-mail: [holger.heyn@cnag.crg.eu](mailto:holger.heyn@cnag.crg.eu)

**Supplementary Material**

**Supplementary Notes**

**Supplementary Figure legends 1-31.**

**Supplementary Figures 1-31.**

**Supplementary Tables 1-8.**

## Supplementary Notes

### Single-cell RNA sequencing library preparation

#### Quartz-Seq2<sup>1</sup>

We isolated single-cells into 1  $\mu$ L of lysis buffer (0.1111  $\mu$ M respective RT primers, 0.12 mM dNTP mix, 0.3% NP-40, 1 unit/ $\mu$ L RNasin plus) in each well of 384-well PCR plates from cell suspension using a MoFlo Astrios EQ (Beckman Coulter) cell sorter. The event-rate in flow cytometry was approximately 200 events per second. The cell sorter was equipped with a 100- $\mu$ m nozzle and a custom-made splash-guard (**Supplementary Fig. 32**). In total, we analyzed 3,072 wells corresponding to eight 384-well PCR plates. Sequence library preparation of Quartz-Seq2 was performed as described previously<sup>1</sup> with the following modifications. For lysis buffer, we used 768 kinds of RT primers corresponding to v3.2A and v3.2B (**Supplementary Table 8**). We prepared two sets of the 384-well PCR plate with lysis buffer containing no ERCC spike-in RNA. We added 1  $\mu$ l of RT premix (2X Thermopol buffer, 1.25 units/ $\mu$ L SuperScript III, 0.1375 units/ $\mu$ L RNasin plus) to 1  $\mu$ l of lysis buffer for each well. After cell barcoding, we collected cDNA solution into one well reservoir from two sets of 384-well plates, which corresponded to 768 wells. For cDNA purification and concentration, we used four Zymo-Spin-I Columns (Zymo Research) for cDNA solution from two 384-well PCR plates. In the PCR step, we amplified the cDNA for 10 cycles under the following conditions: 98 °C for 10 s, 65 °C for 15 s, and 68 °C 5 min. In an additional purification step for amplified cDNA, we added 26  $\mu$ l (0.65X) of resuspended AMPure XP Beads to the cDNA solution. We obtained amplified cDNA of  $32.6 \pm 6.8$  ng ( $n = 4$ ) from the 768 wells. We sequenced the Quartz-Seq2 sequence library with a NextSeq 500/550 High Output Kit v2 (75 cycles). Sequence specification was as follows (Read1, 23 cycles; Index1, 6 cycles; Read2, 63 cycles). The BCL files obtained were converted to FASTQ files using bcl2fastq2 (v2.17.1.14) with

demultiplexing pool barcodes. Each FASTQ file was split into single FASTQ files for each cell barcode using a custom script ([https://github.com/rikenbit/demultiplexer\\_quartz-seq2](https://github.com/rikenbit/demultiplexer_quartz-seq2), DOI: 10.5281/zenodo.2585429).



**Quartz-Seq2 splash-guard design.** **a.** For Quartz-Seq2, MoFlo Astrios EQ (Beckman Coulter) cell sorter was equipped with a custom-made splash-guard (red arrowhead). Splash-guard prevents droplet sorting into unexpected well-position. **b.** Specification of custom-made stainless-steel splash-guard. **c.** Splash-guard was attached to the embedded magnet-bar of the SortRescue tray. **d.** Photograph of splash-guard after single-cell sorting. Prevention of droplet sorting into unexpected well-position resulted in the spots of dried droplets on the splash-guard (purple line).

### inDrop System (1CellBio)<sup>2</sup>

Cells were isolated using an Aria3Fusion (BD Bioscience) cell sorter with a 100 $\mu$ m nozzle and a flow rate of 6-7. The sort rate was 40-50 events per second. In 30 min 80.000-90.000 cells were sorted. The landing buffer was PBS with 1% BSA, 0.6U/ $\mu$ l Ambion RNase, 0.3U/ $\mu$ l SuperaseIN. A total landing buffer volume of 670 $\mu$ l was used. The workflow was carried out using the inDrop instrument and the inDrop single cell RNA-seq kit (Cat. No. 20196, 1CellBio) according to the manufacturer's protocols. Microfluidic chips were prepared by silanization, and barcode labeled hydrogel microspheres (BHMs) were prepared shortly before cell capture, according to protocol (version v2.0., 1CellBio website). Droplet-making oil, single-cell suspension (200 cells/ $\mu$ L), and freshly prepared RT/lysis buffer were loaded onto the chip for droplet generation, according to the

inDrop protocol for single-cell encapsulation and reverse transcription (version 2.1., 1CellBio website). An emulsion corresponding to ~4000 droplets was collected in a cooled tube and irradiated with UV light to release the photo-cleavable barcoding oligos from the BHMs. cDNA synthesis proceeded within the droplets, and the emulsion was subsequently split into equal volumes in such a way as to not exceed ~2000 droplets per reaction tube. The 1CellBio run took 20-30 min (including time to adjust the speed for each fluid and to stabilize the flow). The collection of emulsion for library preparation took 5 min of the total time. After de-emulsification, cDNA contained in the aqueous phase was stored at -80°C. The RT product was further processed according to the InDrop library preparation protocol (version 1.2. 1CellBio website). The cDNA was fragmented by ExoI/HinfI and purified by AMPure XP beads. Second strand synthesis was conducted using NEB second-strand synthesis module (Cat. no. E6111S, NEB). In vitro-transcription was conducted using HiScribe T7 High Yield RNA Synthesis kit (cat. no. E2040S, NEB). Amplified RNA was then fragmented, and the fragments used in a second reverse transcription reaction with random hexamers to convert the sample back into DNA and to add a read primer-binding site to each molecule. Hybrid molecules of RNA and DNA were cleaned up using AMPure beads and amplified by PCR. Final libraries were sequenced using HiSeq4000 and NextSeq (Illumina). Sequence specification was as follows (Read1, 36 cycles; Index1, 6 cycles; Read2, 50 cycles).

#### **ICELL8 SMARTer Single-Cell System (Takara Bio)<sup>3</sup>**

Cells were isolated using an Aria3Fusion (BD Bioscience) cell sorter with a 100µm nozzle and a flow rate of 6-7. The sort rate was 40-50 events per second. In 30 min 80.000-90.000 cells were sorted. The landing buffer was PBS with 1% BSA, 0.6U/µl Ambion RNase, 0.3U/µl SuperaseIN. A total landing buffer volume of 670µl was used.

Hoechst 33342 and propidium iodide co-stained single-cell suspension (20 cells/µL) was distributed in eight wells of a 384-well source plate (Cat. No. 640018, Takara) and dispensed into a barcoded



SMARTer ICELL8 3' DE Chip with 5184 nano-wells (Cat. No. 640143, Takara) using an ICELL8 MultiSample NanoDispenser (MSND, Takara). 4 chips were used to target ~3000 single cells. Nanowells were imaged using the ICELL8 Imaging Station (Takara). Loading of the ICell8 nano-well chip was determined by the pre-defined ICell8 program, which took about 20 min. Subsequent chip imaging took 30-40 min. After imaging, the chip was sealed, placed in a pre-cooled freezing chamber, and stored at  $-80^{\circ}\text{C}$ . CellSelect software was used to identify each nanowell that contained a single cell. These nanowells were then selected for subsequent targeted deposition of a 50 nL/nanowell RT-PCR reaction solution from the SMARTer ICELL8 3' DE Reagent Kit (Cat. No. 640167, Takara) using the MSND. After RT and amplification in a Chip Cyclor, barcoded cDNA products from nanowells were pooled together using the SMARTer ICELL8 Collection Kit (Cat. No. 640048, Takara). cDNA was concentrated using the Zymo DNA Clean & Concentrator kit (Cat. No. D4013, Zymo Research), and purified using AMPure XP beads. cDNA was then used to construct Nextera XT (Illumina) DNA libraries, followed by 0.6X AMPure XP bead purification. Compared to the original 1CellBio protocol the following changes have been made: Step 1 to 26: Surfaces were cleaned with RNase AWAY® decontamination reagent. All tubes and reagents were kept RNase-free. Steps 3./4: Post-RT material volume was measured with a pipette while transferring it into the Costar Spin X tube filters resting on ice. Accordingly, the exact amount of Digestion Mix was calculated and prepared. Step 4: DNA Lobind tubes were used instead of Costar Spin X tubes. After steps 6 and 7: Tubes were vortexed and centrifuged briefly, respectively. Step 8: Agencourt® RNAClean™ XP beads from Beckman Coulter were used. Step 8b: The exact volume of digestion mix/post-RT-material was measured with a pipette to calculate the exact volume of beads needed. Step 8c: The incubation time was 10 min. Step 8i: The eluent was Nuclease-free water. Step 8j: Eluate was transferred into Axygen® 0.2 mL Maxymum Recovery® Thin Wall PCR Tubes. From this point onwards, all steps were performed in these tubes. Step 11: Incubation time was 15 hours. Step 12: Agencourt® RNAClean™ XP beads from Beckman Coulter were used. Step 29: During this purification the bead pellet was dried until it showed cracks

(approximately 2 min) before elution. Step 30: qPCR was performed with triplicates. AccuStart II PCR Tough Mix from QuantBio was used instead of 2x Kapa HiFi Hot Start PCR Mix. Step 32: For the library amplification PCR 1.5-2 cycles more than the Ct value from the diagnostic PCR were used. Step 33: A 50 $\mu$ l-reaction was set up with 10.5  $\mu$ l water; 9.5  $\mu$ l eluate; 25 $\mu$ l PCR Mix; and 5 $\mu$ l PE1/PE2 primer mix. AccuStart II PCR Tough Mix by QuantBio was used instead of 2x Kapa HiFi Hot Start PCR Mix. Step: 36: 50 $\mu$ l Elution buffer was added. Step 37: 70 $\mu$ l Ampure beads were added. The library was eluted in 40 $\mu$ l Elution buffer. The bead pellet was not dried excessively; it was still glossy. After step 37: A second bead purification was performed; 28 $\mu$ l beads were added to the 40 $\mu$ l eluate and processed as usual. The library was eluted in 20  $\mu$ l Elution buffer.

Library quantification and size distribution was done using Qubit, KAPA Library Quantification and Agilent TapeStation. Final libraries were sequenced using HiSeq4000 and NextSeq500 (Illumina). Sequence specification was as follows (Read1, 26 cycles; Index1, 8 cycles; Read2, 100 cycles).

#### **Drop-Seq (Dolomite)<sup>4</sup>**

Cells were sorted using a BD Aria Fusion and a 100 $\mu$ m nozzle (100 events per second). Single-cell RNA Drop-Seq experiments were performed using the scRNA system with P-Pumps and a scRNA-chip (100 $\mu$ m channel width) from Dolomite Bio (Royston, UK). Encapsulation was conducted according to the manufacturer's instructions, and library construction was completed according to the published DropSeq protocol<sup>4</sup>. Briefly, polyT-barcoded beads (MACOSKO-2011-10; ChemGenes) were loaded at a concentration of 600/ $\mu$ l, and cells at a concentration of 450/ $\mu$ l. The pumps were operated at a flowrate of 30  $\mu$ l/min for beads and cell suspension (PBS+2mM EDTA), and at 200  $\mu$ l/min for oil (QX200™ Droplet Generation Oil for EvaGreen; BioRad). After encapsulation, cell lysis, and hybridization of RNA to the beads, droplets were broken using PFO (Sigma-Aldrich) and aliquots of a maximum of 90000 beads were collected. Reverse transcription

was performed in a 200µl volume with Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific) and 2.5 µM TSO-primer (AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG; Qiagen) at room temperature for 30 min, followed by 90 min at 42°C. After exonuclease treatment (ExoI; New England Biolabs) at 37°C for 45 min in 200 µl, to digest the unbound primer, cDNA was amplified by PCR using HiFi HotStart mix (Kapa Biosystems) and amplification primer (AAGCAGTGGTATCAACGCAGAGT; Qiagen) in batches of 4000 beads in a volume of 50 µl (95°C - 3min; 4 cycles: 98°C - 20s, 65°C - 45s, 72°C - 3min; 9 cycles: 98°C - 20s, 67°C - 20s, 72°C - 3min; 72°C - 5min). Libraries were generated using the Nextera XT library Kit (Illumina) in five pooled PCR samples with 600 pg of cDNA and a custom P5-primer (AATGATACGGCGACCAACCGAGATCTACACGCCTGTCCGCGGAAGCAGTTGGTATCAACGCAGAGT\*A\*C; Qiagen). Final library QC was conducted using the BioAnalyzer High Sensitivity DNA Chip (Agilent Technologies). For sequencing on an Illumina HiSeq2500 V4, we used a custom read 1 primer (GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC; Qiagen). Sequence specification was as follows (Read1, 75 cycles; Index1, 8 cycles; Index2, 8 cycles; Read2, 75 cycles).

### **Chromium V2 (10X Genomics): Single-cell RNA sequencing<sup>5</sup>**

Two cell preparations were conducted on two different days: one to prepare 2 libraries for sequencing at high read depth, and one to prepare 8 libraries at low read depth. To prepare the libraries for high read depth, one frozen vial of a Human Cell Atlas reference sample was thawed and prepared as described. At the end of this protocol, the cells were resuspended in PBS with 2 mM EDTA. Since cells showed clumping and low viability, they were centrifuged 3 times at 150 g for 10 min at room temperature, and resuspended in 50% PBS, 2mM EDTA and 50% Iscove's Modified Dulbecco Medium (IMDM, ATCC) supplemented with 10% FBS and filtered through a 40µm FlowMi cell strainer (Sigma-Aldrich) to remove cell aggregates and large cell debris. At the final count before loading, the cell suspension demonstrated a viability of 60%. To prepare the

libraries for low read depth, two frozen vials of a the reference sample were thawed and prepared as described in an updated version of the HCA Benchmark protocol. At the end of this protocol, the cells were resuspended in IMDM, 10% FBS and 1mM EDTA, and filtered through a 40- $\mu$ m FlowMi cell strainer to remove cell aggregates and large cell debris. At the final count before loading, the cell suspension demonstrated a viability of 65%. The cells were not processed using FACS isolation, but run directly on the 10x Chromium system (10x Genomics, Pleasanton, CA, USA).

Cells were mixed with single-cell master mix, and the resulting cell suspensions were loaded on a 10x Chromium system to generate 2 libraries at 5,000 cells each and 5 libraries at 10,000 cells each. The single-cell libraries were generated using 10x Chromium Single Cell gene expression V2 reagent kits according to the manufacturer's instructions (Chromium single cell 3' reagents kits v2 user guide). Single cell 3' RNA-seq libraries were quantified using an Agilent Bioanalyzer with a high sensitivity chip (Agilent), and a Kapa DNA quantification kit for Illumina platforms (Kapa Biosystems). The libraries were pooled according to the target cell number loaded. Sequencing libraries were loaded at 200 pM on an Illumina NovaSeq6000 with Novaseq S2 Reagent Kit (100 cycles) using the following read lengths: 26 bp Read1, 8 bp I7 Index and 91 bp Read2. The 2 libraries of 5,000 cells and the 8 libraries of 10,000 cells were sequenced at 250,000 and 25,000 reads per cell, respectively.

#### **Chromium V2 (10X Genomics): Single-nucleus RNA sequencing**

We isolated nuclei from the cell suspension using a protocol provided by 10x Genomics (Isolation of Nuclei for Single Cell RNA Sequencing - Demonstrated Protocol - Sample Prep - Single Cell Gene Expression - Official 10x Genomics Support). We counted the nuclei using a Countess II (Thermo Fisher Scientific). We made an aliquot containing ~11,000 nuclei in a volume of 33.8  $\mu$ L in RB buffer (1x PBS, 1% BSA, and 0.2U/ $\mu$ l RNaseIn (TaKaRa)) as sample A, and stained the rest of the nuclei suspension with Vybrant DyeCycle Violet Stain (Thermo Fisher Scientific) at a

concentration of 10  $\mu\text{M}$ . We used a MoFlo Astrios EQ cell sorter (Beckman Coulter) and set fluorescence activated cell sorting (FACS) gating on forward scatter plot, side scatter plot and on fluorescent channels to pick Violet-positive (for nuclei), while excluding debris and doublets. We used a 100  $\mu\text{m}$  nozzle to sort 20,000 nuclei at a rate of 340 events per second into 20  $\mu\text{l}$  RB buffer resulting in a final volume of about 70  $\mu\text{l}$ . After sorting, we measured the volume of B with a pipette, spun it at 500 g for 5 min at 4°C, and then carefully removed part of the supernatant to leave ~40 $\mu\text{l}$ . We resuspended B by gentle pipetting 40 times.

Immediately after nuclei isolation, we loaded sample A into one channel of a Chromium Single Cell 3' Chip (10x Genomics, PN-120236), and then processed it through the Chromium Controller to generate GEMs (Gel Beads in Emulsion). We then loaded 33.8  $\mu\text{L}$  of B 25 minutes later after sorting and centrifugation, as described above, into one channel of a second chip, and processed it in the same way as the first chip. We prepared RNA-Seq libraries for both samples in parallel with the Chromium Single Cell 3' Library & Gel Bead Kit V2 (10x Genomics, PN-120237), according to the manufacturer's protocol. We pooled the 2 samples based on molar concentrations and sequenced them on a NextSeq500 instrument (Illumina) with 26 bases for Read 1, 57 bases for Read 2, and 8 bases for Index Read 1.

### Smart-seq2<sup>6</sup>

Cells were sorted using a BD Aria III and a 100 $\mu\text{m}$  nozzle (100 events per second). Smart-seq2 libraries were prepared at half the volume, as described previously<sup>6</sup>, with minor modifications. In brief, 2  $\mu\text{l}$  of lysis buffer containing 0.1 % Triton X-100 (Sigma-Aldrich), 1 U/ $\mu\text{l}$  RNase inhibitor (Takara), 2.5 mM dNTPs (Thermo Fisher) and 2  $\mu\text{M}$  oligo-dT primer (5'-AAGCAGTGGTATCAACGCAGAGTACT30VN-3'; IDT) were dispensed into each well of a 384-well plate (4titude). Lysis plates were stored at -20°C until cell sorting, after which single-cell lysates were kept at -80 °C. Before reverse transcription, cell lysates were denatured at 72 °C for 3 min and immediately placed on ice. The RT reaction was performed in a 5  $\mu\text{l}$  total volume, with

final reagent concentrations of 1x Superscript first-strand buffer (Thermo Fisher), 5 mM DTT (Thermo Fisher), 1 M Betaine (Sigma-Aldrich), 9 mM MgCl<sub>2</sub> (Sigma-Aldrich), 1 U/μl RNase inhibitor (Takara), 1 μM LNA template-switching oligo (5'-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3'; Exiqon), and 10 U/μl Superscript II RT enzyme. Next, pre-amplification PCR was performed for 22 cycles at final concentrations of 1x KAPA HiFi HotStart ReadyMix (Roche) and 0.08 μM ISPCR primer (5'-AAGCAGTGGTATCAACGCAGAGT-3'; IDT) in a total reaction volume of 11 μl. The cDNA was cleaned up by adding 10 μl of SPRI beads (bead stock composition: 19.5 % PEG, 1 M NaCl, 1 mM EDTA, 0.01 % IGEPAL CA-630), washing twice with 20 μl 80 % ethanol, and eluting in 10 μl H<sub>2</sub>O. The cDNA concentration was measured for all wells using Picogreen dsDNA assay (Thermo Fisher), and diluted to 200 pg/μl using a Mantis liquid handler (Formulatrix). Next, 1 μl of cDNA was used as input for the Nextera XT library preparation kit (Illumina) at 1/5 volume, according to the manufacturer's instructions. During the 12 cycles library PCR, custom i7 and i5 indexing primers (IDT) were added at 0.5 μM each. Finally, 5 μl of library per well were pooled, cleaned and concentrated using SPRI beads (19.5 % PEG; see above). Final libraries were sequenced using HiSeq2500 V4 (Illumina). Sequence specification was as follows (Read1, 75 cycles; Index1, 8 cycles; Index2, 8 cycles; Read2, 75 cycles).

#### **CEL-Seq<sup>2</sup>**<sup>7,8</sup>

Single-cell RNA sequencing was performed using a modified version of the mCEL-Seq2 protocol, an automated and miniaturized version of CEL-Seq2, on a Mosquito nanoliter-scale liquid-handling robot (TTP LabTech). A detailed step-by-step protocol is available<sup>8</sup>. Briefly, cells were sorted using a BD Aria Fusion and a 100μm nozzle (100 events per second) into 384-well plates (Bio-Rad) containing 240 nl of lysis buffer containing polyT primers and 1.2 μl of mineral oil (Sigma-Aldrich). Sorted plates were centrifuged at 2200 x g for several minutes at 4°C, snap-frozen in liquid nitrogen and stored at -80°C until processing. On the day of processing, sorted plates were

thawed on ice and heat lysed at 95°C for 3 min prior to cDNA synthesis. 160nl of reverse transcription reaction mix and 2.2 µl of second strand reaction mix were used to convert RNA into cDNA. cDNA from 96-cells were pooled together before clean up and in vitro transcription, generating 4 libraries from one 384-well plate. 11 PCR cycles were used for library amplification. During all purification steps, including the library cleanup, we used 0.8 µl of AMPure/RNAClean XP beads (Beckman Coulter) per 1 µl of sample. Sixteen libraries with 96 cells each (one of the libraries contained 30,000 RNA molecules from ERCC spike-in mix per cell) were sequenced on an Illumina HiSeq3000 sequencing system (pair-end multiplexing run). Sequence specification was as follows (Read1, 30 cycles; Read2, 75 cycles).

### **MARS-Seq<sup>9</sup>**

To construct single-cell libraries from poly(A)-tailed RNA, we used massively parallel single-cell RNA sequencing (MARS-Seq). Briefly, single cells were FACS-isolated with a BD Aria III and a 100µm nozzle (100 events per second) into 384-well plates containing lysis buffer (0.2% Triton X-100 (Sigma-Aldrich); RNase inhibitor (Invitrogen)) and reverse-transcription (RT) primers. Single-cell lysates were denatured and immediately placed on ice. The RT reaction mix, containing SuperScript III reverse transcriptase (Invitrogen), was added to each sample. After RT, the cDNA was pooled using an automated pipeline (epMotion, Eppendorf). Unbound primers were eliminated by incubating the cDNA with exonuclease I (NEB). A second stage of pooling was performed through cleanup with SPRI magnetic beads (Beckman Coulter). Subsequently, pooled cDNAs were converted into double-stranded DNA using the Second Strand Synthesis enzyme (NEB), followed by clean-up and linear amplification by T7 *in vitro* transcription overnight. The DNA template was then removed by Turbo DNase I (Ambion), and the RNA purified using SPRI beads. Amplified RNA was chemically fragmented using Zn<sup>2+</sup> (Ambion), and then purified using SPRI beads. The fragmented RNA was ligated with ligation primers containing a pool barcode and partial Illumina Read1 sequencing adapter using T4 RNA ligase I (NEB). The ligated products were reverse-

transcribed using the Affinity Script RT enzyme (Agilent Technologies) and a primer complementary to the ligated adapter, partial Read1. The cDNA was purified using SPRI beads. Libraries were completed by a PCR step using the KAPA Hifi Hotstart ReadyMix (Kapa Biosystems) and a forward primer containing the Illumina P5-Read1 sequence, and a reverse primer containing the P7-Read2 sequence. The final library was purified using SPRI beads to remove excess primers. Library concentration and molecular size were determined with a High Sensitivity DNA Chip (Agilent Technologies). Multiplexed pools were run on Illumina HiSeq2500 Rapid flow cells (Illumina). Sequence specification was as follows (Read1, 52 cycles; Index1, 7 cycles; Read2, 15 cycles).

#### **C1 High-Throughput (HT-IFC)<sup>10</sup>**

Cells were sorted into 15-ml tubes containing 7 ml of PBS with 5% FBS, using a Sony SH800 Cell Sorter. Cells were concentrated by centrifugation at 350 x g for 5 minutes at 4°C (recovery 81%). The supernatant was removed, and cells were counted and diluted to 900 cells/ul for the Fluidigm C1 HT Small-Cell Integrated Fluidic Circuits (IFCs), and 450 cells/ul for the Fluidigm C1 HT Medium-Cell IFCs. A total of eight small-cell and seven medium-cell IFCs were used to generate cDNA on the Fluidigm C1 System. cDNA generation and the subsequent preparation of sequencing libraries were performed according to the recommended Fluidigm C1 HT protocols. Enrichment Primers from the Fluidigm reagent kit were replaced with NEBNext i5xx primers from NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1 & 2) (New England BioLabs), to enable library pooling. Libraries from fifteen IFCs were pooled and sequenced on the NovaSeq6000 system (Illumina) in two runs on the S2 flow cell. Sequence specification was as follows (Read1, 26 cycles; Index1, 8 cycles; Read2, 85 cycles).



### **ddSEQ (Bio-Rad)**

Flow cytometry analysis and cell sorting were performed on the S3e Cell Sorter using ProSort Software (Bio-Rad Laboratories, #12007058) for acquisition and sorting. 41,749 viable cells were sorted with a 100  $\mu\text{m}$  nozzle at 231 events per second into 1x PBS with + 0.1% BSA and kept at 4°C until scRNA-Seq (approx. 1 h). Cell concentration of sorted cells was determined using the TC20 Automated Cell Counter (Bio-Rad Laboratories, #1450102) and adjusted to a final concentration of 2,500 cells/ $\mu\text{l}$ . Cells were then prepared for single-cell sequencing using the Illumina Bio-Rad SureCell WTA 3' Library Prep Kit for the ddSEQ (Illumina, #20014280). Cells were loaded onto ddSEQ cartridges and processed in the ddSEQ Single-Cell Isolator (Bio-Rad Laboratories, #12004336) to isolate and barcode single cells in droplets. First-strand cDNA synthesis occurred in droplets, which were then disrupted for second strand cDNA synthesis in bulk. Libraries were prepared according to manufacturer's instructions and then sequenced on the NextSeq500 system (Illumina).

### **gmcSCRB-seq<sup>11</sup>**

Cells were sorted and processed using the alternative lysis (Guanidine Hydrochloride) condition (gmcSCRB-seq) as described suitable for PBMCs in Bagnoli et al (2018). Briefly, single cells ("3 drops" purity mode) were sorted (Sample pressure: 5, 2-20 events per second) into 96-well DNA LoBind plates (Eppendorf) containing 5  $\mu\text{l}$  lysis buffer using a Sony SH800 sorter (Sony Biotechnology #LE-SH800SZGCPL; Chip series: LE-C32, 100  $\mu\text{m}$ ). 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). Samples were processed in six batches, with one batch of two plates and five batches of six plates. SPRI Beads (GE Healthcare) were prepared and diluted 50-fold (final concentration 1 mg/mL) in bead-binding buffer (22% PEG8000 (w/v), 1M NaCl, 10mM Tris-HCl pH 8.0, 1 mM EDTA, 0.01% IGEPAL, 0.05% Sodium Azide ). Each well was cleaned up using a ratio of 2:1 of 1  $\mu\text{g}/\mu\text{L}$  beads (10  $\mu\text{L}$  beads and 5  $\mu\text{L}$  lysate) and

resuspended in 4  $\mu$ l H<sub>2</sub>O (Invitrogen) and a mix of 5  $\mu$ l reverse transcription master mix, consisting of 20 units Maxima H- enzyme (Thermo Fisher), 2  $\times$  Maxima H- Buffer (Thermo Fisher), 2 mM each dNTPs (Thermo Fisher), 4  $\mu$ M template-switching oligo (IDT), and 15% PEG 8000 (Sigma-Aldrich). For libraries containing ERCCs, 30,000 molecules of ERCC spike-in Mix 1 (Ambion) was used and the H<sub>2</sub>O (Invitrogen) was adjusted accordingly. After the addition of 1  $\mu$ l 2  $\mu$ M barcoded oligo-dT primer (E3V6NEXT, IDT), cDNA synthesis and template switching was performed for 90 min at 42  $^{\circ}$ C. Barcoded cDNA and remaining beads were then pooled in 2 ml DNA LoBind tubes (Eppendorf) and an equal volume of bead-binding buffer was added. Purified cDNA was eluted in 17  $\mu$ l and residual primers digested with Exonuclease I (Thermo Fisher) for 20 min at 37  $^{\circ}$ C. After heat inactivation for 10 min at 80  $^{\circ}$ C, 30  $\mu$ l PCR master mix consisting of 1.25 U Terra direct polymerase (Clontech) 1.66  $\times$  Terra direct buffer and 0.33  $\mu$ M SINGV6 primer (IDT) was added. PCR was cycled as given: 3 min at 98  $^{\circ}$ C for initial denaturation followed by 19 cycles of 15 s at 98  $^{\circ}$ C, 30 s at 65  $^{\circ}$ C, 4 min at 68  $^{\circ}$ C. Final elongation was performed for 10 min at 72  $^{\circ}$ C. Batch 4 was erroneously denatured for 10 min due to a cyclers error, but left in as we consider such errors as possible batch variation errors.

Following pre-amplification, all samples were purified using SPRI beads at a ratio of 1:0.8 of 1  $\mu$ g/ $\mu$ L beads (40  $\mu$ L beads and 50  $\mu$ L sample) with a final elution in 10  $\mu$ l of H<sub>2</sub>O (Invitrogen). The cDNA was then quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher). Size distributions were checked using high-sensitivity DNA Fragment Analyzer kits (AATI) and high-sensitivity DNA Bioanalyzer kits (Agilent). As the samples had large primer peaks, they were purified a second time using SPRI beads at a ratio of 1:0.8 and then pre-amplified for an additional 3 cycles, as above. The cDNA was then purified and reanalyzed as above. Samples passing the quantity and quality controls were used to construct Nextera XT libraries from 0.8 ng of pre-amplified 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. Libraries were sequenced on high output flow cells of an Illumina HiSeq 1500 instrument. Sequence specification was as follows (Read1, 16 cycles; Index1, 8 cycles; Read2, 50 cycles). 16 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. An additional 8 base i7 barcode read was done to allow multiplexing.

## Data analysis

### Primary data preprocessing

FASTQ files for each technique were collected and processed in a unified manner. We developed a snakemake<sup>12</sup> workflow that streamlines all steps, including read filtering and mapping, quantification, downsampling and species deconvolution, and provides a Single Cell Experiment Object<sup>13</sup> output with detailed metadata. We used zUMIs<sup>14</sup>, a single-cell processing tool compatible with all major scRNA-Seq protocols for filtering, mapping and quantification, ensuring comparable primary data processing between all methods. First, we discarded low-quality reads (barcodes and UMI sequences with more than 1 base below the Phred quality threshold of 20) and removed barcodes with less than 100 reads.

For techniques with known barcodes, we provided zUMIs with these barcode sequences, and used the automatic barcode detection function to detect the sequenced cells for other techniques. Next, cDNA reads were mapped to the human GRCh38, mouse GRCm38, and a human-mouse-dog mixed (for species level doublet detection) reference genomes using STAR<sup>15</sup>. Reads were then assigned to exonic and intronic features using featureCounts<sup>16</sup> and counted using the default parameters of zUMIs for human-only, mouse-only and mixed bam-files, separately. The output expression matrix of reads mapping to both exonic and intronic regions was selected for the downstream analysis. Of note, we included intronic counts in the expression quantification to improve gene detection and to enable a comparison with the snRNA-seq derived dataset. To deconvolute species, detect doublets and low quality cells, the mixed-species mapped data was used. Cells for which >70% of the reads mapped to only one species were assigned to the corresponding species. The remaining cells (those for which <70% of the reads mapped to only one species) were removed from the downstream analysis. Finally, for each technique, a *human* and *mouse* Single Cell Experiment object was created by combining the expression matrix and the metadata.

For subsequent data analysis, we discarded cells with <10,000 total number of reads as well as the cells having <65% of the reads mapped to their reference genome. Cells in the 95th percentile of the number of genes/cell and those having <25% mitochondrial gene content were included in the downstream analyses. Genes that were expressed in less than five cells were removed.

For the comparative viability analysis, we used EmptyDrops<sup>17</sup> to determine the inflection point on the ranked barcodes vs number of detected UMIs for each library separately. We assigned all barcodes before the inflection point as cells and the remaining as empty drops.

### Clustering

Filtering, normalization, selection of highly variable genes (HVG), and clustering of cells were performed using the Seurat<sup>18</sup> package (version 2.3.4). We normalized the gene expression measurements for each cell by the total expression, multiplied by a scale factor (10e4), and log-transformed the result. We used 10e4 (instead of 10e6 more commonly used in bulk RNA-seq) due to the reduced number of transcripts present in single-cell data. To avoid spurious correlations, the library sizes were regressed out, and the genes were scaled and centered. The scaled Z-score values were then used as normalized gene measurement input for clustering and for visualizing differences in expression between cell clusters. We selected HVGs by evaluating the relationship between gene dispersion ( $y.cutoff = 0.5$ ) and the log mean expression. The clustering procedure projects cells onto a reduced dimensional space, and then groups them into subpopulations by computing a shared-nearest-neighbour (SNN) based on the Euclidean distance (finding highly interconnected communities). The algorithm is a variant of the Louvain method, which uses a resolution parameter to determine the number of clusters.

In this step, the dimension of the subspace was set to the number of significant principal components (PC) based on the distribution of the PC standard deviations and by inspecting the ElbowPlot graph. For downsampled data, the number of PCs was set to 8 after inspecting all ElbowPlot separately. The number of clusters was aligned to the expected biological variability, and

cluster identities were assigned using previously described gene markers<sup>5,19</sup>. T-SNE and UMAP were used to visualize the clustering distribution of cells. Cluster-specific markers were then identified using the Wilcoxon rank-sum test.

Trajectory analysis and pseudo-ordering of cells was performed using the Monocle<sup>20</sup> package (version 2.8.0) with the previously identified HVGs. Monocle works with the raw data and allows to specify the family distribution of gene measurements, which was set to a negative binomial, as defined in the family function from the VGAM package. As for the clustering, the expression space was reduced before ordering cells using the DDRTree algorithm. To validate cell populations, and for cell type identification and annotation, we used pseudotime ordering of single cells derived from the mouse colon.

#### **Sample deconvolution and annotation**

To identify and annotate cell types and states, we analyzed the individual single-cell experiments separately, taking advantage of the original sequencing depth. Gene expression counts were log-normalized to identify HVGs, as input to compute cell-to-cell distances and graph-based clustering (see Clustering). Cell clusters were visualized in two-dimensional space using t-SNE and UMAP, and then annotated by examining previously described cell population marker genes<sup>5,19</sup>

(**Supplementary Fig. 8 and 9**). All methods were able to recapitulate most cell types in both human and mouse samples, although in different proportions and resolutions.

In human samples, the T-cell marker CD3 was used to differentiate T-cells from other populations. While the CD4 T-cells cluster was clearly identifiable (with non-overlapping expression of markers), CD8 T-cells and Natural Killer (NK) were often intermixed. Monocytes were the second most abundant cell type, including subpopulations of CD14 and FCGR3A monocytes. High levels of CD79A and CD79B allowed the clear identification of B-cells. HEK293T cells generally fell into the same cluster, separate from blood subpopulations. They were clearly identifiable by the high number of detected genes (up to six-fold higher than PBMC populations). However, there was a

correlation between the expression profiles of immune cells, leading in some instances to mixtures of PBMCs and HEK293T cells.

With few exceptions (Chromium), significantly fewer cells mapped to the mouse genome (half that of human cells, on average), leading to poorer clustering performance. However, the expected subpopulation composition of the colon was maintained overall. A small set of putative intestinal stem cells (*Lgr5* and *Smoc2* expression) were close (in transcriptional space) to rapidly proliferating transit amplifying (TA) cells (showing high ribosomal genes). Secretory cells (e.g. *Muc2*, *Tff3*, *Agr2*) resulted in a well-defined cluster. Enterocytes were more heterogeneous and ordered along their grade of lineage commitment. Notably, in some experiments two distinct clusters of enterocytes were identified, as well as a very small group of enterocyte progenitors. In addition to colon cells, fibroblasts and immune-cells were detected in all samples.

### Reference datasets

To compare the efficiency of scRNA-seq protocols in describing the structure of a mixed population, we produced a reference dataset with 30,807 human and 19,749 mouse cells. Cells were clustered and annotated as described above. Due to the high number of cells, major cell types were clustered and clearly identifiable using population marker genes (**Supplementary Fig. 4a-b**)<sup>5,19</sup>.

However, to improve cell-to-cell annotations, we combined clustering with additional analyses. To annotate human blood cells, we used *matchScore2* (see Methods) using an annotated set of 2700 PBMCs<sup>5</sup> as reference (**Supplementary Fig. 4c-d**). We used cluster-specific markers of annotated populations as input to create a multinomial logistic model according to the *matchScore2* algorithm. For each unknown cell, we assigned probability values for any possible cell identity, and the most likely identity was used for the classification (where this probability was >0.5; otherwise the cell was considered unclassified). Cell identities inferred by *matchScore2* were highly consistent with clusters, with agreement ranging from 96% for CD4 T-cells to 100% for B-cells. Cell-by-cell prediction helped to identify smaller cell subsets, such as FCGR3A monocytes,

dendritic cells and megakaryocytes. For all clusters, 17% of the cells remained unclassified (**Supplementary Fig. 4c**). Half of these were previously annotated as HEK293T cells, which split into three different clusters because they varied in number of genes (**Supplementary Fig. 4d**). Cells with fewer genes (cluster HEK293T cell2 and partially HEK293T cell3) were classified as CD4+ T-cells, although these did not show expression of any of the key blood markers. For the purposes of subsequent analysis, we removed the *unclear* cluster, representing 1% of the total number of cells, as well as the unclassified cells (except cells in HEK293T clusters). To further validate annotations, we assigned a score to each cell, corresponding to the overall expression of cell type signatures from the list of the top 100 computational markers (**Supplementary Fig. 4d**). Transcriptional signatures revealed a set of cells from the HEK293T cell1 and HEK293T cell2 clusters showing high scores ( $>0.5$ , range 0-1) for multiple signatures. We considered these as potential doublets, and removed them. The remaining cells were then used to compute an unbiased set of cell-type specific markers.

In the case of the mouse reference sample, we used clustering to dissect the colon subpopulation structure (excluding immune cells and fibroblasts). The largest cluster was formed by immature enterocytes (**Supplementary Fig. 5a-b**). Other clusters included similar proportions of mature enterocytes, secretory cells, transit-amplifying cells and other undifferentiated cells. To refine annotations of immature cells, we ordered cells by intermediate states and projected them along a trajectory (see Clustering). The trajectory analysis (**Supplementary Fig. 5c-d**) revealed 9 different states, ranging from intestinal stem cells and transit-amplifying cells (expressing high levels of *Lgr5*, *Smoc2*, *Top2a*) to enterocytes (*Slc26a3*, *Saa1*). Based on the pseudo-ordering and expression levels of previously described markers, states were merged into four major groups (**Supplementary Fig. 5d**). For annotation, we labeled these four groups as Intestinal Stem cells (ISC), Transit Amplifying cells (TA), Enterocyte progenitors (Epr), and Enterocyte (E). We combined this finer-grained annotation with the remaining cell types, and then computed population-specific gene markers for training the reference model.



## Supplementary Notes References

1. Sasagawa, Y. *et al.* Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, (2018).
2. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187–1201 (2015).
3. Goldstein, L. D. *et al.* Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics* **18**, 519 (2017).
4. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
5. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
6. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
7. Herman, J. S., Sagar, null & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379–386 (2018).
8. Sagar, null, Herman, J. S., Pospisilik, J. A. & Grün, D. High-Throughput Single-Cell RNA Sequencing and Data Analysis. *Methods Mol. Biol. Clifton NJ* **1766**, 257–283 (2018).
9. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
10. Barriga, F. M. *et al.* Mex3a Marks a Slowly Dividing Subpopulation of Lgr5+ Intestinal Stem Cells. *Cell Stem Cell* **20**, 801-816.e7 (2017).
11. Bagnoli, J. W. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCRBS-seq. *Nat. Commun.* **9**, 2937 (2018).
12. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinforma. Oxf. Engl.* **28**, 2520–2522 (2012).
13. SingleCellExperiment: S4 Classes for Single Cell Data version 1.4.1 from Bioconductor. <https://rdrr.io/bioc/SingleCellExperiment/>.
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**, (2018).
15. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).
16. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma. Oxf. Engl.* **30**, 923–930 (2014).
17. Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
18. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
19. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
20. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).

### Supplementary Figure legends

#### Supplementary Fig. 1. The effect of viability sorting on data quality (human cells).

**a,b.** Quality control displaying the number of detected genes and the relative proportion of reads mapped to mitochondrial transcripts (**a**; indicating cell damage) or total number of mapped reads (**b**). Cells with a mitochondrial proportion >25% and <1,778 ( $\log_{10}=3.25$ ) sequencing reads were considered low-quality cells. **c.** T-SNE visualizations of unsupervised clustering in human samples with (left, 4,941 cells) or without (right, 4,094 cells) viability selection. Each dataset was analyzed separately and cells are colored by cell types inferred by *matchScore2*. Cells that did not reach a probability score of 0.5 for any cell type were considered unclassified. **d.** Cell type composition of samples with or without viability selection with annotations from the reference dataset.

#### Supplementary Fig. 2. The effect of viability sorting on data quality (mouse cells).

**a,b.** Quality control displaying the number of detected genes and the relative proportion of reads mapped to mitochondrial transcripts (**a**; indicating cell damage) or total number of mapped reads (**b**). Cells with a mitochondrial proportion >25% and <1,778 ( $\log_{10}=3.25$ ) sequencing reads were considered low-quality cells. **c.** Relationship between the number of mapped reads and detected genes for high-quality cells color-coded by cell type. **d.** T-SNE visualizations of unsupervised clustering in mouse samples with (left, 1,159 cells) or without (right, 4,245 cells) viability selection. Each dataset is analyzed separately and cells are colored by cell types inferred by *matchScore2*. **d.** Cell type composition of samples with or without viability selection with annotations from the reference dataset.

#### Supplementary Fig. 3. Gene expression levels of selected marker genes.

UMAP visualization of normalized expression levels for selected marker genes of the most common PBMC (**a**, 30,807 cells) and colon (**b**, 19,749 cells) populations. Maps are shown for CD4+ T-cell markers IL7R and CD4 (expressed also in monocytes), the CD8+ T-cell marker CD8A, the B-cell marker CD79A, NK cell markers GNLY and NKG7, and monocyte-specific markers LYZ, CD14 and FCGR3A. In (**b**) maps are shown for markers of Intestinal Stem cell and proliferation (*Smoc2*, *Miki67* and *Top2a*), secretory markers (*Muc2*, *Agr2* and *Tff3*), enteroendocrine cell markers (*Chga* and *Chgb*), and enterocyte markers (*Slc26a3*, *Car1* and *Fabp2*).

#### Supplementary Fig. 4. Identifying PBMC cell types using unsupervised clustering and classification.

**a.** UMAP visualization of 38,195 human PBMC and HEK293T human cells colored according to their assignment to clusters. Cluster labels are defined by examining the expression levels of known markers. **b.** Heatmap indicating the relative expression and gene detection rates for most common PBMC marker genes. **c.** UMAP visualization of 38,195 PBMC and HEK293T cells color coded by cell classification inferred by *matchScore2*. 17% of cells were unclassified and were removed from the analysis. **d.** UMAP visualization of 38,195 PBMC and HEK293T cells showing the number of genes per cell, and scores for transcriptional signatures obtained by computing cell-type-specific markers (*lightgray*: low-score, *blue*: high score).

#### Supplementary Fig. 5. Identifying colon cell types by unsupervised clustering and trajectory analysis.

**a.** UMAP visualization of 17,558 mouse colon cells. Cells are colored by their assignment to clusters. Annotations are defined by examining the expression of known markers and differentially expressed genes (DEG). **b.** Heatmap of top DEG per cluster. Key markers of common colon cell populations are shown. **c.** Trajectory and pseudotime analysis of 8,716 immature enterocytes (IE) showing the transition from intestinal stem cells (ISC) to enterocytes. Trajectories with the relative expression of known markers are shown (yellow: low, gray: mid, blue: high). **d.** (Top) Ordered 17,558 colon cells are grouped into four different states according to their differentiation stage: intestinal stem cell (ISC), transit amplifying (TA), enterocyte progenitor (Epr), Enterocytes (E). (Bottom) UMAP visualization of IE cells colored according to the four resulting states.

#### Supplementary Fig. 6. Comparison of PBMC human reference with PBMC data from Zheng et al., (Nature Communications 2017).

**a.** UMAP visualization of 2,700 PBMCs from the Zheng et al. Chromium PBMC-3k dataset (left) and our human reference dataset (right). The colors indicate the cell types based on the annotation of the PBMC-3k dataset. Cell labels are transferred from the PBMC-3k data using the *matchScore2* classification. **b.** Jaccard Indexes (JI) of cell type-specific markers from the two datasets (30,807 vs 2,700). For each annotated cluster,

the top 100 ranked markers were considered. **c.** Cell type composition of our human reference clusters with annotations from the PBMC-3k dataset.

**Supplementary Fig. 7. Comparison of our mouse colon reference with the Tabula Muris (TM) colon dataset (Nature 2019).**

**a.** UMAP visualization of 3,938 colon cells from the Smart-seq2 TM dataset (left) and our mouse reference dataset (right). Colors indicate the cell type based on the annotation provided by the TM Consortium. Cell labels of the mouse reference are transferred from the TM using the *matchScore2* classification. **b.** Jaccard Indexes (JI) of cell type-specific markers from the two datasets (19,749 vs 3,938). For each annotated cluster, the top 100 ranked markers were considered. **c.** Cell type composition of our mouse reference clusters with annotations from the TM dataset.

**Supplementary Fig. 8. Clustering analysis of 13 sc/snRNA-seq methods.**

T-SNE visualizations of unsupervised clustering in human samples from 13 different methods. Each dataset is analyzed separately by taking advantage of its original sequencing depth. Cells are colored by cell type inferred by *matchScore2*. Cells that did not reach a probability score of 0.5 for any cell type were considered unclassified.

**Supplementary Fig. 9. Clustering analysis of 11 sc/snRNA-seq methods.**

T-SNE visualizations of the unsupervised clustering in mouse samples from 11 different methods. Each dataset is analyzed separately by taking advantage of its original depth. Cells are colored according to cell type inferred by *matchScore2*. Cells that did not reach a probability score of 0.5 for any cell types were considered unclassified.

**Supplementary Fig. 10. Downsampling iterations.**

Number of detected molecules per cell type (HEK293T, monocytes and B cells) with 5 downsampling iterations and at different downsampling thresholds (5K, 10K, 15K, 20K, 50K).

**Supplementary Fig. 11. Performance comparison of 13 scRNA sequencing methods.**

**a.** Boxplots comparing the number of detected genes across protocols on downsampled data (20K), in mouse secretory and transit amplifying cells. Cell identities were defined by cell projection onto the reference. **b.** Number of genes detected at step-wise downsampled sequencing depths. Points represent the average number of genes detected for all cells of the corresponding cell type at the corresponding sequencing depth. **c,d.** Boxplots comparing the number of detected genes from countification of reads mapping to only exonic regions (**c**) and UMI (**d**, from exonic and intronic counts) across protocols on downsampled data (20K) of human HEK293T cells, monocytes and B-cells. All the boxplots display the minimum, 1st, 2nd, 3rd quantiles and maximum values.

**Supplementary Fig. 12. Performance across Chromium versions and application types (sc/snRNA-seq).**

**a,b.** Boxplots comparing the number of molecules (**a**) and genes (**b**), in downsampled (10K) HEK293T cells, monocytes and B-cells. The results are displayed for gene quantification including (open boxes) or excluding (filled boxes) intronically mapping reads. **c.** Cumulative gene counts per protocol as the average of 50 randomly sampled HEK293T cells, monocytes and B-cells on downsampled data (10K). **d.** Overlap of detected genes using cumulative gene counts from the maximum of consistently detected cells numbers (HEK293T: 46, Monocytes: 50, B-cells: 13) on downsampled (10K) data from different cells types. All the boxplots display the minimum, 1st, 2nd, 3rd quantiles and maximum values.

**Supplementary Fig. 13. Technical reproducibility within sc/snRNA-seq protocols. a,b.**

Boxplots comparing the number of genes detected across processing units (e.g. plates, droplet lanes and IFCs), in downsampled (20K) HEK293T (**a**) and B-cells (**b**). Each protocol was stratified into processing units and only replicates with >5 cells were included. All the boxplots display the minimum, 1st, 2nd, 3rd quantiles and maximum values. **c,d.** Pearson correlation plots across replicates using the expression of all genes and cells per replicate for HEK293T (**c**) and B-cells (**d**). Protocols are ordered by Ward agglomerative hierarchical clustering. **e.** R-squared measures of the PC regression model using KBET to quantify variation in the total human dataset introduced by processing units (Online Methods).

**Supplementary Fig. 14. T-SNE representation of human cell types using highly variable genes.**

**a,b.** T-SNE representation (calculated on first 8 principle components) on downsampled data (20K) using highly variable genes across protocols, separated by HEK293T cells, monocytes and B-cells and color coded by protocols (**a**) or the number of detected genes per cell (**b**).

**Supplementary Fig. 15. PCA representation of human cell types using cell type markers.**

**a,b.** PCA analysis on downsampled data (20K) for HEK293T cells, monocytes and B-cells separately using the corresponding cell type's reference markers and color coded by protocols (**a**) or number of detected genes per cell (**b**).

**Supplementary Fig. 16. Gene expression correlations across 13 sc/snRNA-seq methods.**

Pearson correlation plots between protocols using gene expression of cell-type-specific signatures for HEK293T cells (**a**), monocytes (**b**) and B-cells (**c**). For a fair comparison, cells were downsampled to the same number for each method (B cells=32, Monocytes = 57, HEK293T= 55). Cells are ordered by agglomerative hierarchical clustering.

**Supplementary Fig. 17. Comparison of cell type-specific markers across protocols.**

**a.** Jaccard Indexes (JI) of B-cells, monocytes and HEK293T cell markers comparison across protocols. For each protocol, the top 100 ranked markers were considered for the JI computation. **b.** Evaluation of human marker accuracy. Protocols are compared in their ability to identify cell type-specific markers (as defined from the human reference). Jaccard Indexes are shown per cell type for each protocol (left) and their averages are displayed in relation with the clustering accuracy (right). **c.** Evaluation of mouse marker accuracy. Protocols are compared in their ability to identify cell type-specific markers (as defined from the mouse reference).

**Supplementary Fig. 18. Marker overlap across protocols.**

Overlap percentages of B-cells, monocytes and HEK293T markers across protocols considering the top 100 ranked markers.

**Supplementary Fig. 19. Data integration using Seurat.**

**a,b.** UMAP visualization of clusters after the integration of technologies for 18,034 human (**a**) and 7,902 mouse (**b**) cells. Cluster annotations are assigned on the basis of the most frequent cell type. **c,d.** Barplots showing normalized and method-corrected (integrated) expression scores in cell type specific signatures for CD4+ and CD8+ T-cells (**c**) and enterocytes 1, enterocytes 2 and intestinal stem cells (**d**). Bars are colored by method. **e.** Evaluation of dataset mixability after integration. Protocols are compared in their ability to mix with other technologies within same cell types. Barplots correspond to the mixability scores and colors are indicating the level of sequencing depths (10K and 20K), highlighting the drop of integratability at lower depth.

**Supplementary Fig. 20. Integration of human sc/snRNA-seq datasets (original sequencing reads).**

**a,b.** UMAP visualization of cells after Seurat integrations for 20,237 human sc/snRNA-seq datasets without downsampling. Cells are colored by cell type (**a**) and protocol (**b**).

**Supplementary Fig. 21. Integration of human sc/snRNA-seq.**

**a,b.** UMAP visualization of 18,034 cells after harmony (**a**) and scMerge (**b**) integrations for human sc/snRNA-seq datasets (downsampled to 20K). Cells are colored by cell type (**left**) and protocol (**right**). **c,d.** Evaluation of protocol integratability in harmony (**c**) and scMerge (**d**). Protocols are compared according to their ability to group cell types into clusters (after integration) and mix with other technologies within the same clusters. Points are colored by sc/snRNA-seq protocol.

**Supplementary Fig. 22. Integration of mouse sc/snRNA-seq downsampled datasets.**

**a,b.** UMAP visualization of 7,902 cells after harmony (**a**) and scMerge (**b**) integrations for mouse sc/snRNA-seq datasets (downsampled to 20K). Cells are colored by cell type (**left**) and protocol (**right**). **c,d.** Evaluation of protocol integratability in harmony (**c**) and scMerge (**d**). Protocols are compared according to their ability to group cell types into clusters (after integration) and mix with other technologies within the same clusters. Points are colored by sc/snRNA-seq protocol.

**Supplementary Fig. 23. Integration of human Chromium (V2) sc/snRNA-seq datasets.**

**a,b.** UMAP visualization of cells after data integration with scMerge, Seurat and harmony for human Chromium scRNA-seq (1,599 cells) and snRNA-seq (856 cells) datasets (downsampled to 20K). Cells are colored by cell type (**a**) and protocol (**b**). **c.** Evaluation of protocol integratability based on the clustering accuracy after merging (separately for the three integration tools). The boxplots display the minimum, 1st, 2nd, 3rd quantiles and maximum clustering accuracies obtained for the different cell types. For all three alignment methods, Seurat was applied to perform clustering and UMAP after the protocol correction, in order to minimize the variability related to the downstream analysis. Results were consistent across tools with Chromium (single-cell) showing the highest clustering accuracy and Chromium (single-nuclei) displaying higher variability. While B-cells, monocytes and T-cells were robustly clustered, NK cells were grouped with CD8+ T-cells in Chromium scRNA-seq. CD14+ monocytes, CD8+ T-cells and HEK293T cells were poorly clustered in Chromium snRNA-seq.

**Supplementary Fig. 24. Integration of mouse Chromium (V2) sc/snRNA-seq datasets.**

**a,b.** UMAP visualization of 7,902 cells after data integration with scMerge, Seurat and harmony for mouse Chromium scRNA-seq and snRNA-seq datasets (downsampling to 20K). Cells are colored by cell type (**a**) and protocol (**b**). **c.** Evaluation of protocol integratability based on the clustering accuracy after merging for the three integration tools. Boxplots displaying the minimum, 1st, 2nd, 3rd quantiles and maximum clustering accuracies obtained for the different cell types. For all three alignment methods, Seurat was applied to perform clustering and UMAP after the protocol correction, in order to minimize the variability related to the downstream analysis. After integration, the clustering accuracy was largely conserved. Of note, transit amplifying cells were divided into two main cluster pointing to a heterogeneity between the protocols and potentially due to the decreased frequency of highly abundant ribosomal genes when sampling from the nucleus.

**Supplementary Fig. 25. Comparison of mappability scores across technologies.**

Boxplots displaying minimum, 1st, 2nd, 3rd quantiles and maximum probabilities values (scores) obtained by *matchScore2* in classifying most common cell types in human (**a,b**) and mouse (**c**) samples. B-cells, HEK293T cells and CD14+ monocytes are shown with data downsampled to 20K (**a**) and 10K (**b**) sequencing reads.

**Supplementary Fig. 26. Comparing column and bead purification in Quartz-seq2.**

**a.** Sequential processing steps from poly-A tailed RNA to sequencing-ready libraries common to most sc/snRNA-seq protocols. **b.** Experimental design to systematically compare the yield of amplified cDNA using column or bead cDNA purifications, at different bead concentrations. **c.** Relative amount of amplified cDNA using different concentrations of beads. **d.** Comparing the yield of amplified cDNA using column and bead purification.

**Supplementary Fig. 27. FACS sample processing strategy.**

Representative FACS plot (BD Aria III) displaying sample composition and viability statistics for the HCA reference sample.

**Supplementary Fig. 28. Human reference signature scores for plate-based protocols.**

Boxplots comparing the distribution of B cell, monocyte and HEK293T signature scores across the different human cell types. For each cell, a score is computed by combining z-scores of genes in each signatures.

**Supplementary Fig. 29. Human reference signature scores for microfluidic-based protocols.**

Boxplots comparing the distribution of B cell, monocyte and HEK293T signature scores across the different human cell types. For each cell, a score is computed by combining z-scores of genes in each signatures.

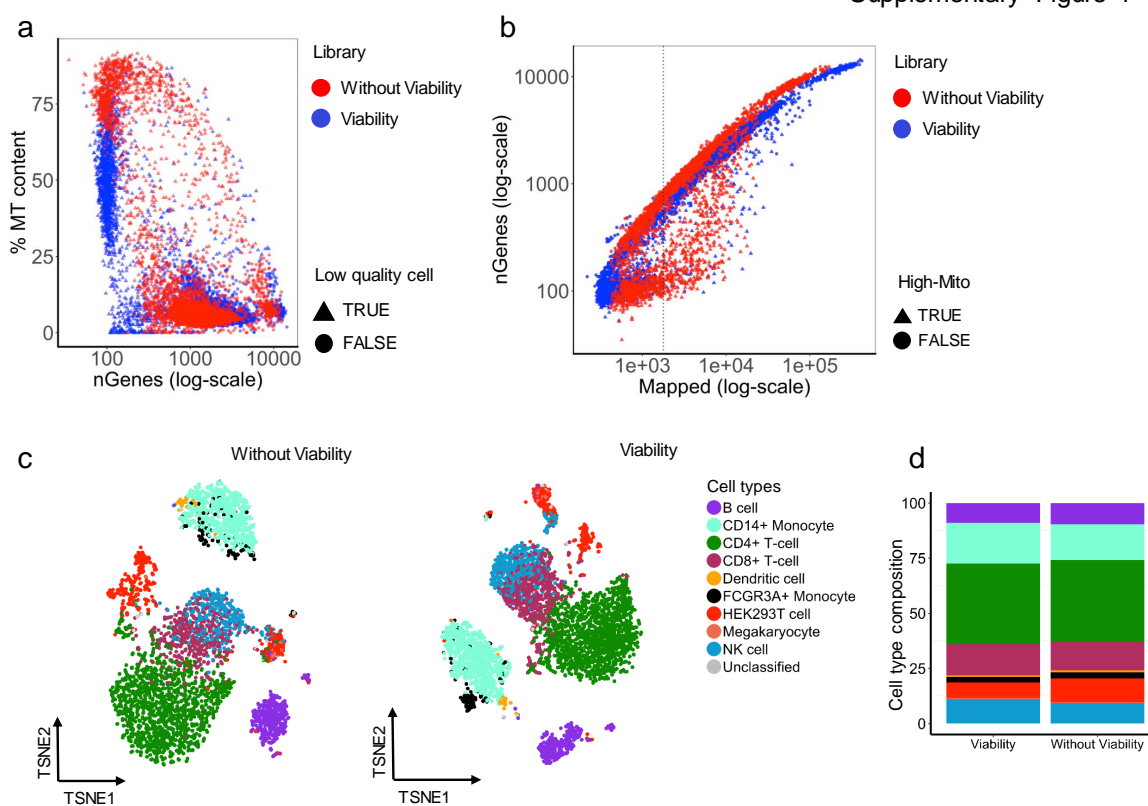
**Supplementary Fig. 30. Merging of human and mouse sc/snRNA-seq datasets.**

**a,b.** T-SNE (left) and UMAP (right) visualization of 18,034 cells after the datasets were combined and normalized by library size. Cells are colored by cell type (**a**) and protocol (**b**), showing a strong protocol-specific distribution.

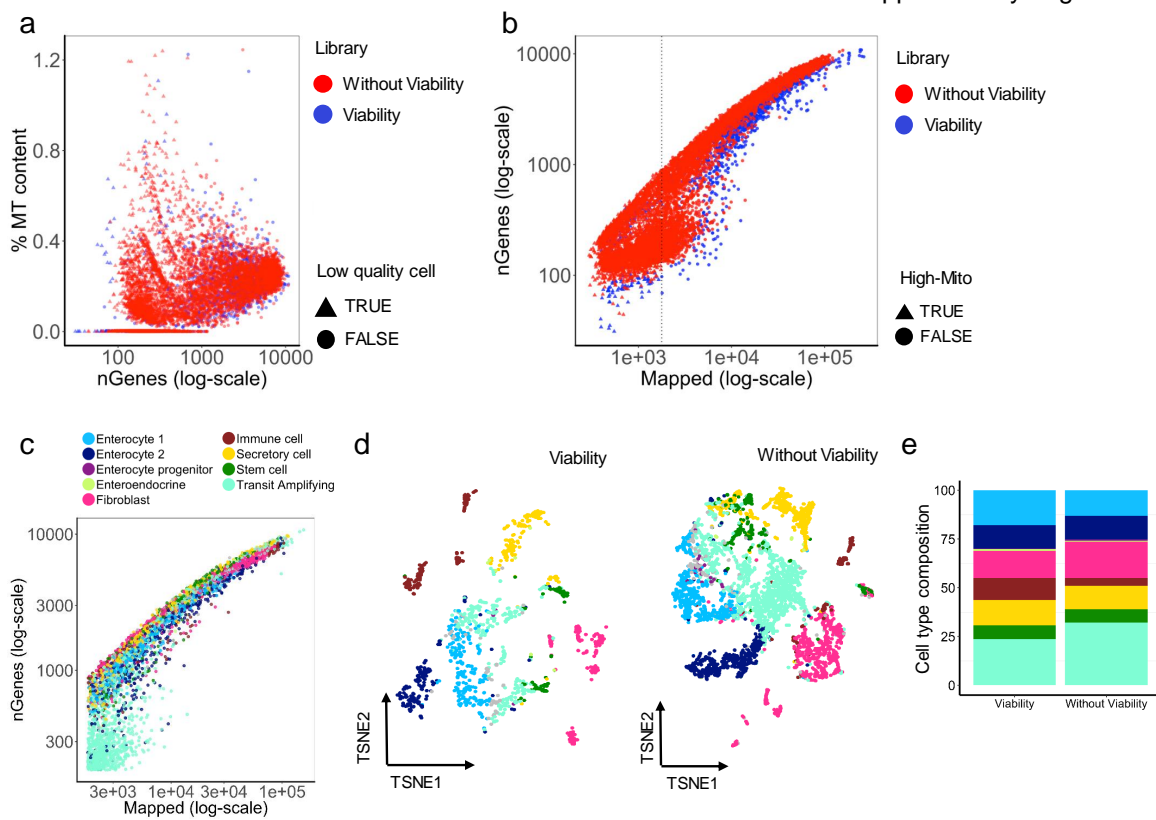
**Supplementary Fig. 31. Protocol performance with Chromium or inDrops as reference dataset.**

**a.** Mappability comparison assigning the human Chromium or inDrop datasets as reference. High similarity in the ranking of mappability for B-cells, monocytes and HEK293T cells. **b.** Similar overall performance despite the reduced dataset size of the inDrop reference. **c.** Comparison of the protocol ranking to detect cell type-specific marker expression levels (using Chromium or inDrop as reference datasets). Scaled values of the averaged expression levels (data downsampled to 20K) between B-cells, monocytes and HEK293T cells are displayed.

Supplementary Figure 1

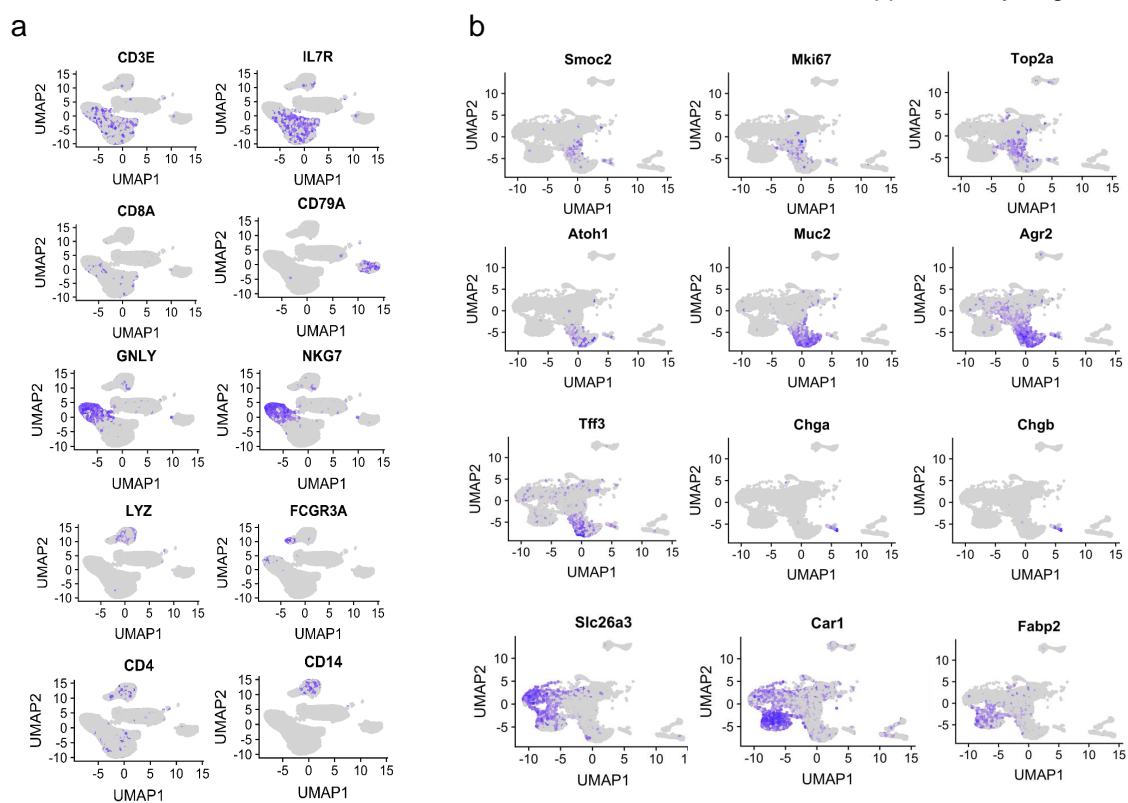


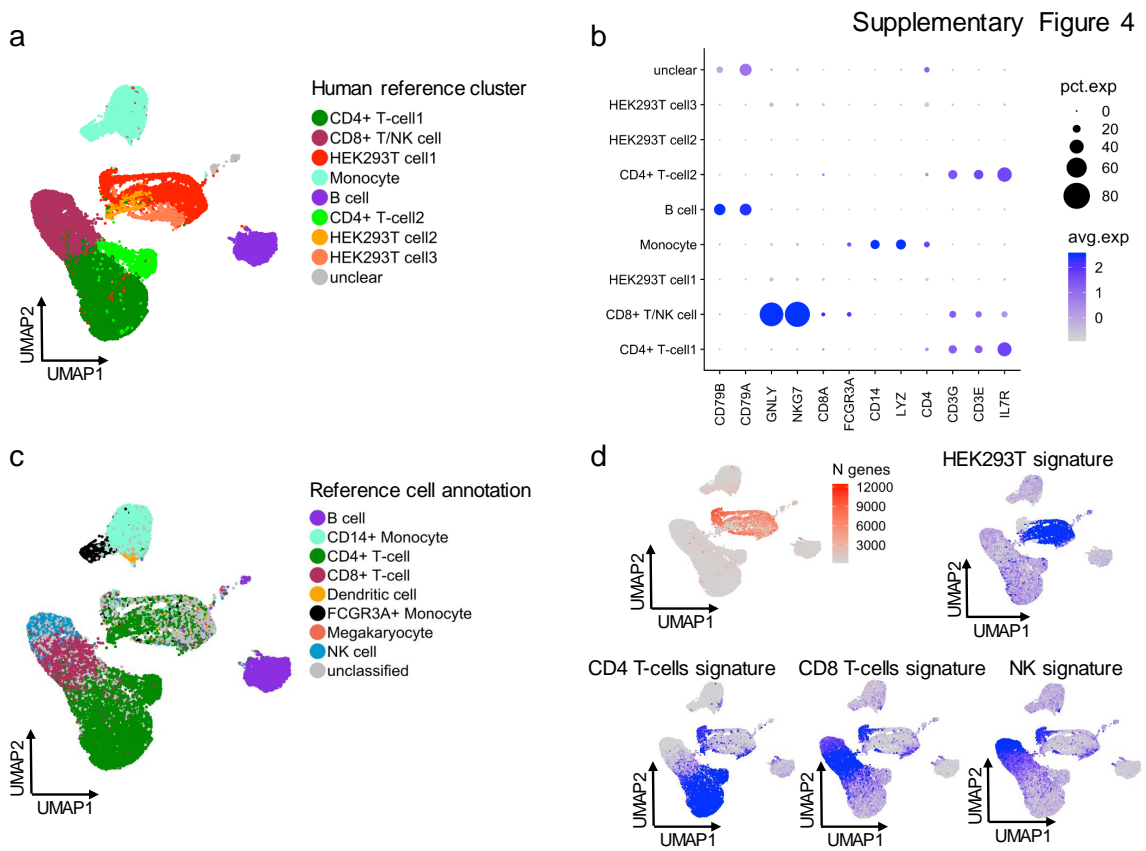
Supplementary Figure 2

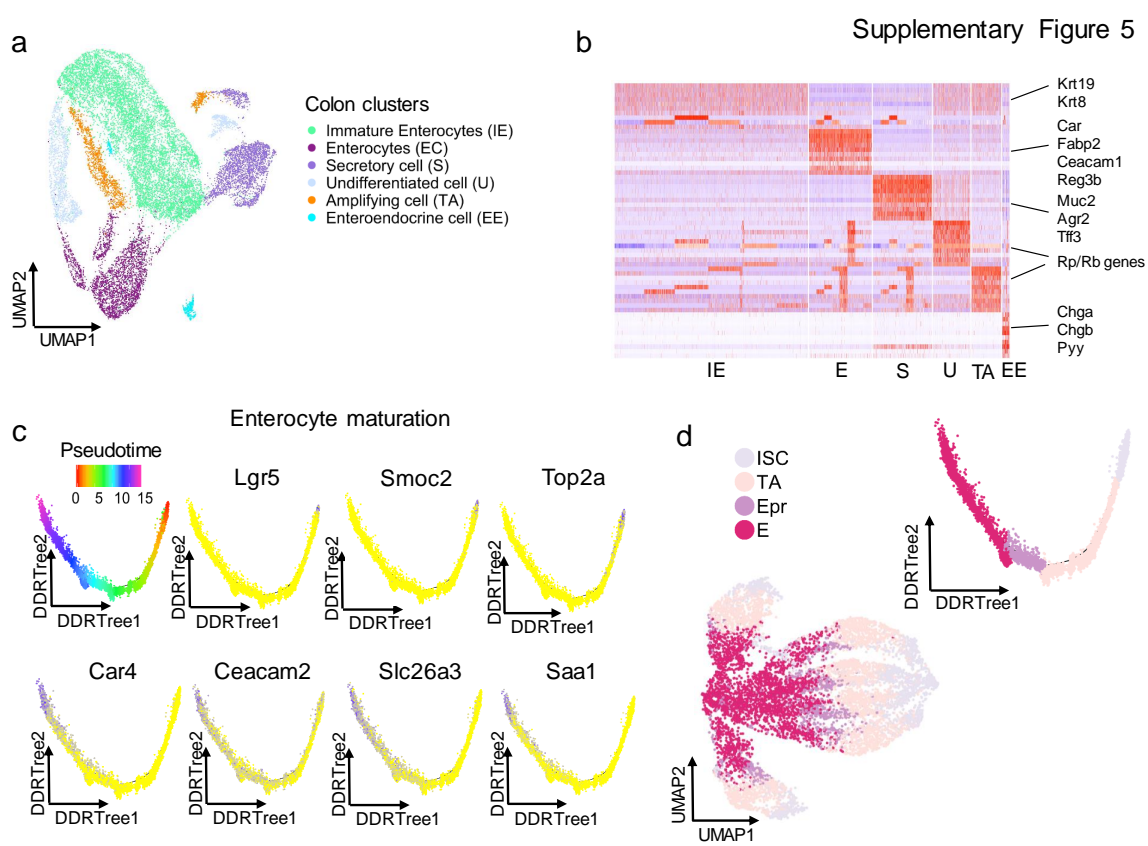




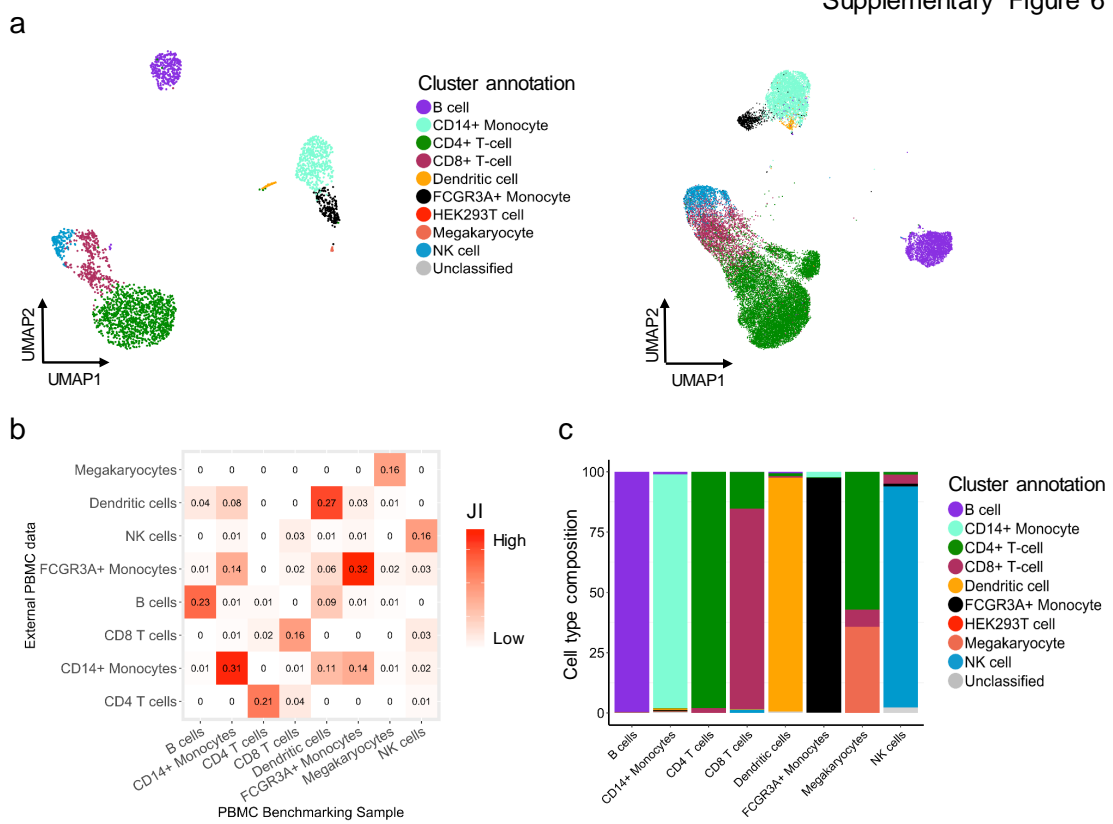
Supplementary Figure 3



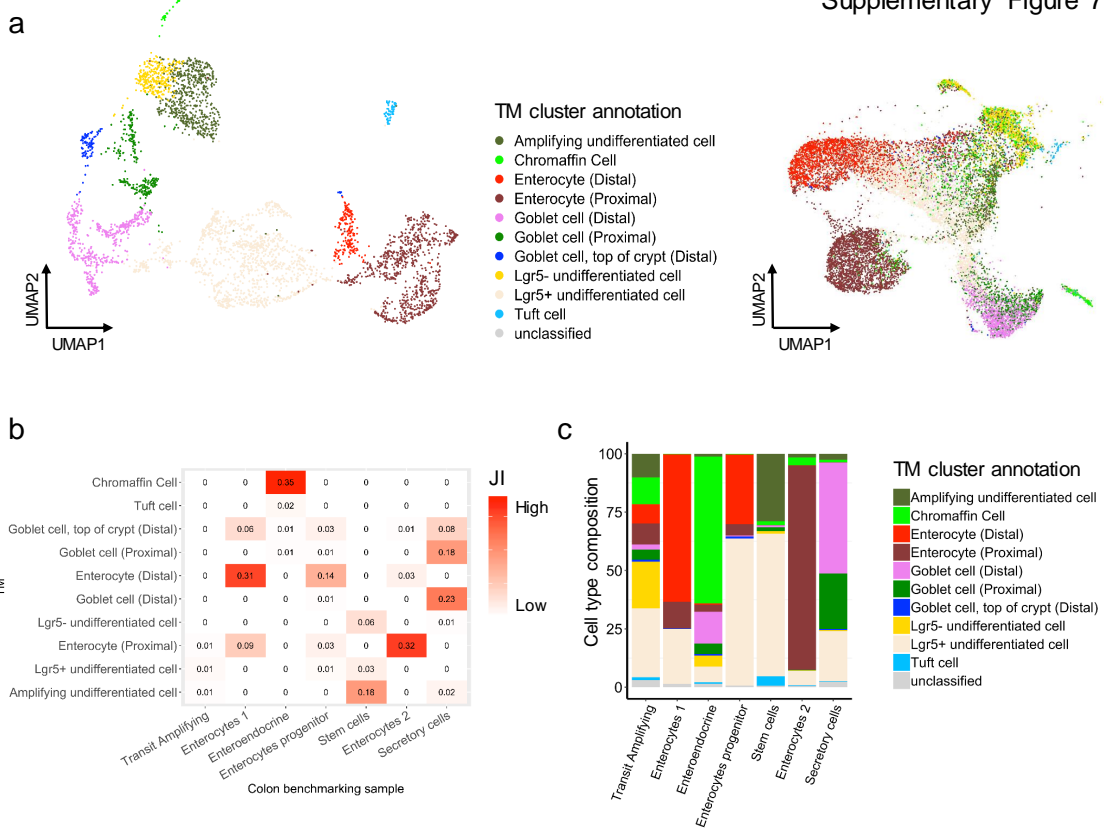


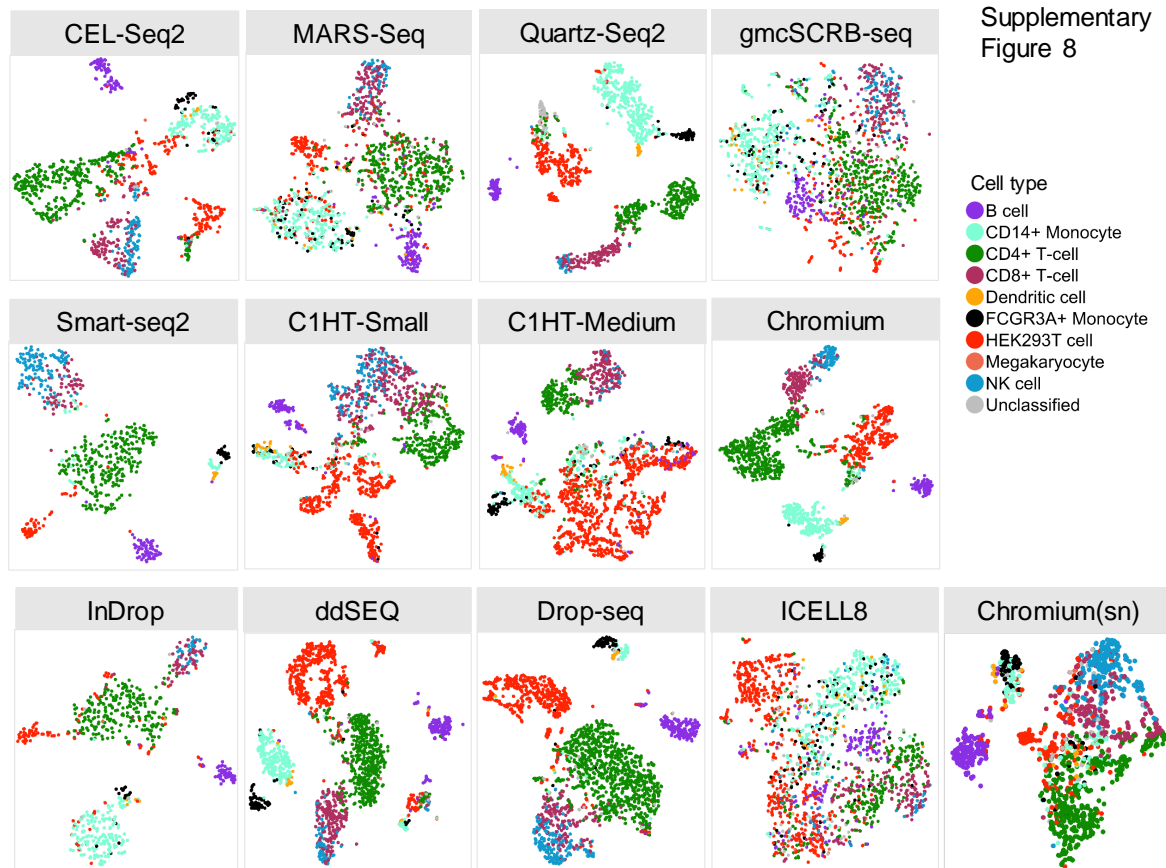


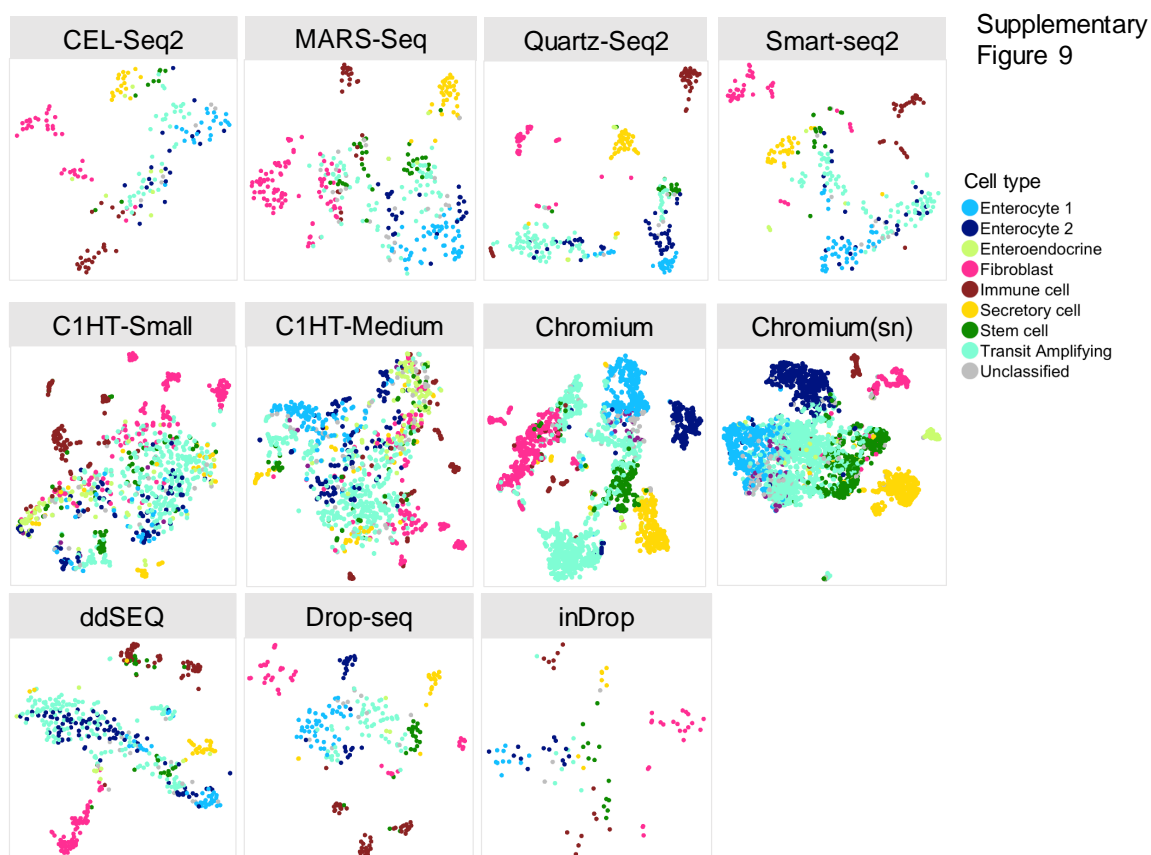
Supplementary Figure 6



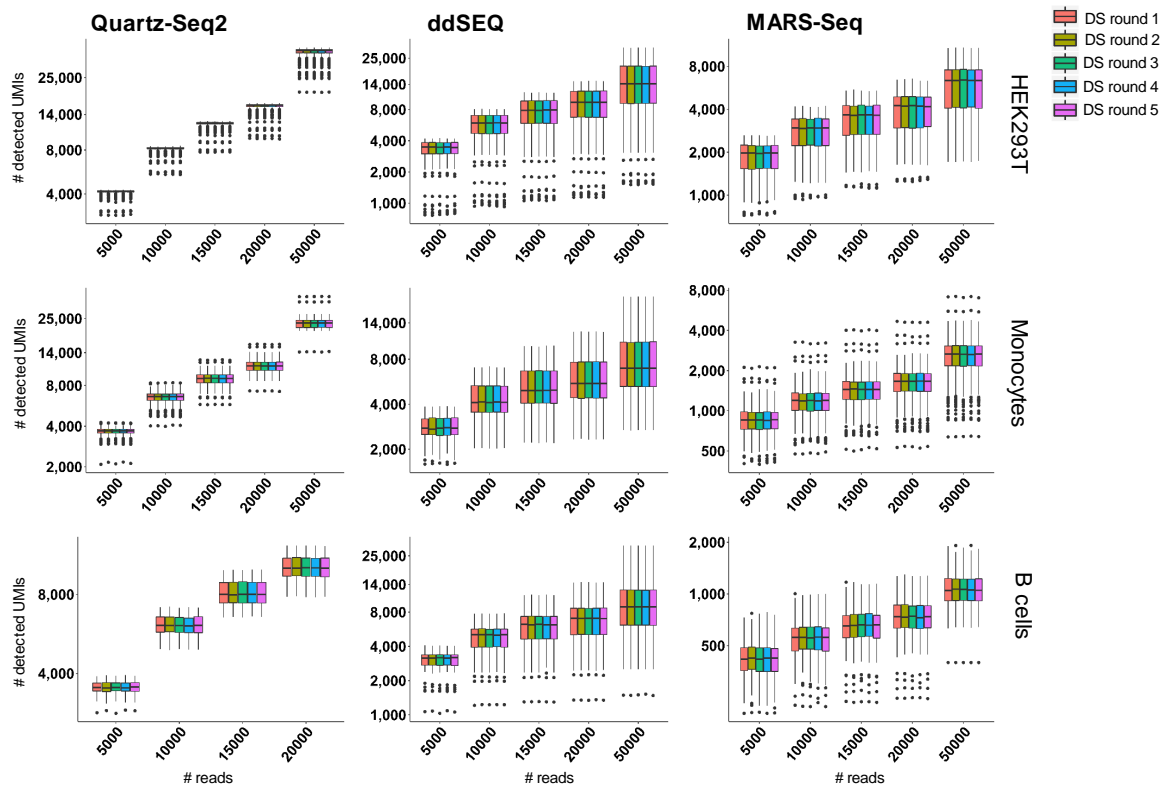
Supplementary Figure 7





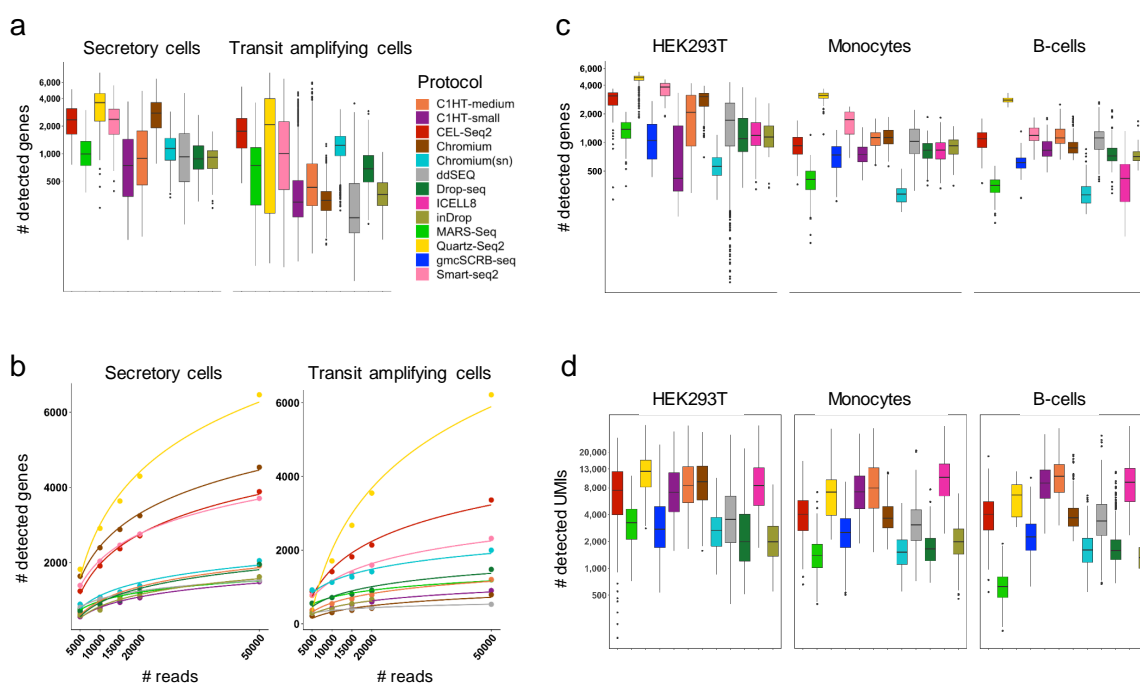


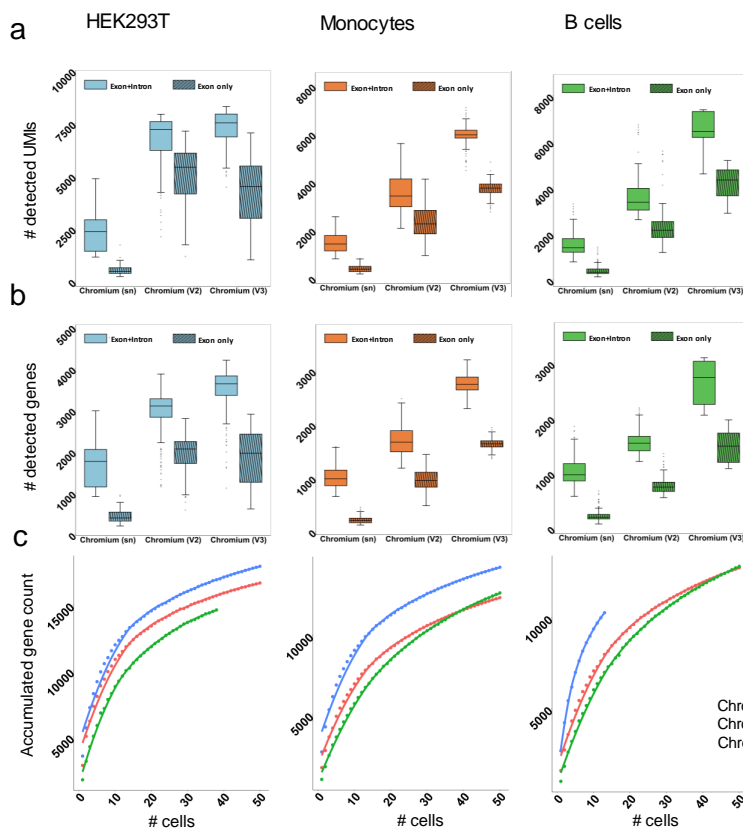
Supplementary Figure 10



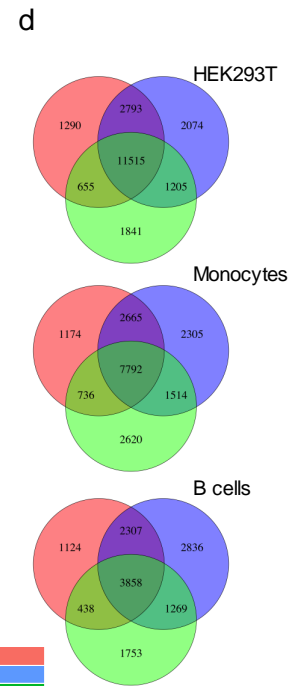


Supplementary Figure 11

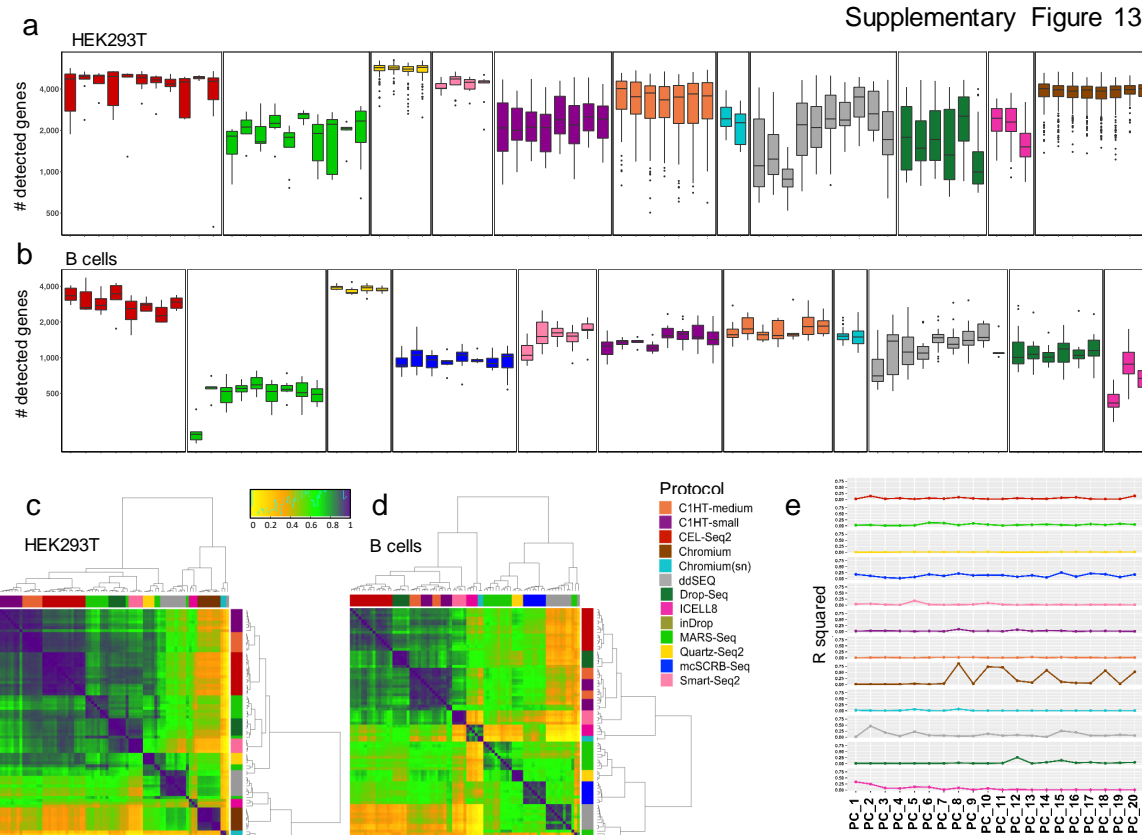




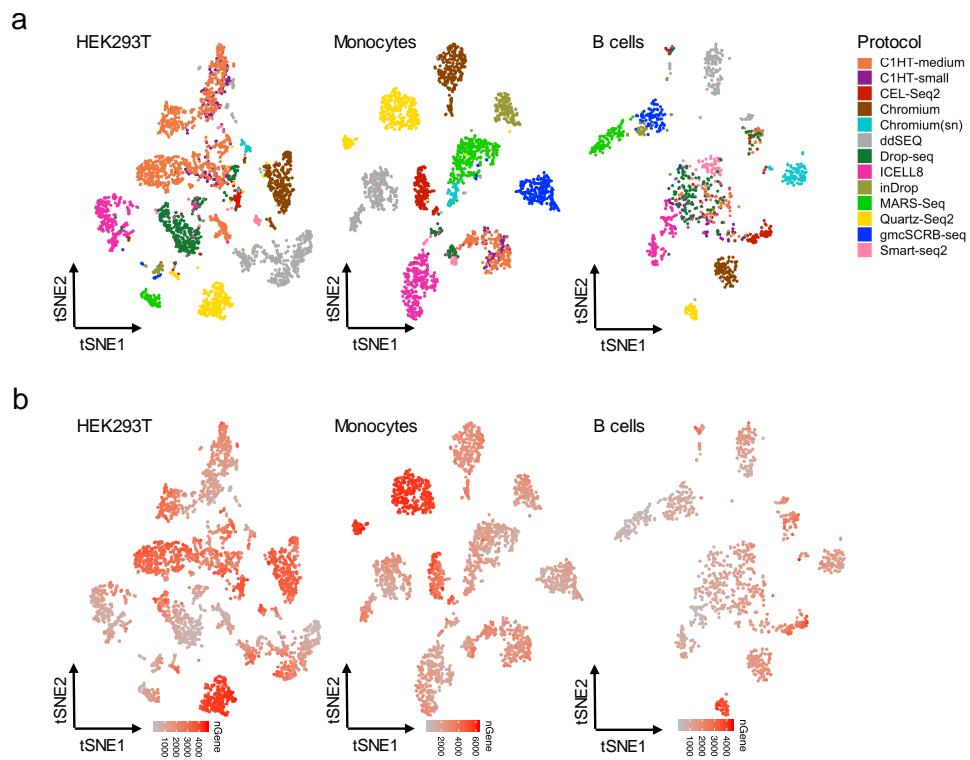
Supplementary Figure 12



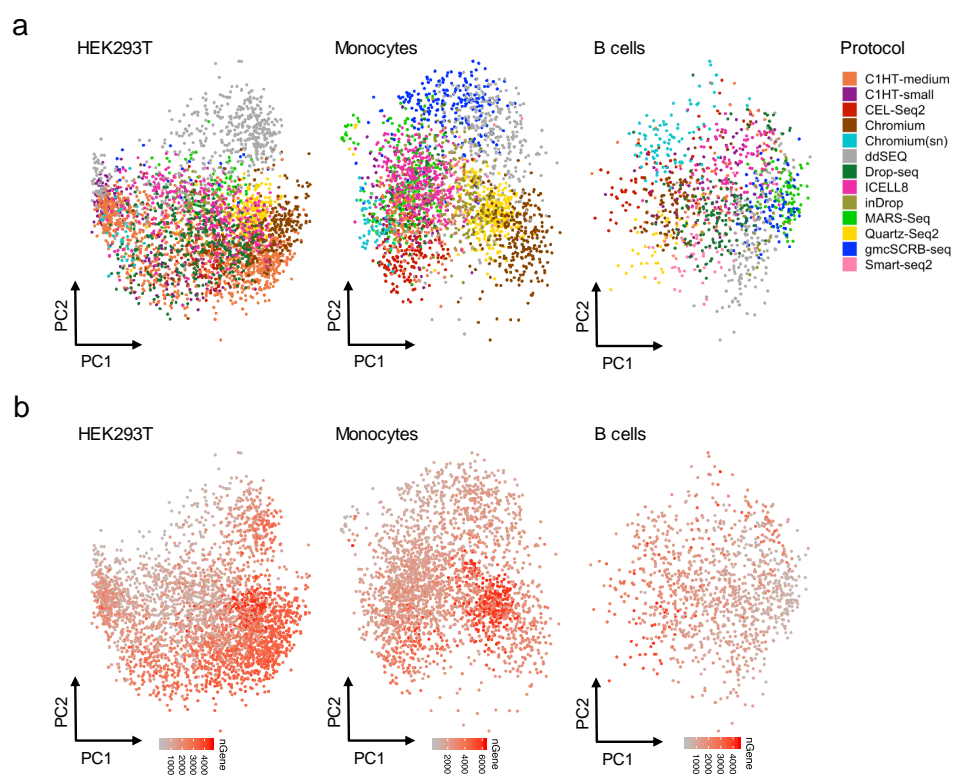
Supplementary Figure 13



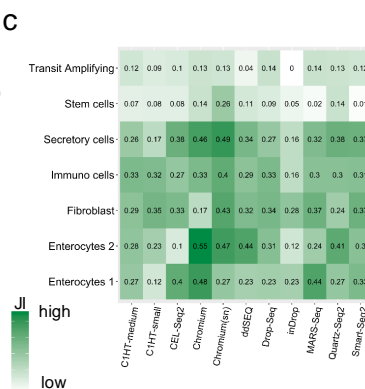
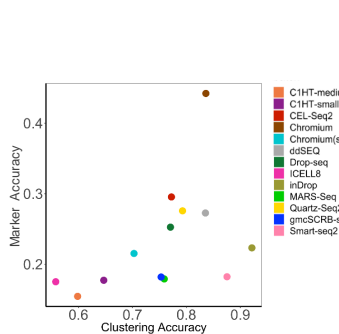
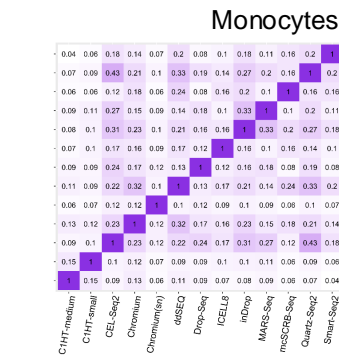
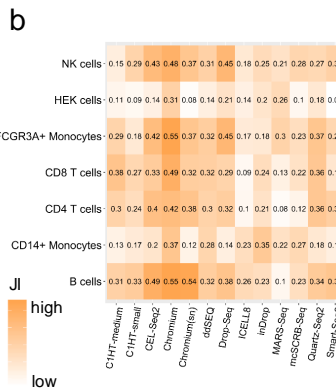
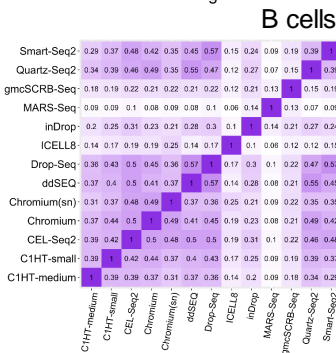
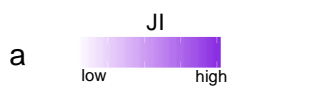
Supplementary Figure 14



Supplementary Figure 15

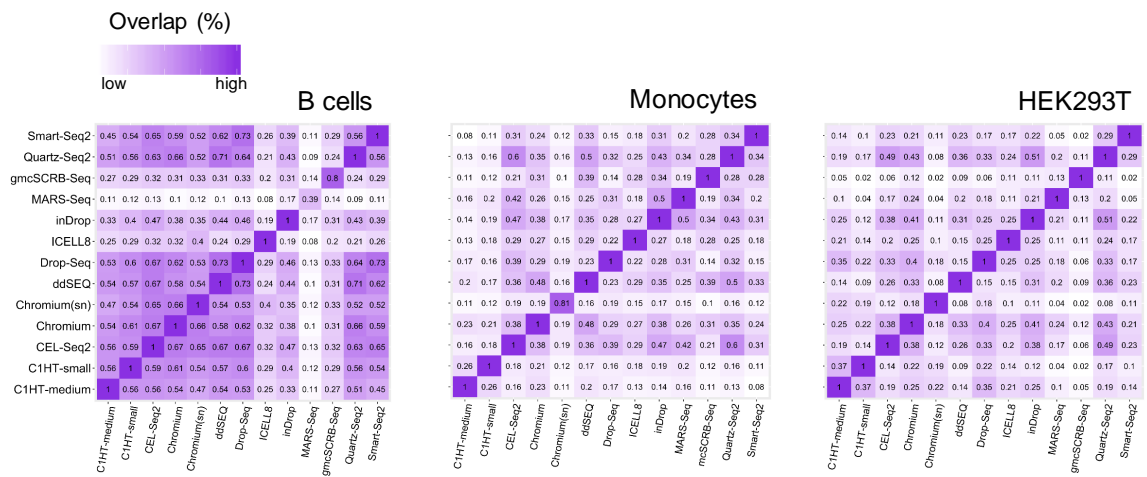




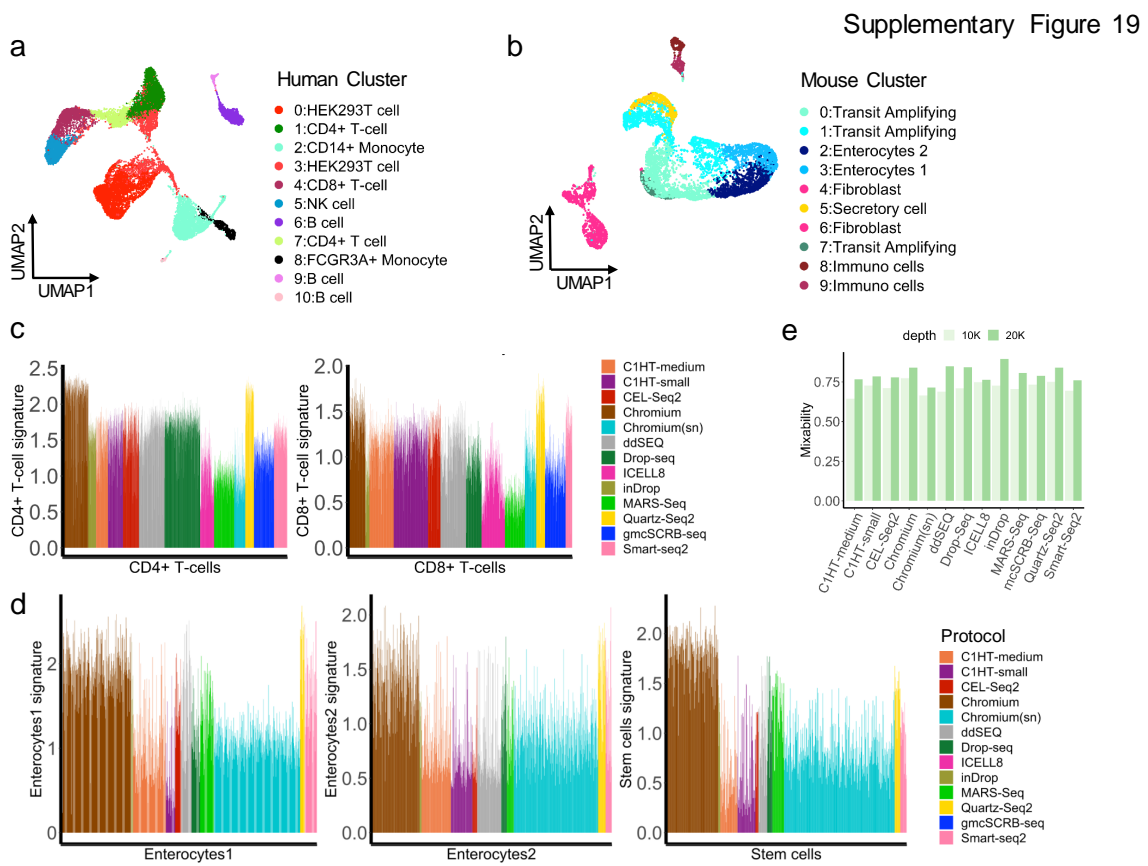


Supplementary Figure 17

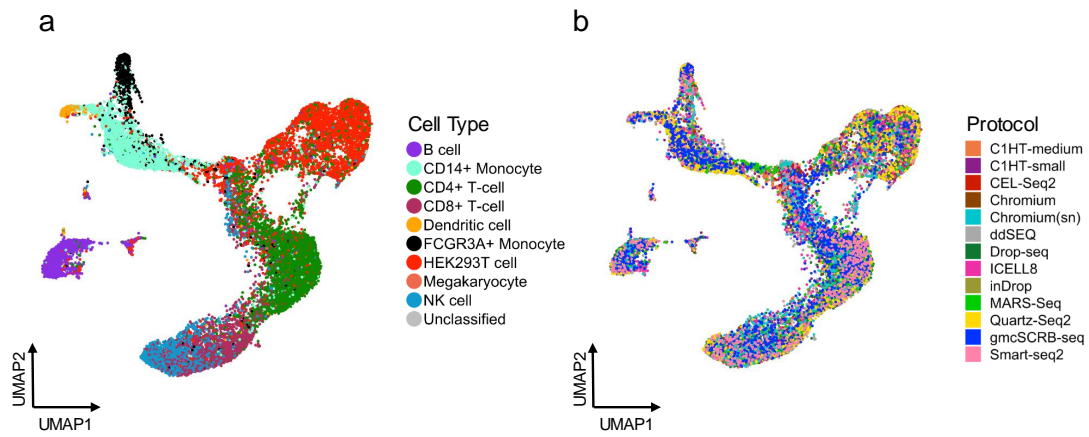
Supplementary Figure 18

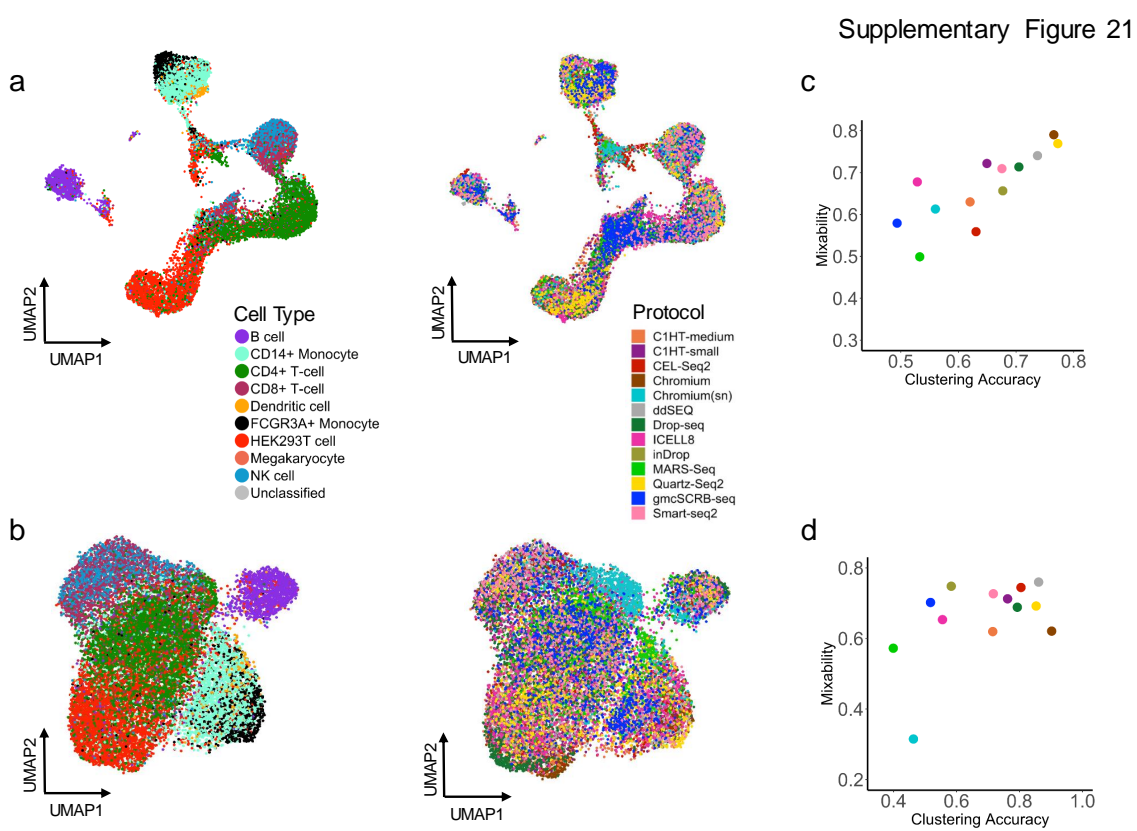


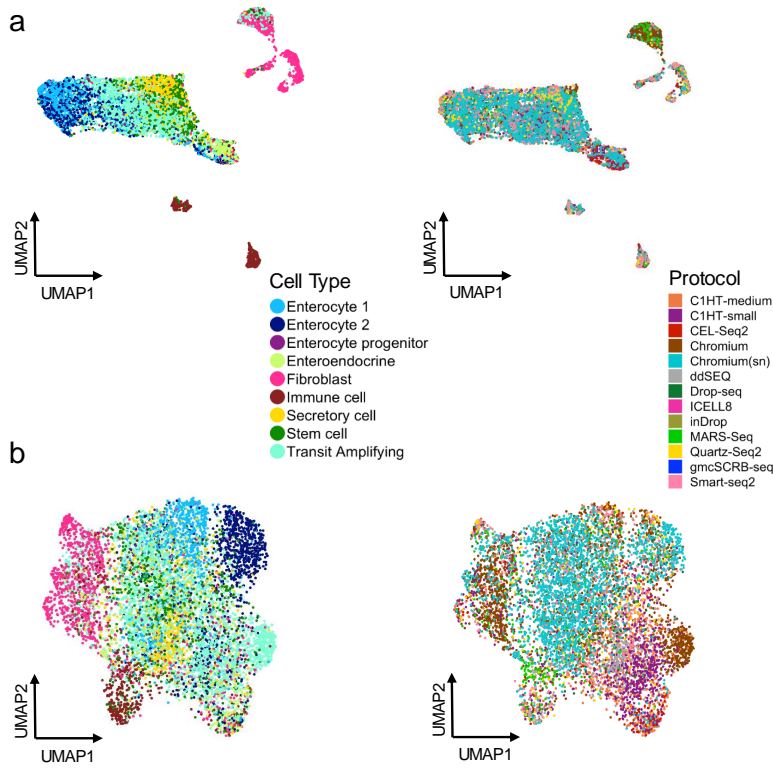




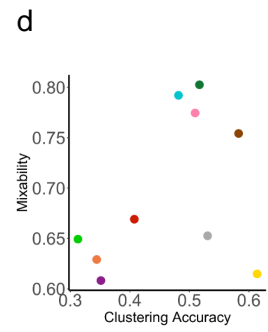
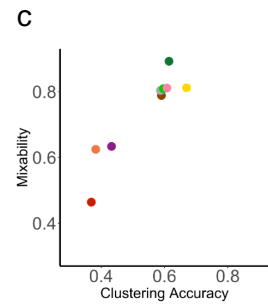
Supplementary Figure 20

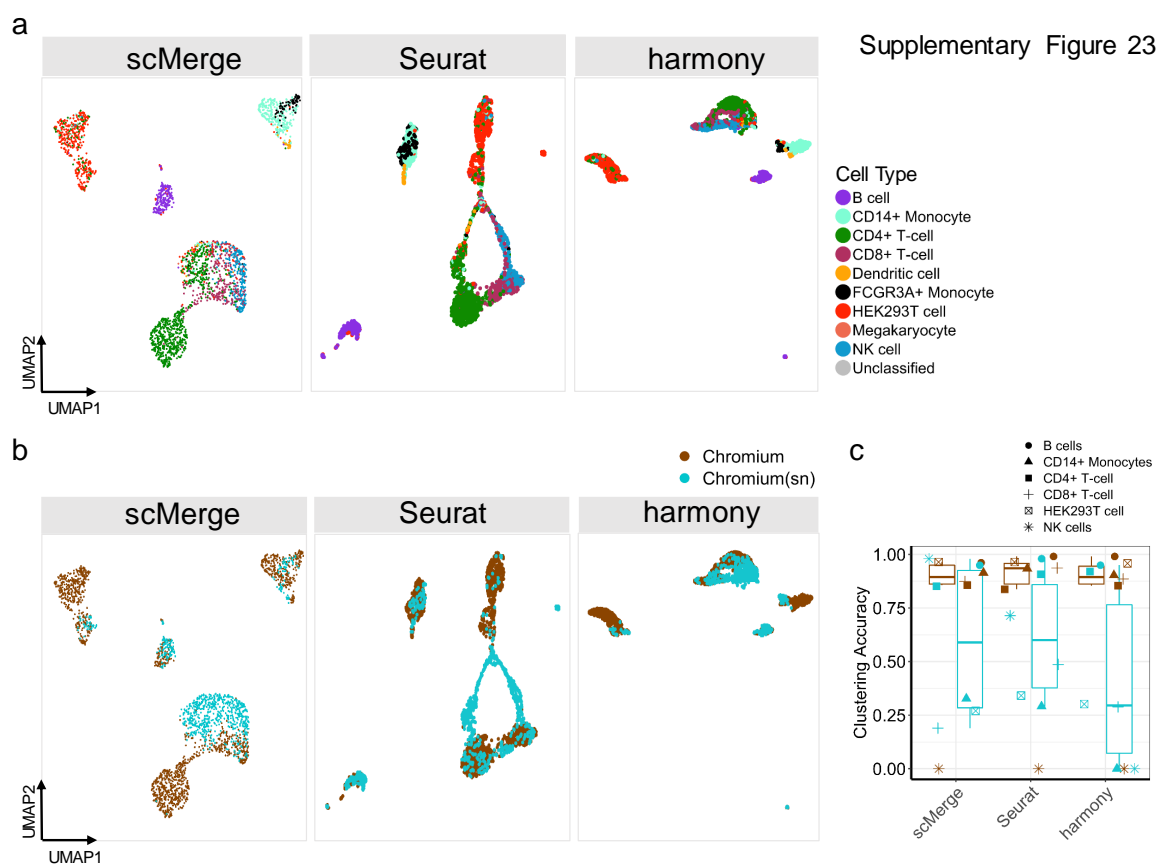


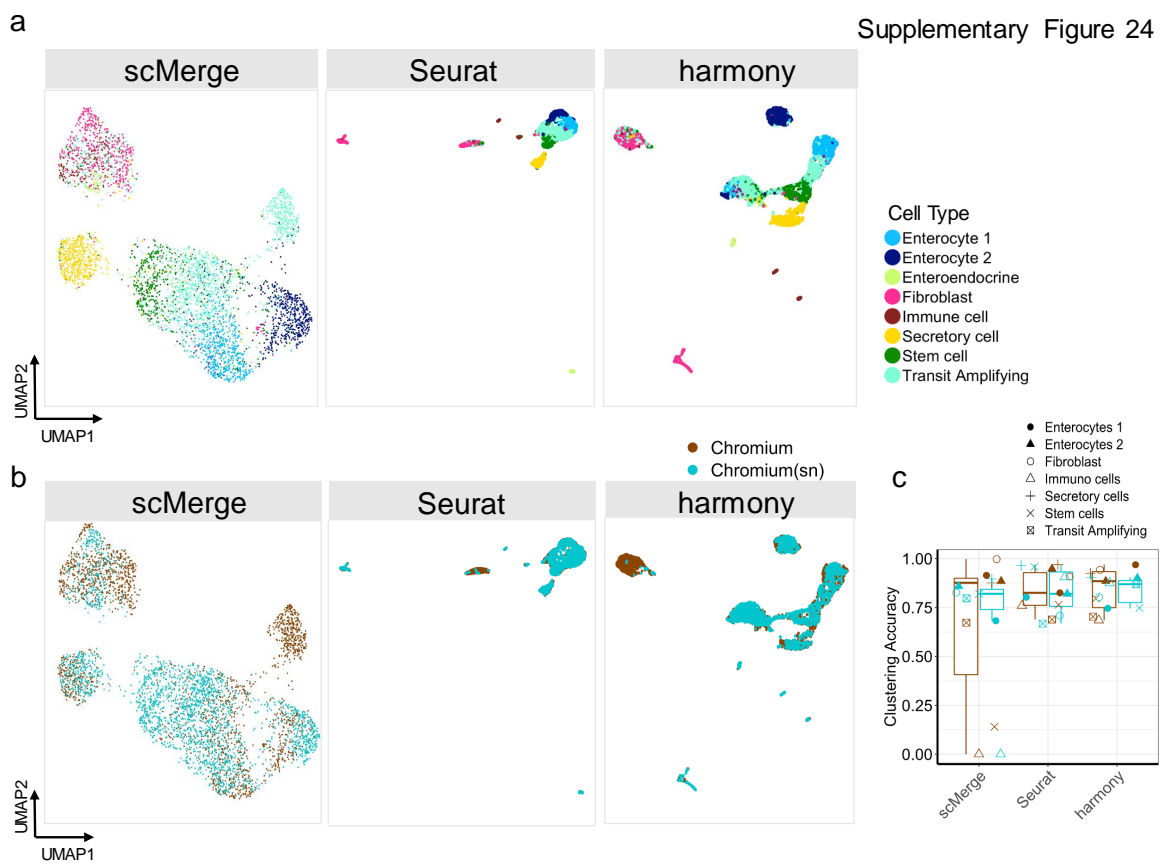




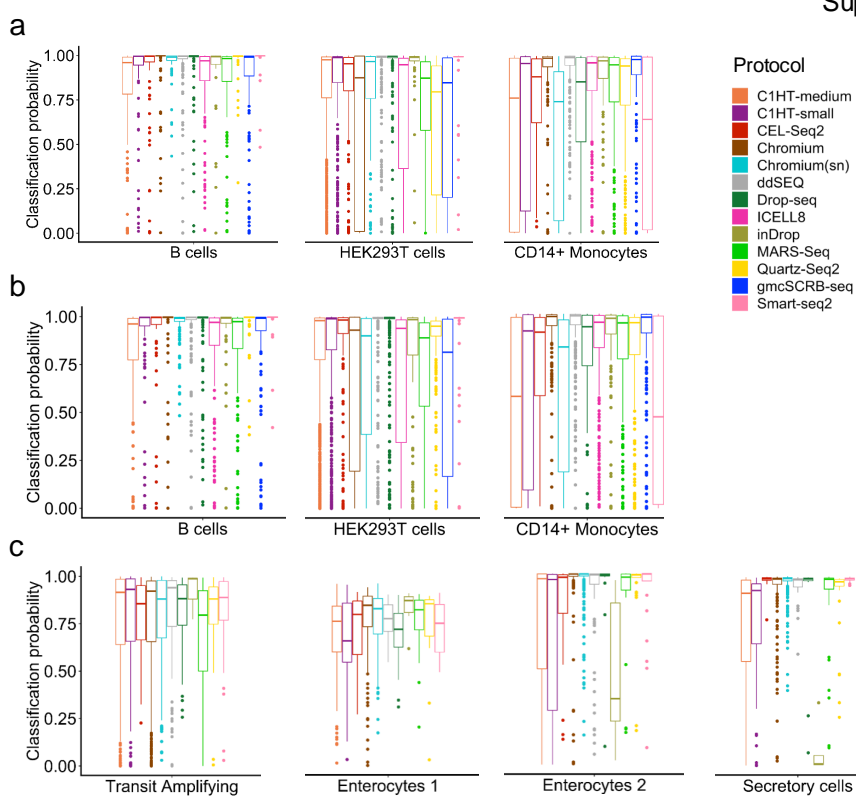
Supplementary Figure 22



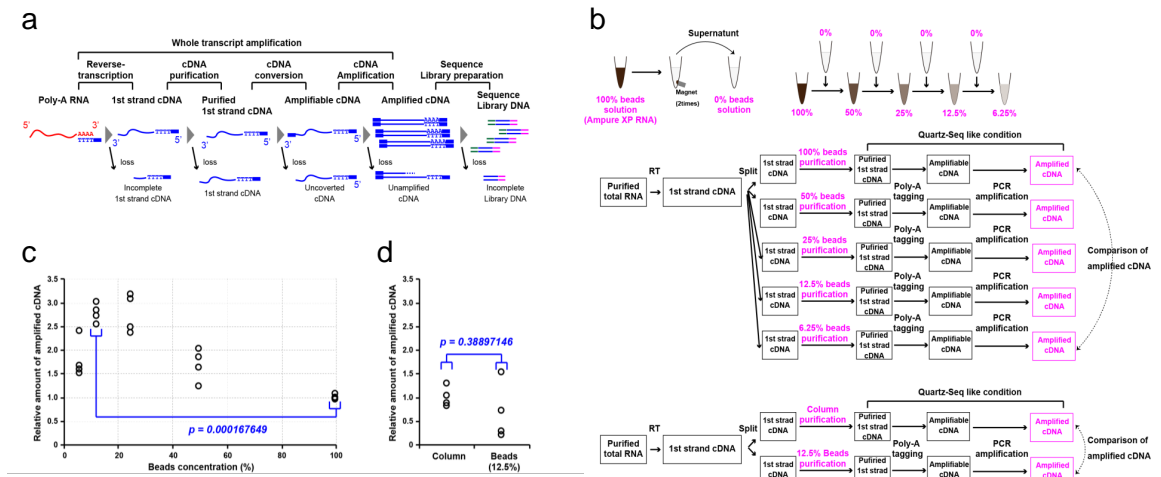




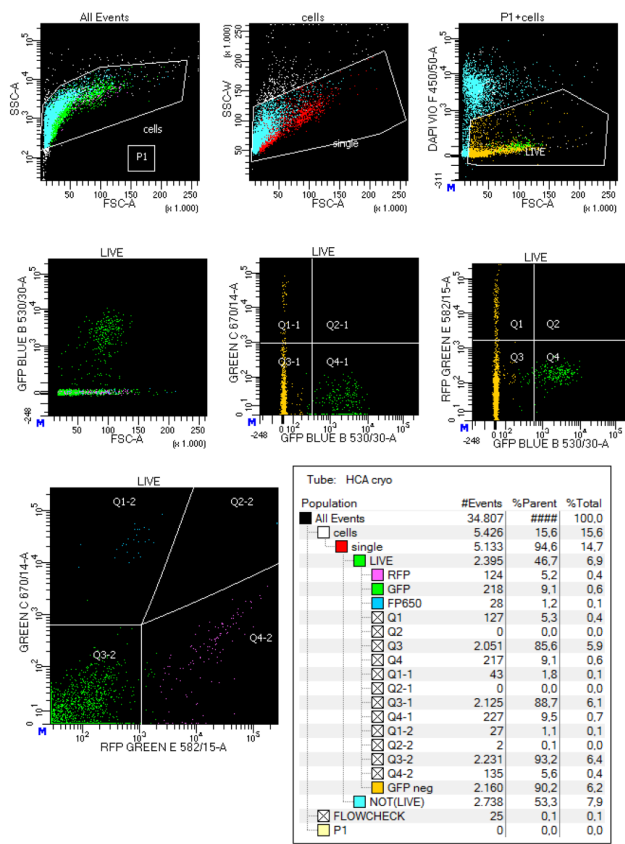
Supplementary Figure 25



Supplementary Figure 26

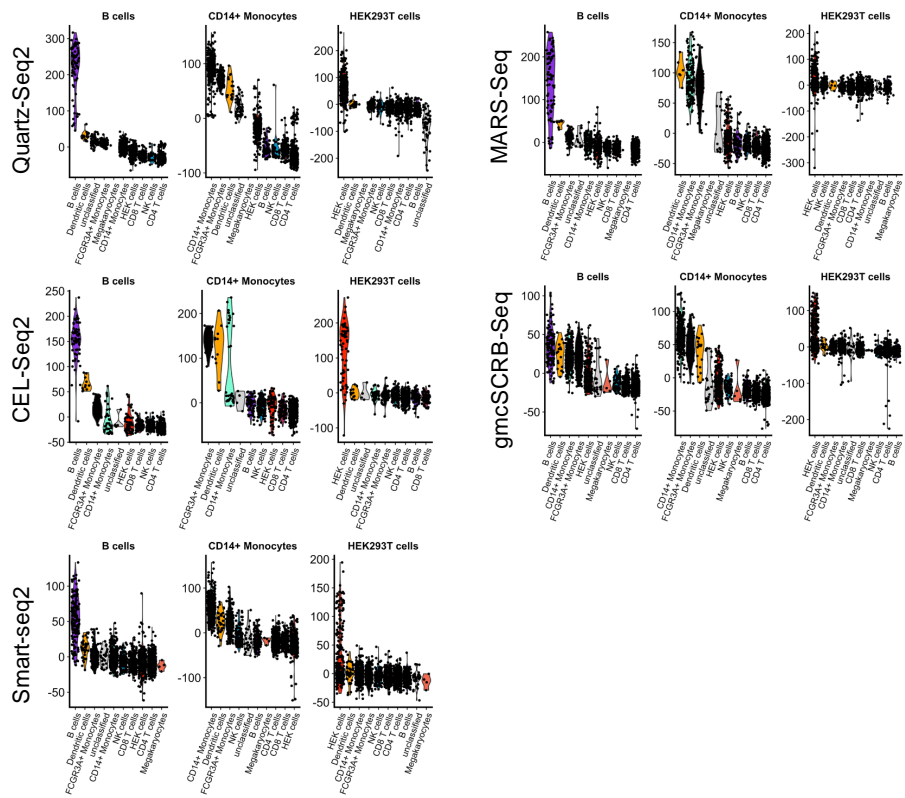




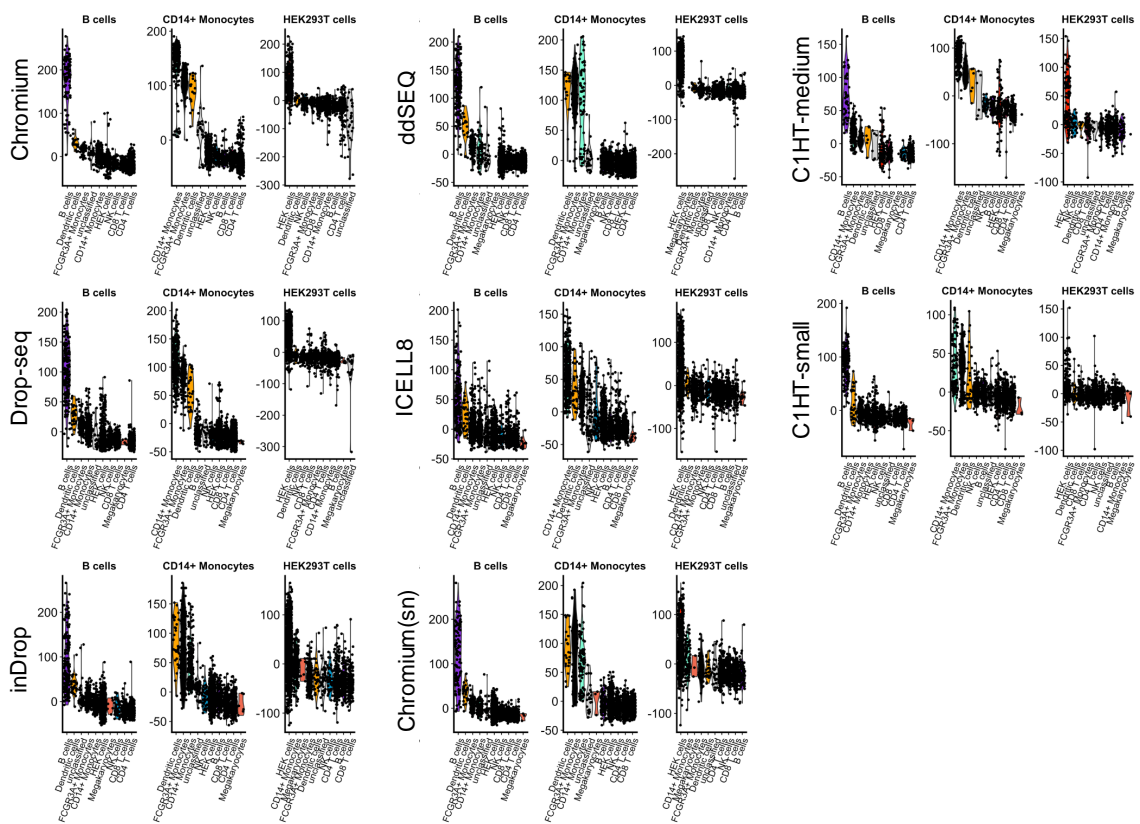


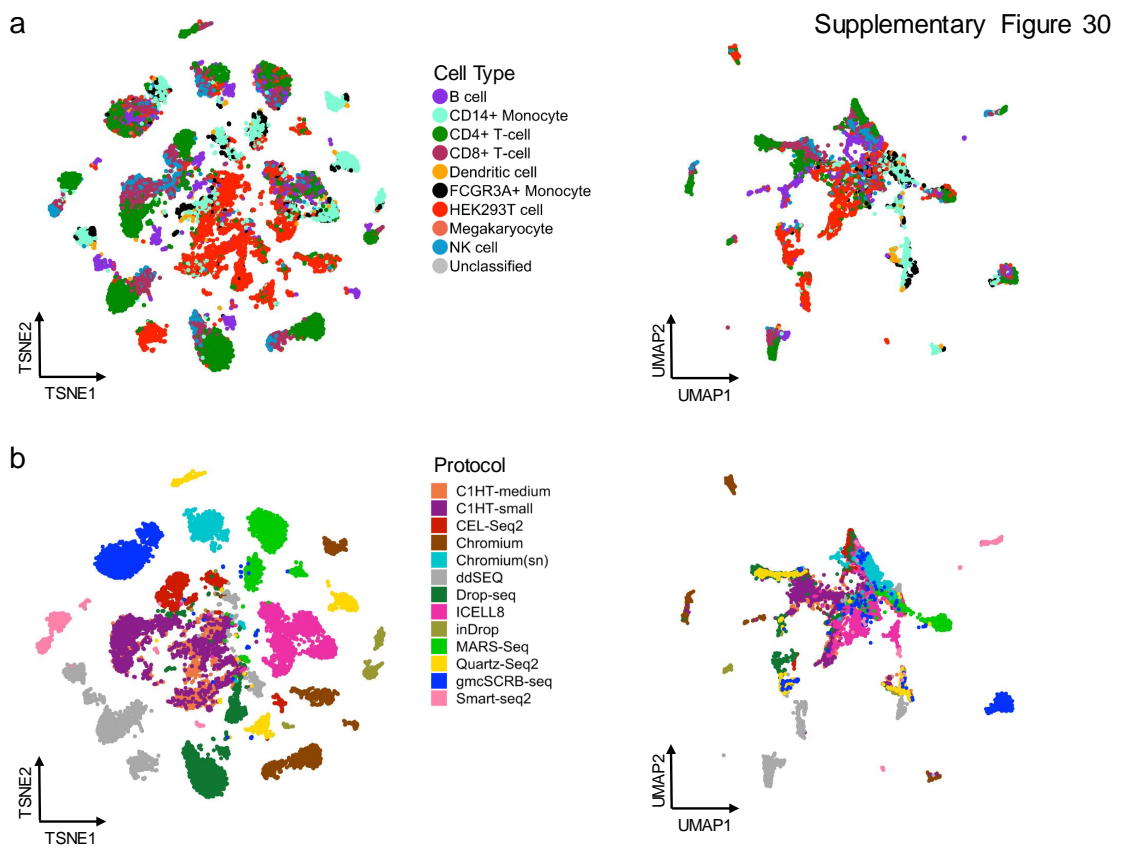
Supplementary Figure 27

Supplementary Figure 28

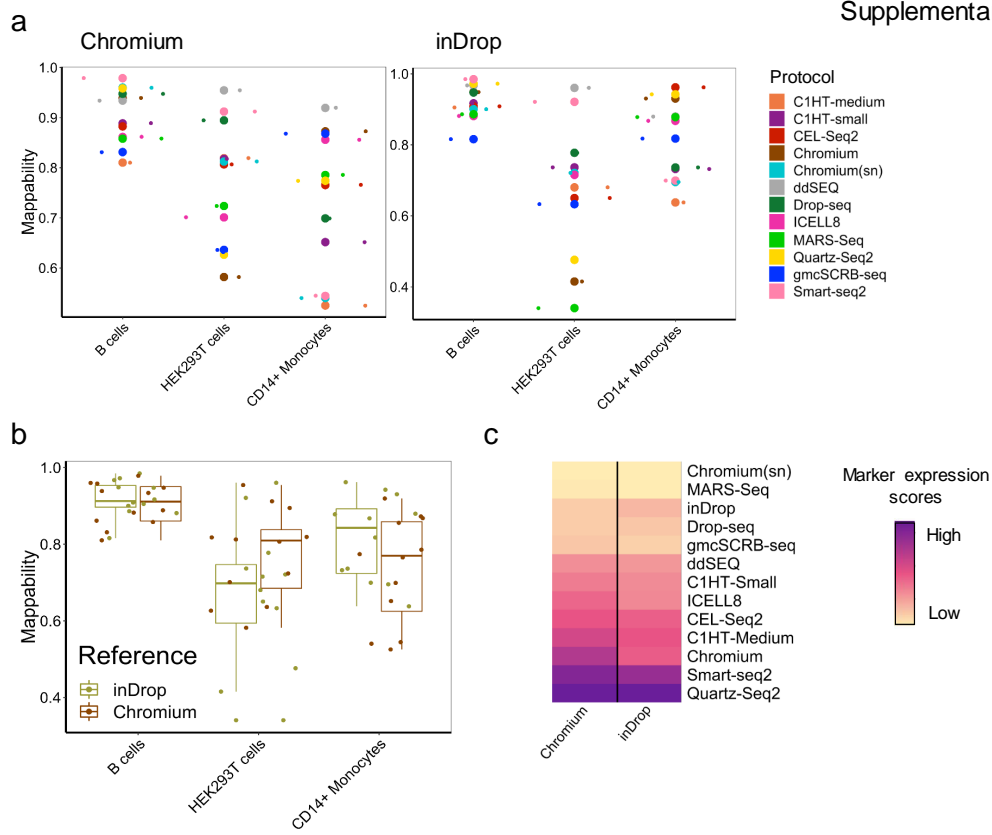


Supplementary Figure 29





Supplementary Figure 31





## 5.3 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.

## METHOD

## Open Access

# Prime-seq, efficient and powerful bulk RNA sequencing



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

\*Correspondence:

enard@bio.lmu.de

<sup>†</sup>Aleksandar Janjic and Lucas E. Wange contributed equally to this work.

<sup>1</sup>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,

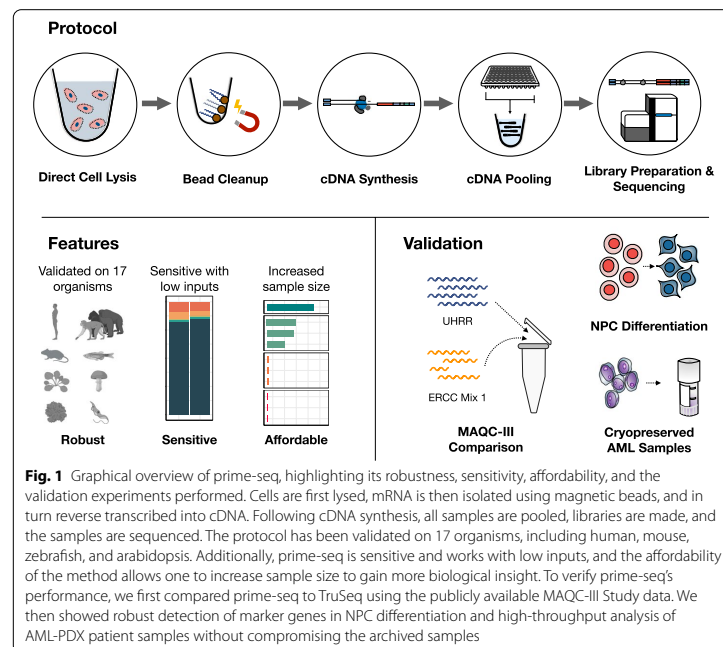
and/or setup costs can prevent their implementation. Accordingly, further developments of bulk RNA-seq protocols are still useful.

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.

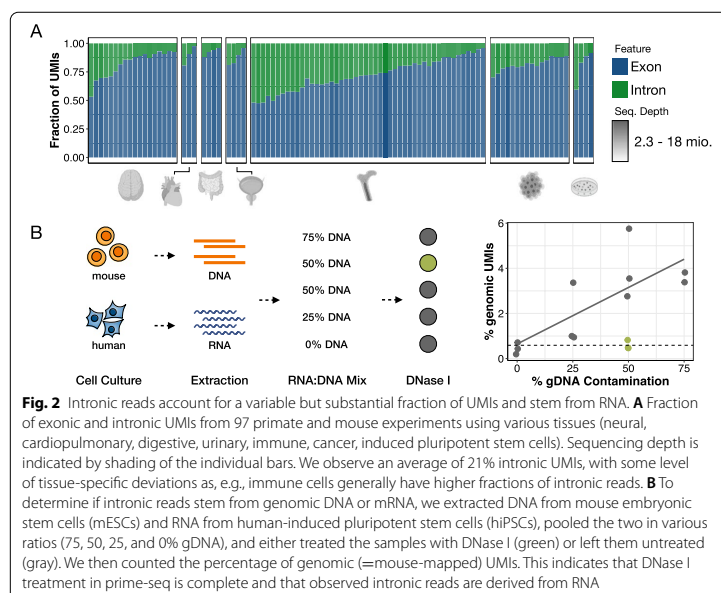
## Results

### Development of the prime-seq protocol

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



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



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.

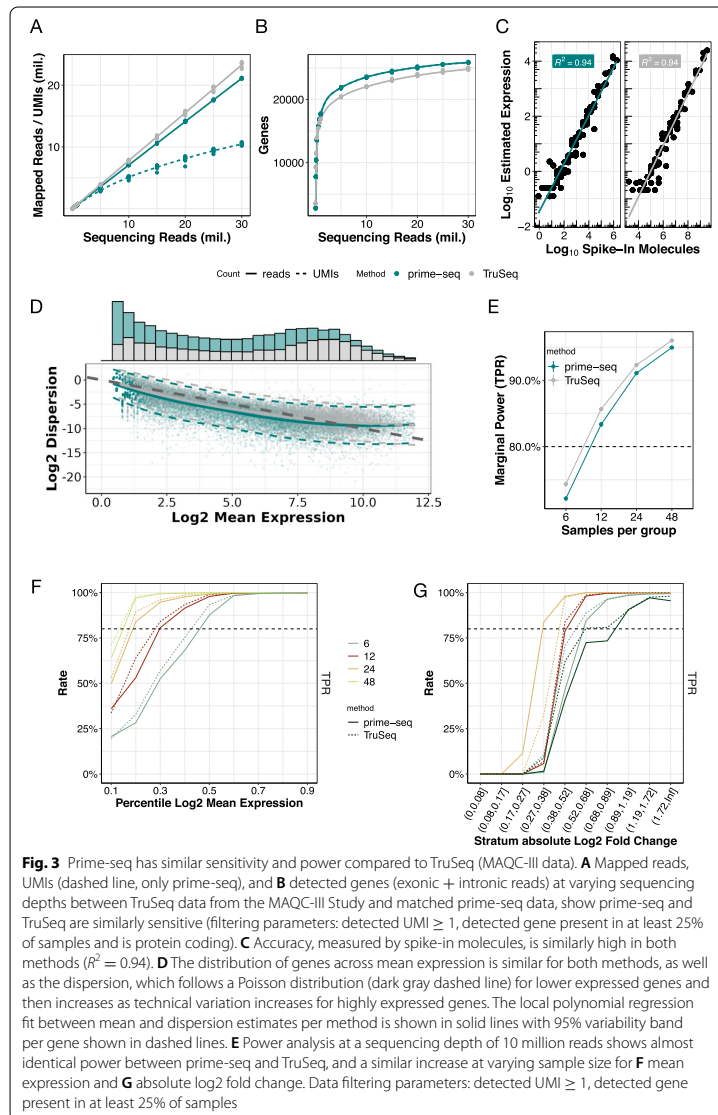
#### Prime-seq performs as well as TruSeq

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



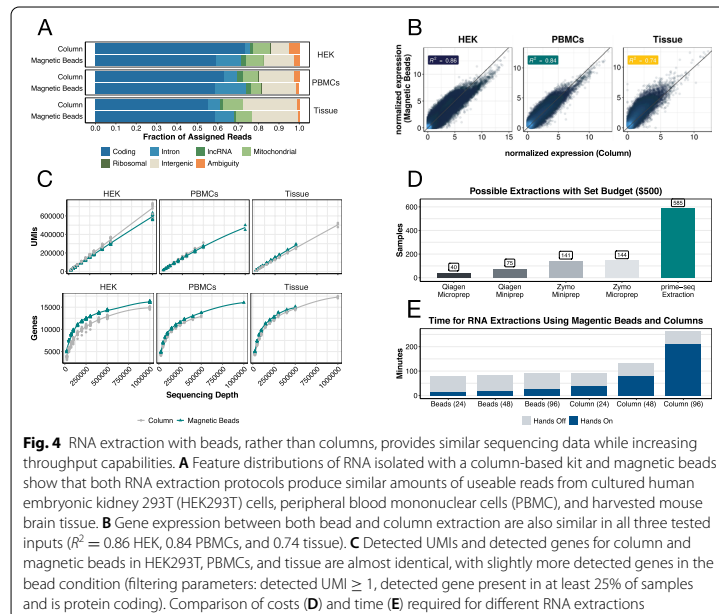
data on sampling (Poisson expectation) at low expression levels and an increasing influence of the additional technical variation at higher expression levels (Fig. 3D). Prime-seq shows a slightly lower variance for medium expression levels where most genes are expressed (Fig. 3D). To quantify to what 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).



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.

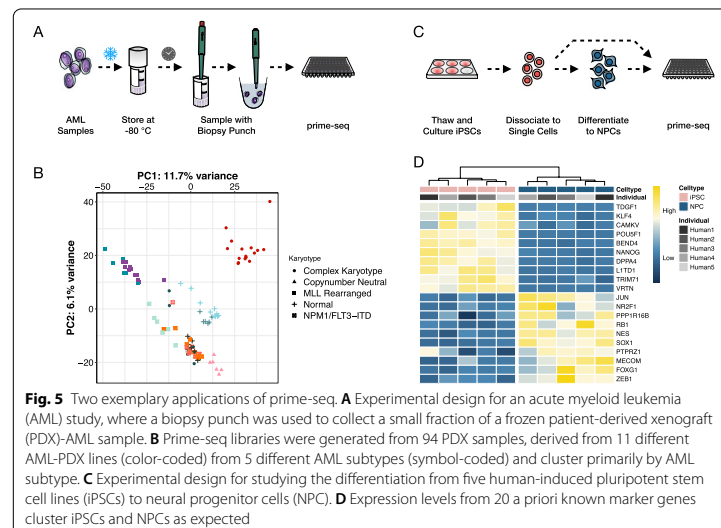
### Two exemplary applications of prime-seq

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  $-80^{\circ}\text{C}$  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



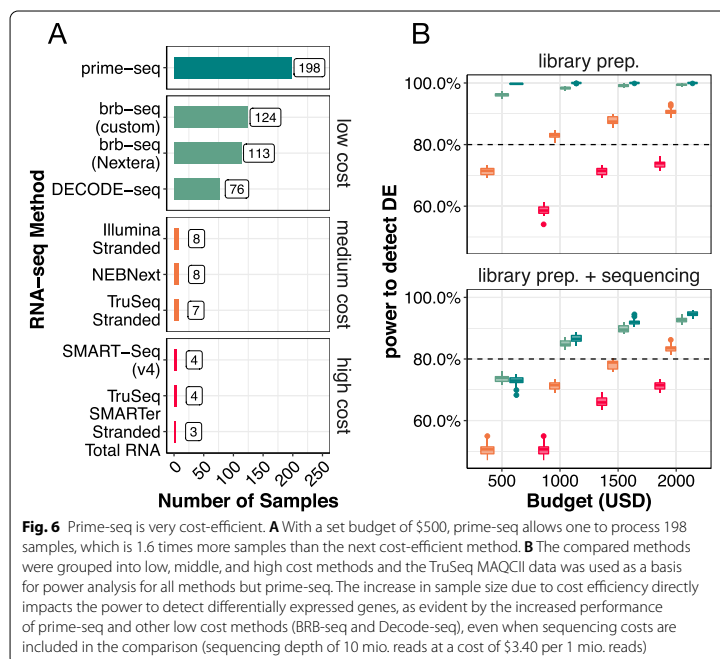
summary, one could argue that prime-seq performs as well as TruSeq for quantifying gene expression levels. Other methods that generate tagged cDNA libraries using early barcoding have also been developed [16, 22, 61–64]. This includes BRB-seq that uses poly(A) priming and DNA-Pol I for second-strand 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



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 NEB-Next 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<sub>2</sub>O. 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<sub>2</sub>O.

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<sub>2</sub>O. 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  $\mu$ L reactions. Following tagmentation, the libraries were amplified using 0.1  $\mu$ M P5NextPT5 primer (IDT) and 0.1  $\mu$ M i7 index primer (IDT) in a reaction volume of 50  $\mu$ L. 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% β-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  $\mu$ L of lysis buffer, RLT Plus, and 1%  $\beta$ -mercaptoethanol. The tubes were snap frozen and stored at  $-80^{\circ}\text{C}$  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  $50^{\circ}\text{C}$  for 15 min and  $75^{\circ}\text{C}$  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  $\mu$ L of ddH<sub>2</sub>O, 2.8  $\mu$ L of UHRR (100 ng/ $\mu$ L) (Thermo Fisher, QS0639), and 2.5  $\mu$ L 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  $\mu$ L 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 QuantiFlour 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  $9 \times 10^3$  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 $\times$  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  $\mu$ M 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  $2 \times 10^4$  were lysed in 100  $\mu$ L of lysis buffer (RLT Plus (Qiagen, 1053393) and 1%  $\beta$ -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  $-80^\circ\text{C}$  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  $\mu$ L 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  $\mu$ L of cryopreserved cells corresponding to roughly  $1 \times 10^5$  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.** Intronic 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 Maïke 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](mailto: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**

<sup>1</sup>Anthropology & Human Genomics, Faculty of Biology, Ludwig-Maximilians University, Großhaderner Str. 2, 82152 Martinsried, Germany. <sup>2</sup>Graduate School of Systemic Neurosciences, Faculty of Biology, Ludwig-Maximilians University, Martinsried, Germany. <sup>3</sup>Research Unit Apoptosis in Hematopoietic Stem Cells, Helmholtz Zentrum München, German Research Center for Environmental Health (HMGU), Munich, Germany. <sup>4</sup>German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany. <sup>5</sup>Department of Pediatrics, Dr. von Hauner Children's Hospital, Ludwig-Maximilians University, Munich, Germany. <sup>6</sup>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, Alvarez-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 mCSCR-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, Nemeshe 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, Munafó 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, Sönmez 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 Danaif N, Schimpf 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, Yigit 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, Metzler 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, Esглеас 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, Bhat 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. Tidy: 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:

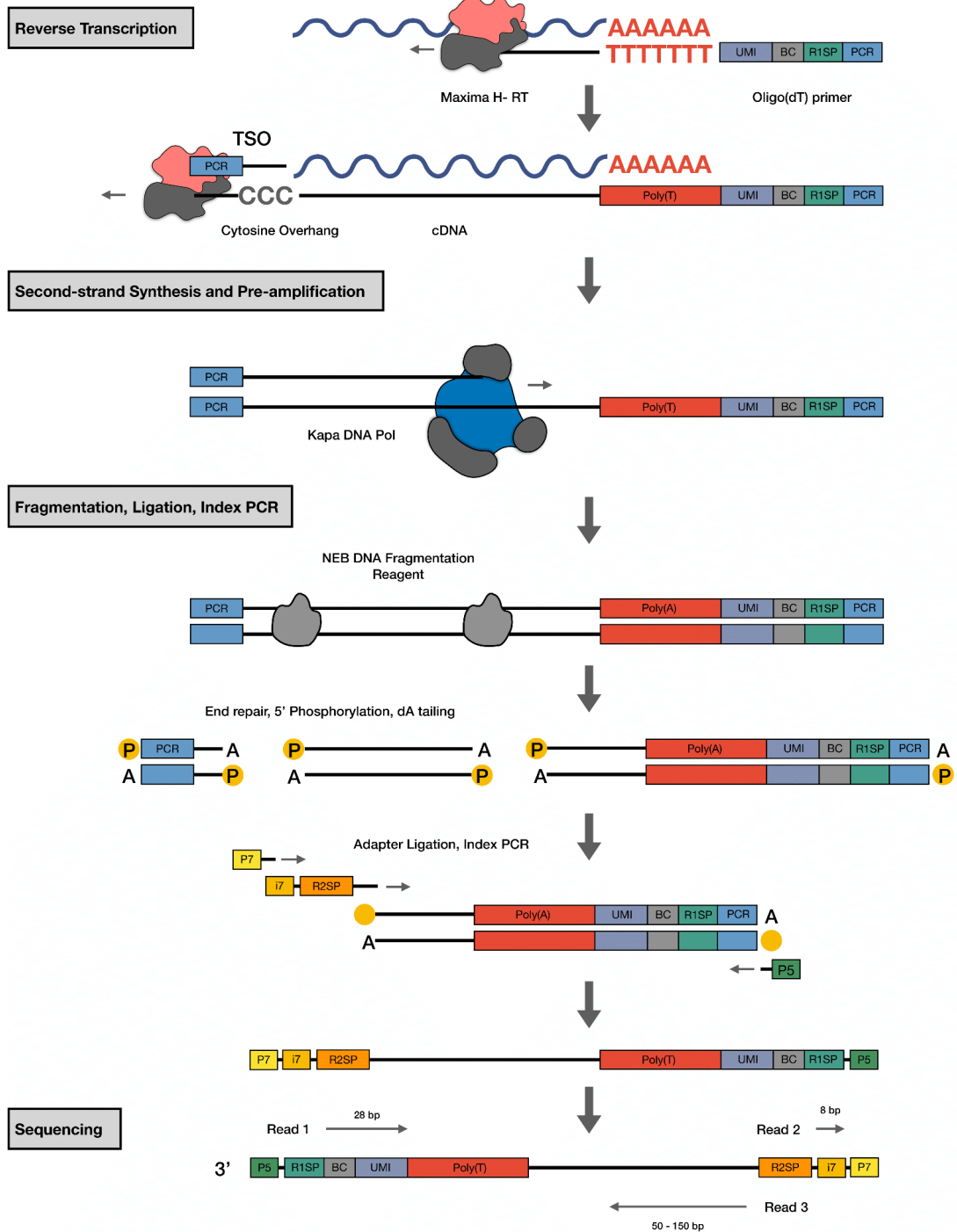
- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility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

At BMC, research is always in progress.

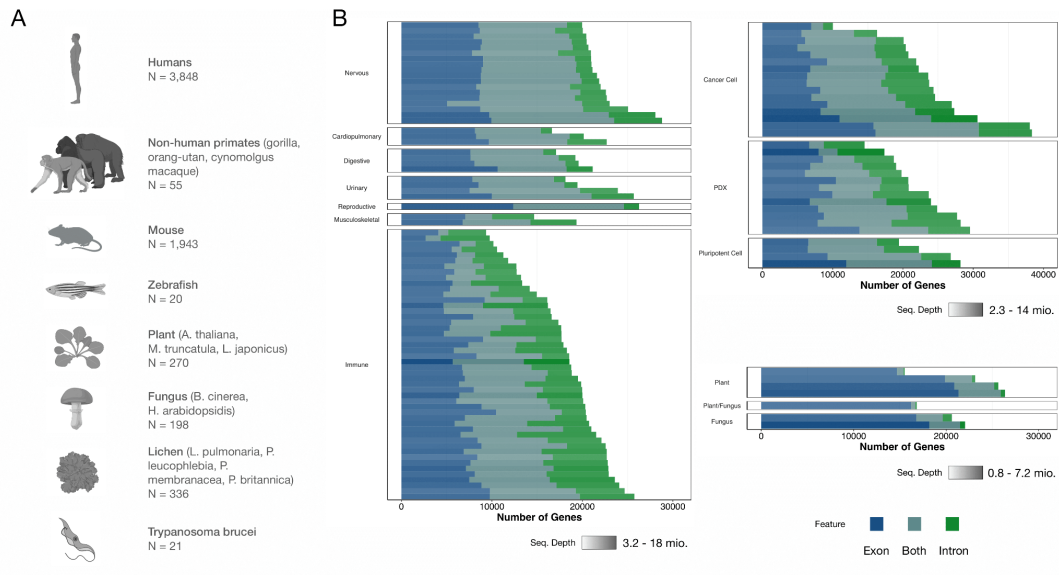
Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)



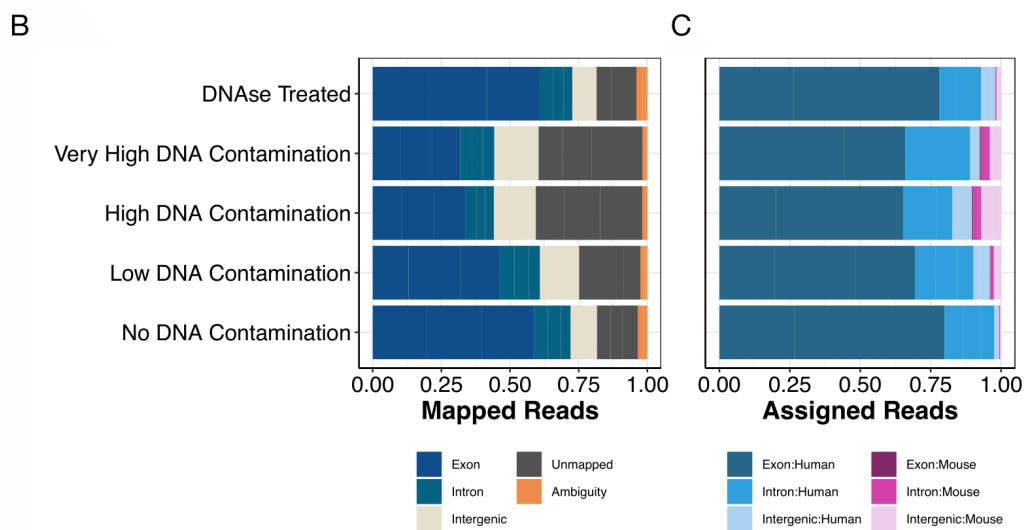
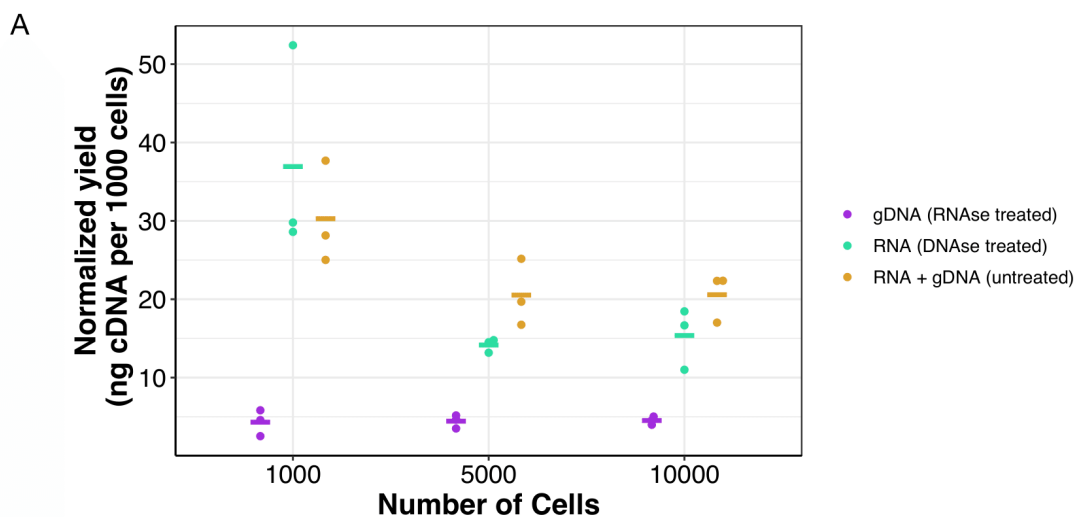
## Supplementary Information



**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.

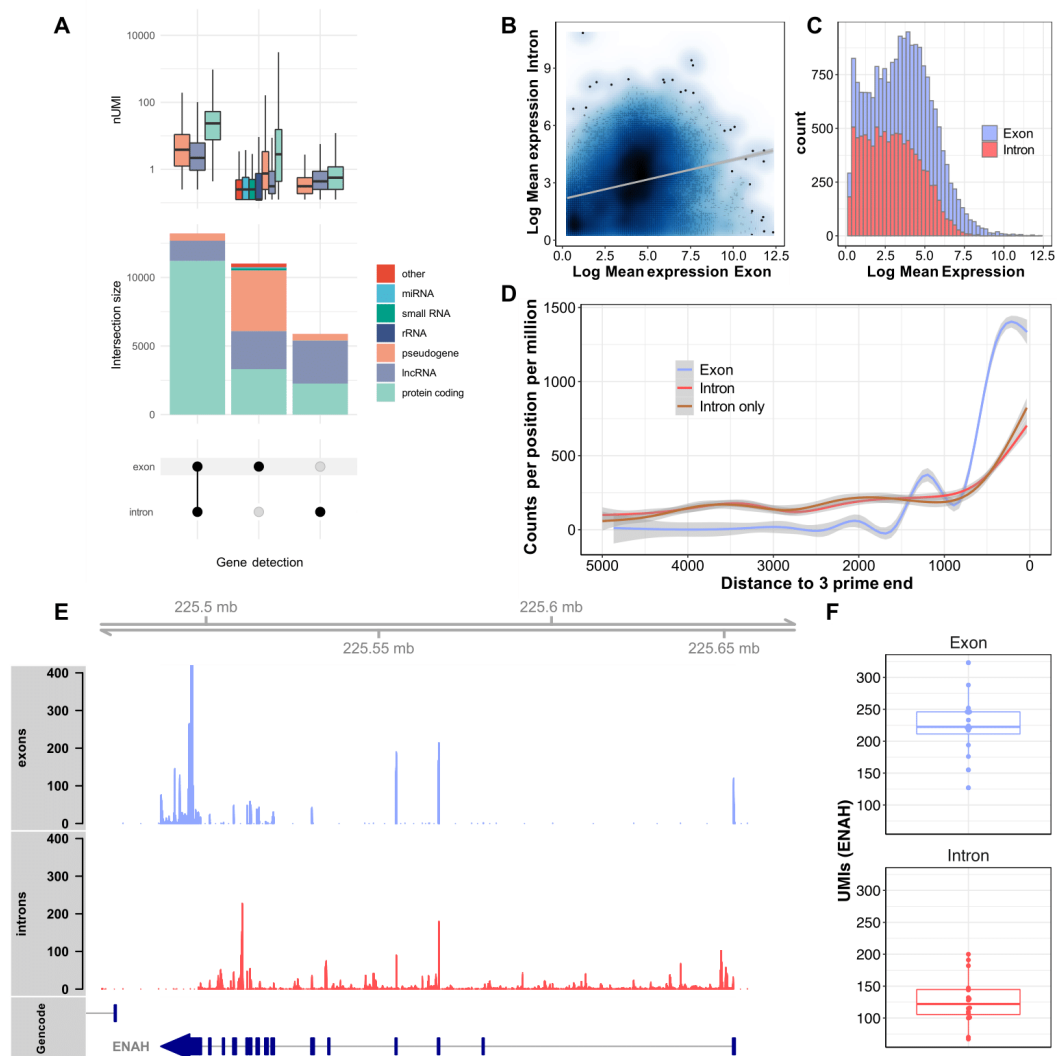


**Fig. S2. prime-seq is a robust protocol and has been validated with numerous organisms.** (Related to Figure 2A) (A) To date, 132 experiments consisting of 6,691 samples from 17 different organisms, ranging from arabidopsis to zebrafish, have been processed with prime-seq. (B) Data from experiments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

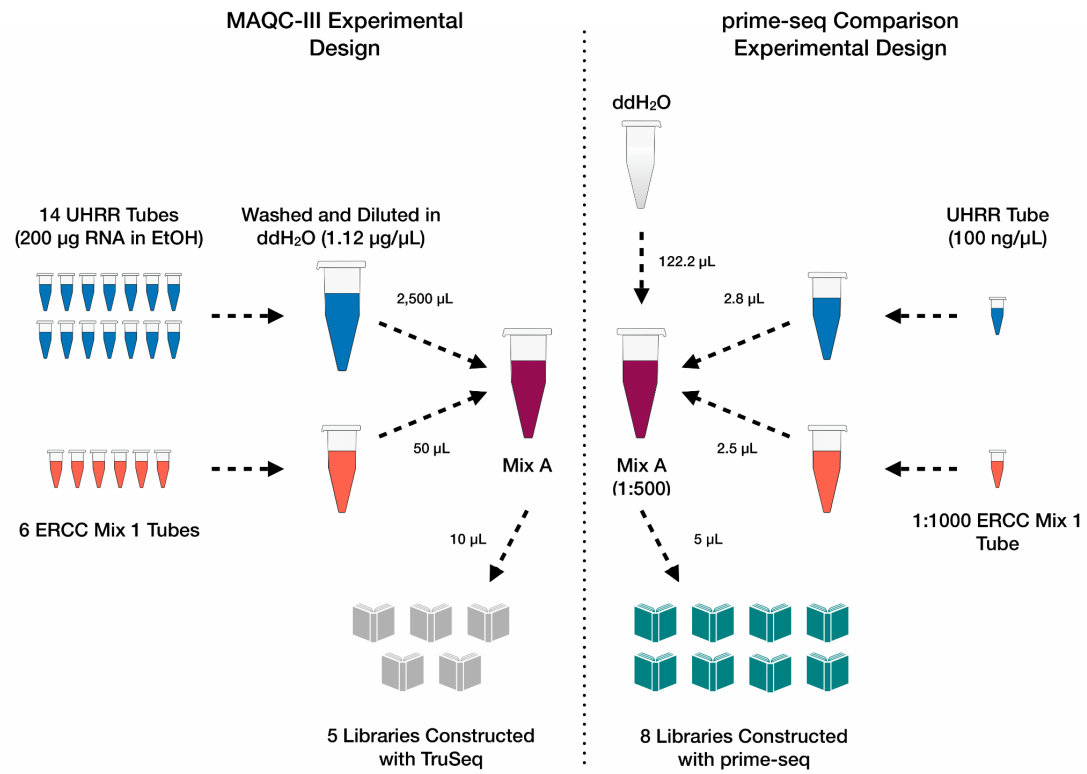




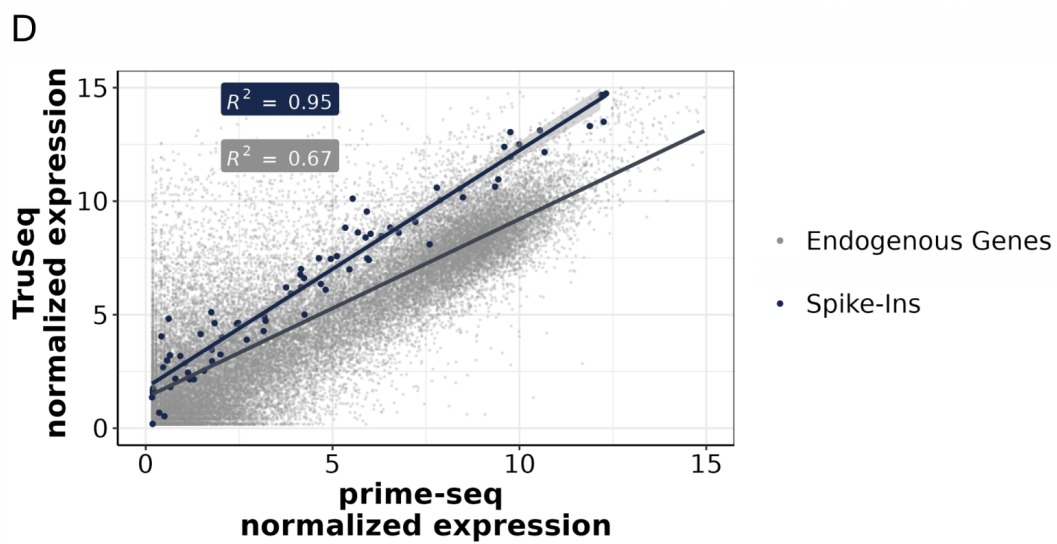
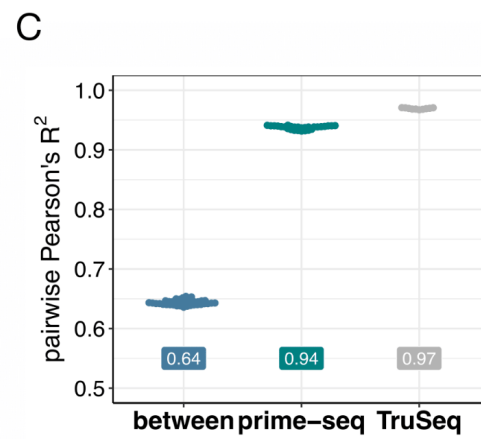
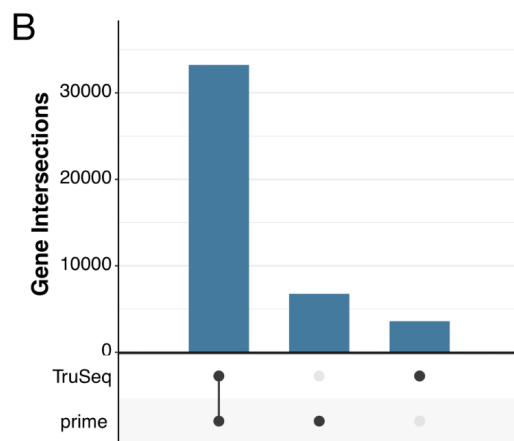
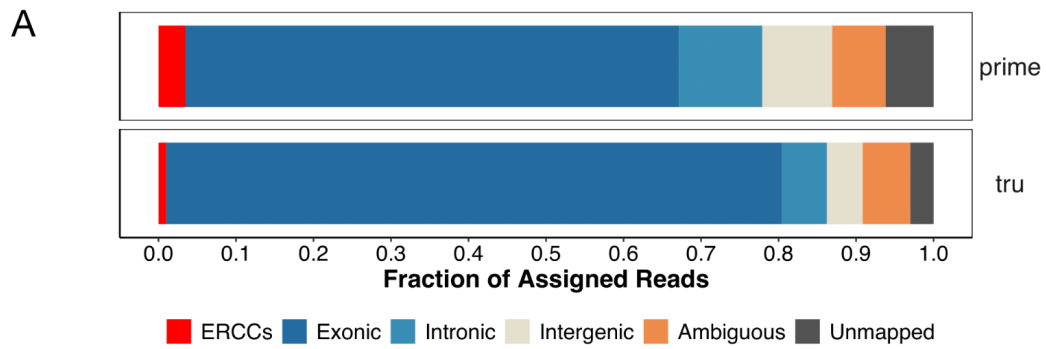
**Fig. S3. Intronic reads are not derived from contaminating gDNA.** (A) Samples containing total nucleic acids were either treated with RNase A or DNase I, or remained untreated. Untreated samples had the highest concentration, showing that genomic DNA is also used as a template when not removed, albeit 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.



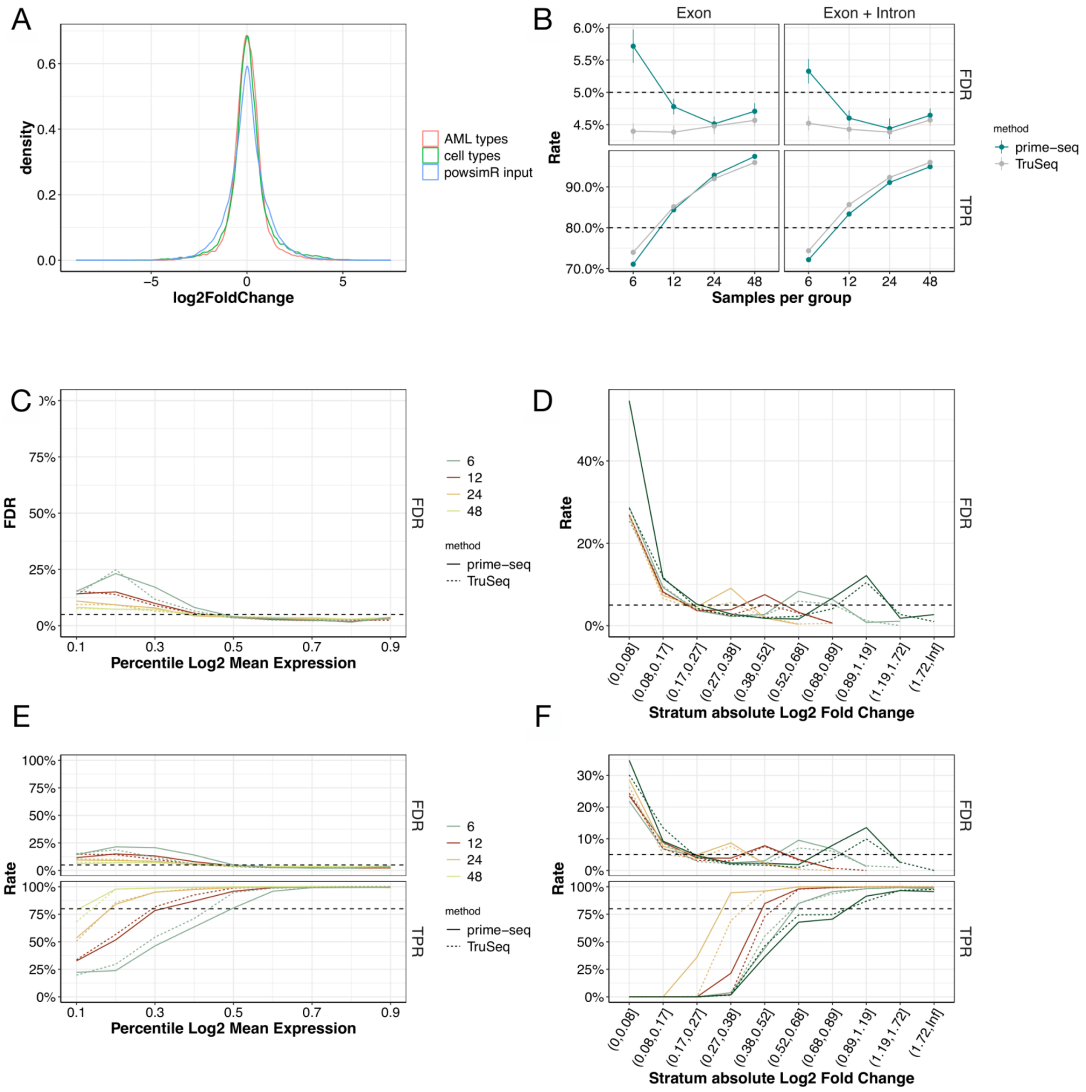
**Fig. S4. Intron counts are enriched at the 3' prime end and correlate with exon counts.** (A) Upset plot showing the intersection of genes detected with reads mapped to exons, introns or both. Most genes are detected in both introns and exons, followed by exons and introns only. Color represents the 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.



**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.

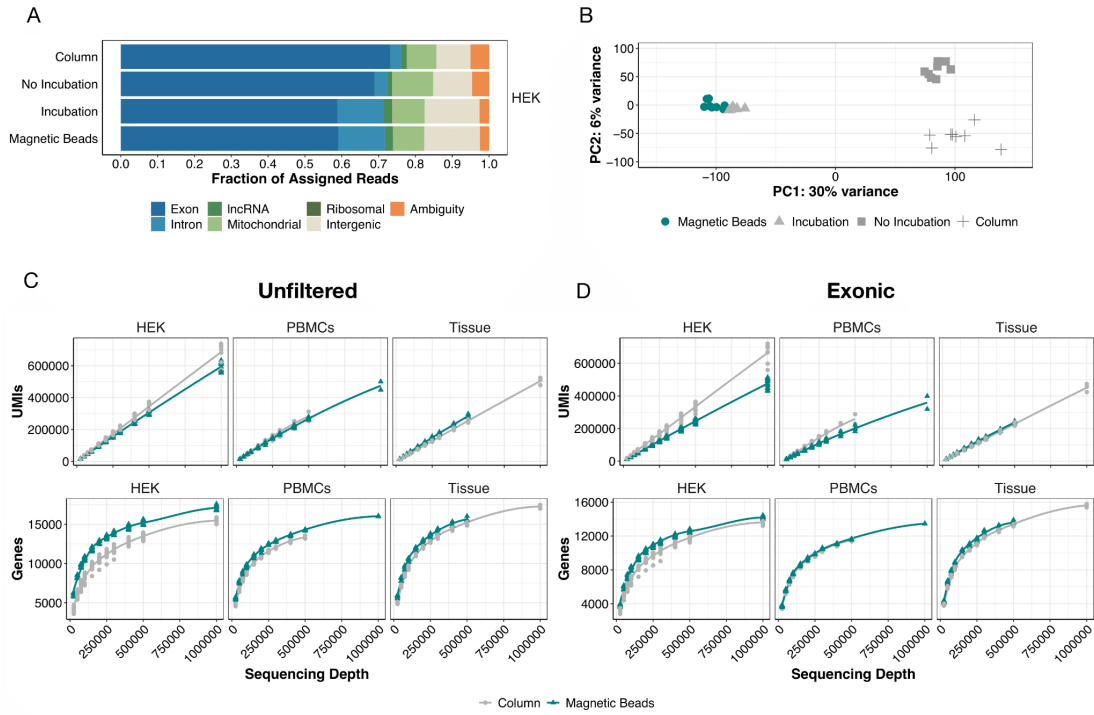


**Fig. S6. prime-seq and TruSeq have similar mapping, gene detection, and expression.** (Related to Figure 3)  
(A) Feature distribution from prime-seq and TruSeq shows 78% and 85% of reads are exonic, intronic, and ERCCs, respectively. (B) TruSeq and prime-seq exhibit a strong overlap of detected genes (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 ( $R^2 = 0.95$ ).



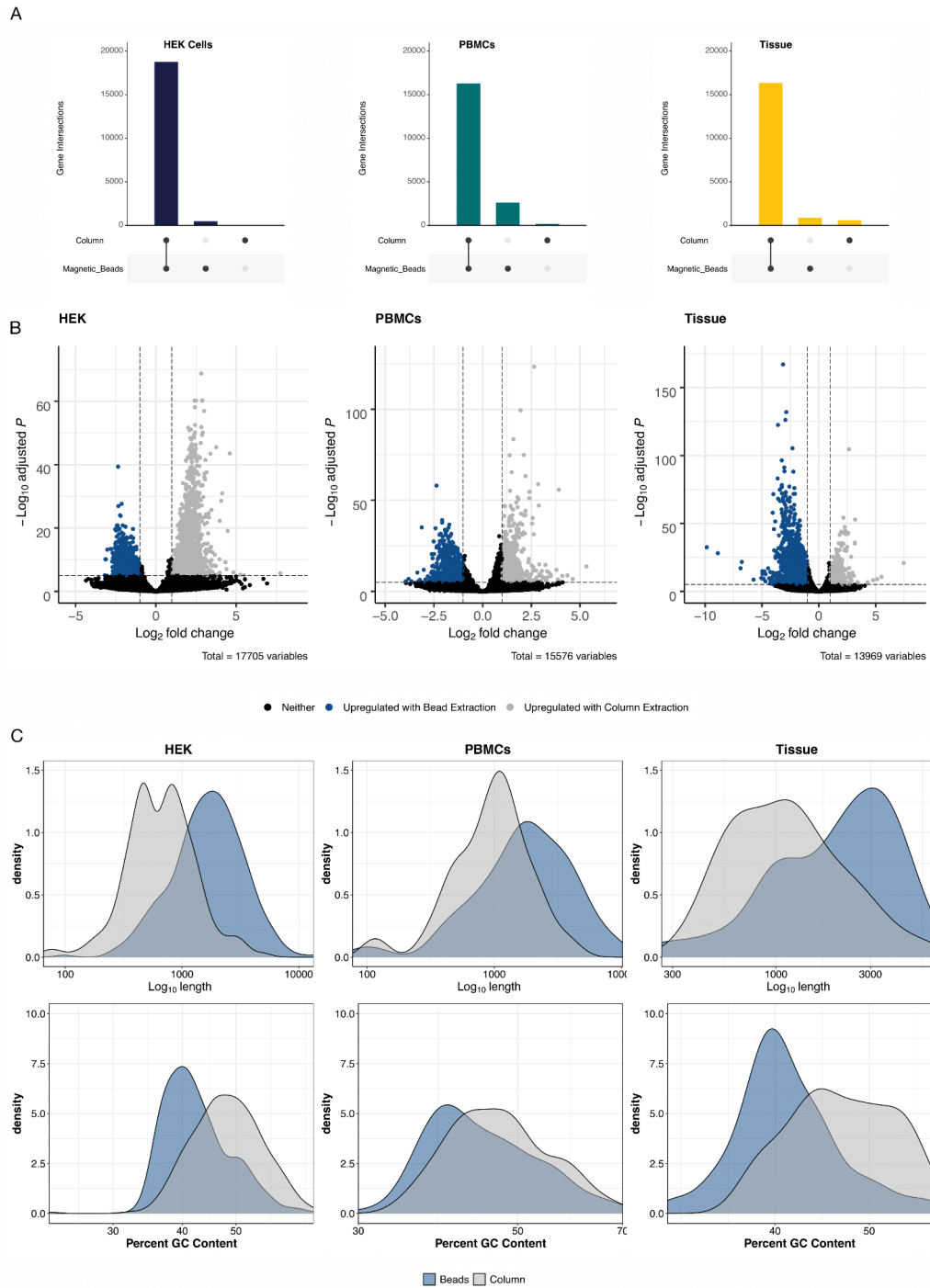


**Fig. S7. Power and FDR mostly depend on sample size and are similar between prime-seq and TruSeq.** (Related to Figure 3) (A) Log<sub>2</sub> fold change distribution from the AML and NPC differentiation experiment (Figure 4) compared to the log<sub>2</sub> fold change distribution used in powsimR 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 log<sub>2</sub> 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 log<sub>2</sub> fold change strata. Filtering parameters: detected UMI  $\geq 1$ , detected gene present in at least 25%.

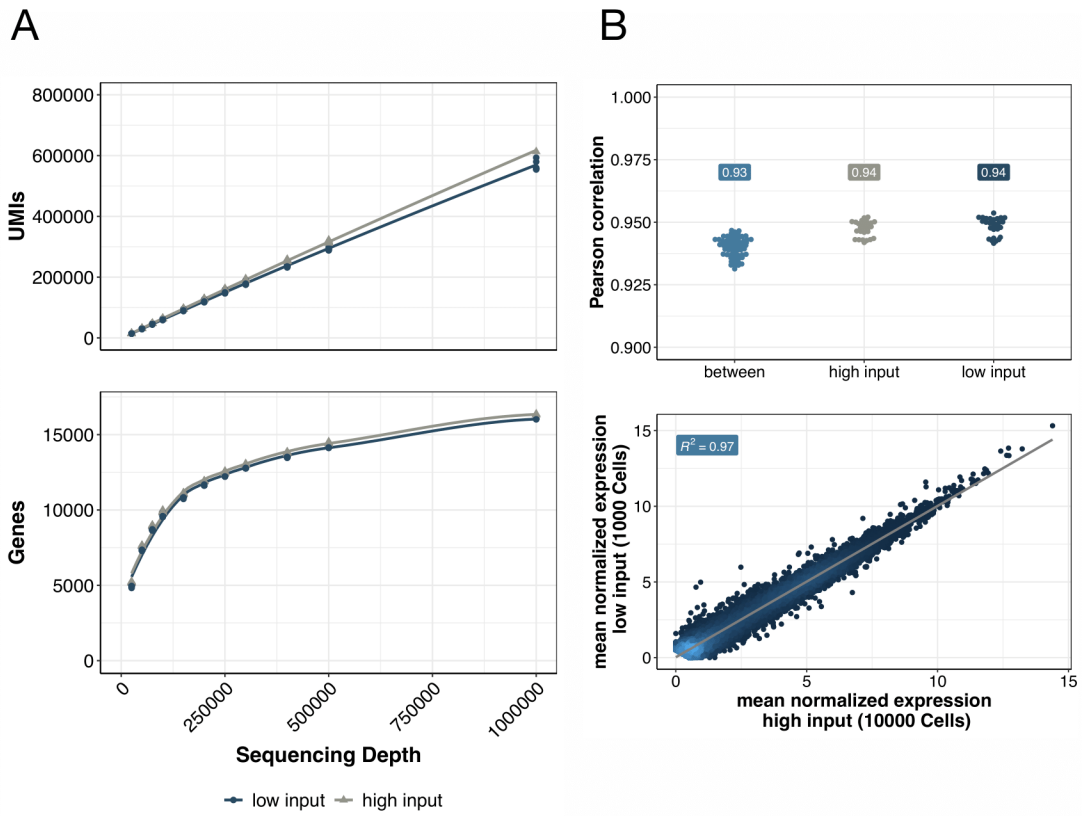


**Fig. S8. Performance of isolation methods is similar independent of prefiltering or usage of only Exon data.**

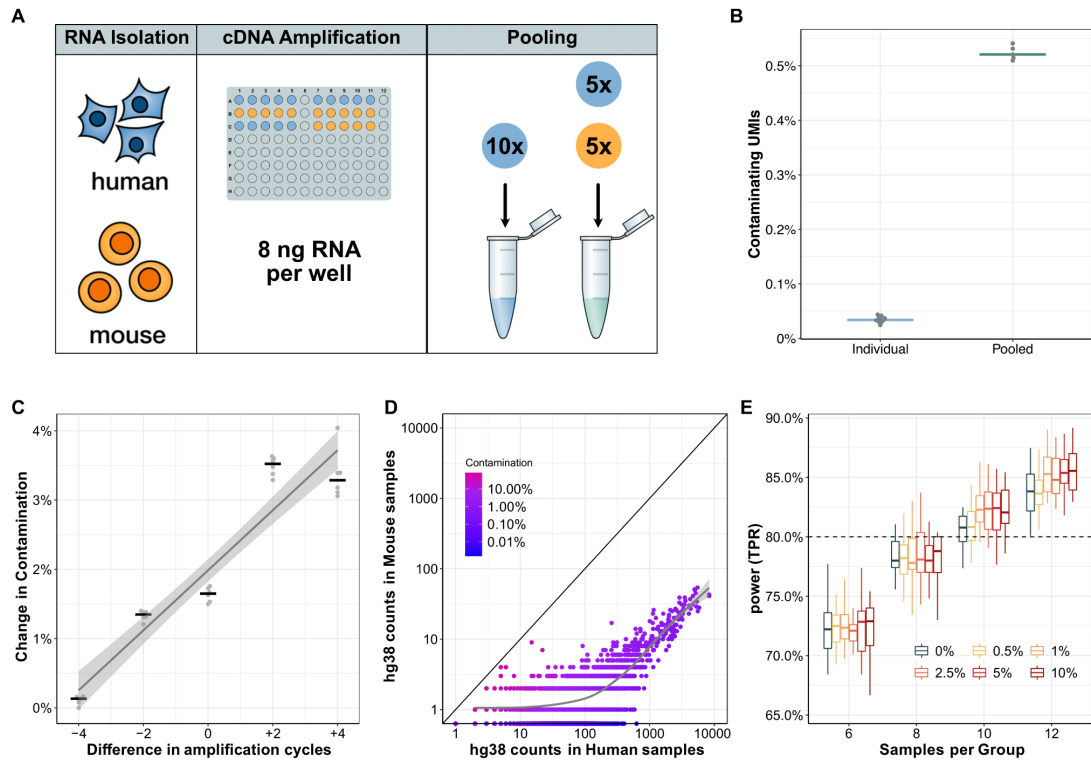
(Related to Figure 4) (A) HEK293T cell samples were extracted using columns and magnetic beads, employing the standard prime-seq protocol (“Magnetic Beads”), as well as variant protocols without proteinase K digestion (“No Incubation”) and a proteinase K digestion control without enzyme (“Incubation”). 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  $\geq 1$ , detected gene present in at least 25% of samples and is protein coding.



**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 ( $p > 0.05$ ,  $\log_2 \text{FC} > 2$ ). (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.

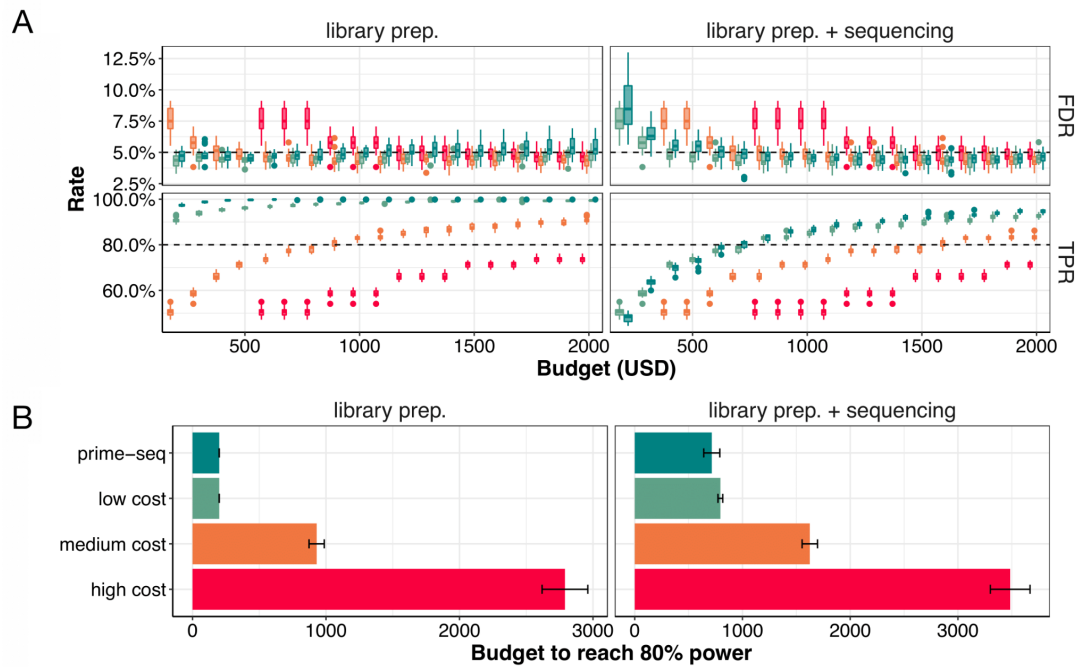


**Fig. S10. prime-seq performs equally well with high- and low-input samples.** (Related to Figure 5) (A) Sensitivity, measured in detected UMIs and genes, is similar between high input (10,000 HEK293T cells) and low input (1,000 HEK293T cells) conditions at various sequencing depths (filtering parameters: detected UMI  $\geq 1$ , 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,  $R^2 = 0.97$ ).





**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.



**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).



## 5.4 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.



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

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

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<sup>1</sup>. 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<sup>2,3</sup>. However, obtaining comparable samples from different primates, especially during development, is practically and—more importantly—ethically very difficult or even impossible.

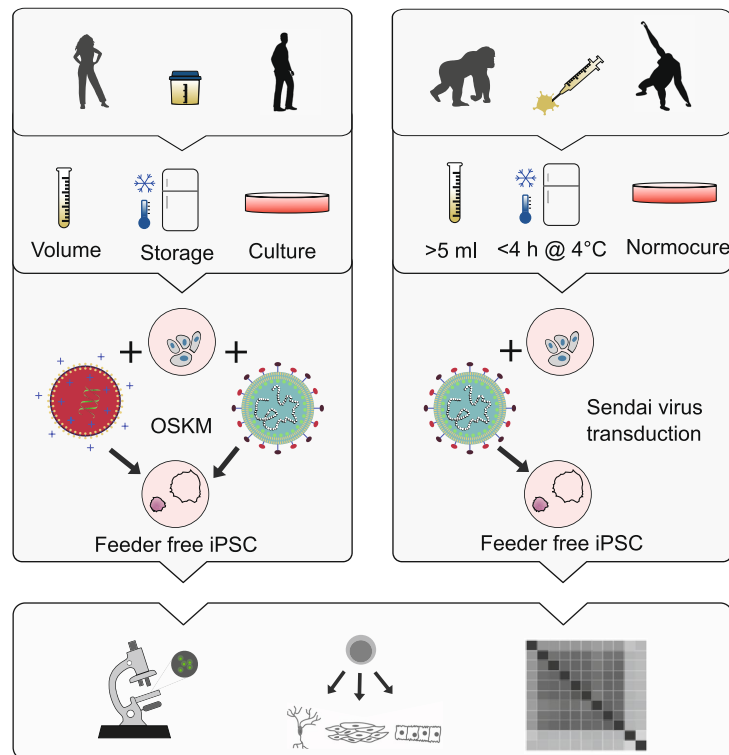
Embryonic stem cells have the potential to partially overcome this limitation by their ability to differentiate into all cell types *in vitro* and divide indefinitely<sup>4</sup>. However, the necessary primary material collection from an embryo is in most cases impossible. Fortunately, a pluripotent state can also be induced in somatic cells by ectopically expressing four genes<sup>5</sup>. Since this discovery of induced pluripotency, great efforts have been made to identify suitable somatic cells<sup>6</sup> and optimize reprogramming methods<sup>7</sup>. 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<sup>8–10</sup> and other mammals<sup>11,12</sup> 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<sup>8–10,13,14</sup>. 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<sup>15,16</sup> and applied it to non-human primates (Fig. 1). We find that primary cells can be isolated from

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

www.nature.com/scientificreports/



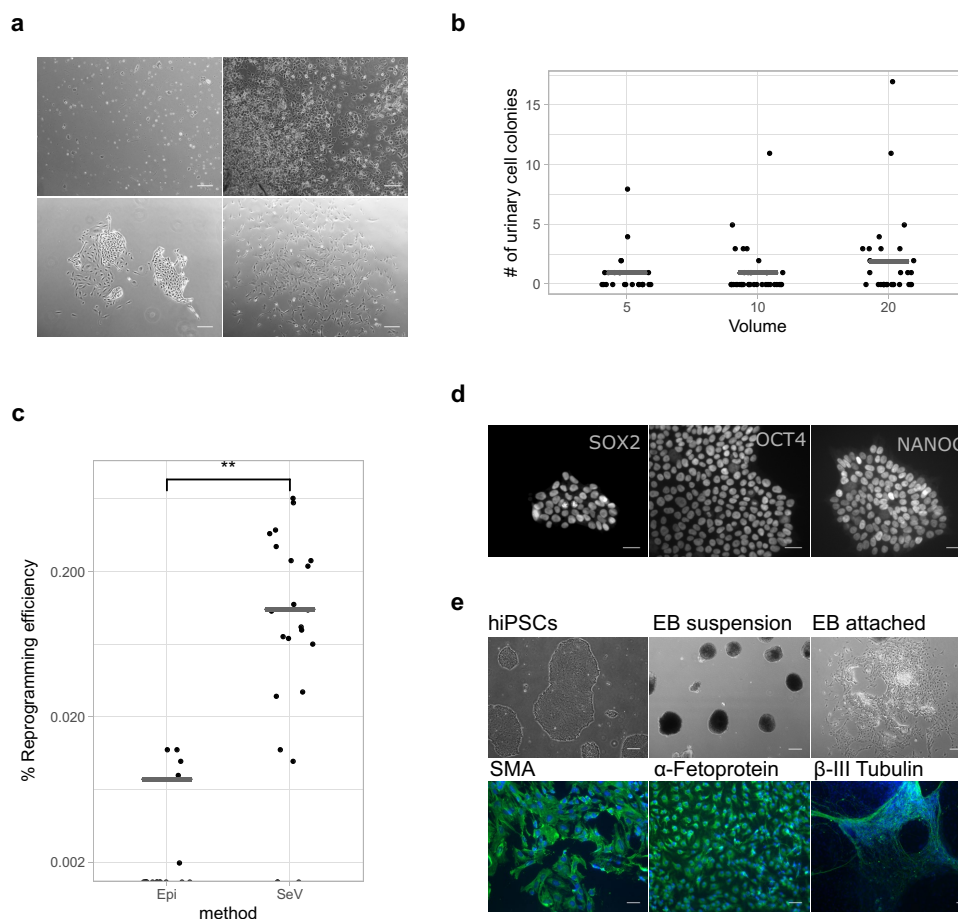
**Figure 1.** Workflow overview for establishing iPSCs from primate urine. We established the protocol for iPSC generation from human urine based on a previously described protocol<sup>16</sup>. We tested volume, storage and culture conditions for primary cells and compared reprogramming by overexpression 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<sup>17</sup> 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.<sup>15,16</sup>. 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<sup>18,19</sup>, 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<sup>20</sup>, 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  $\mu\text{m}$ . (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:  $p = 1.1 \times 10^{-5}$ ). (d) Established human colonies transduced with SeV expressed Nanog, Oct4 and Sox2; Scale bars represent 50  $\mu\text{m}$  and (e) differentiated to cell types of the three germ layers; scale bar represents 500  $\mu\text{m}$  in the phase contrast pictures and 100  $\mu\text{m}$  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.

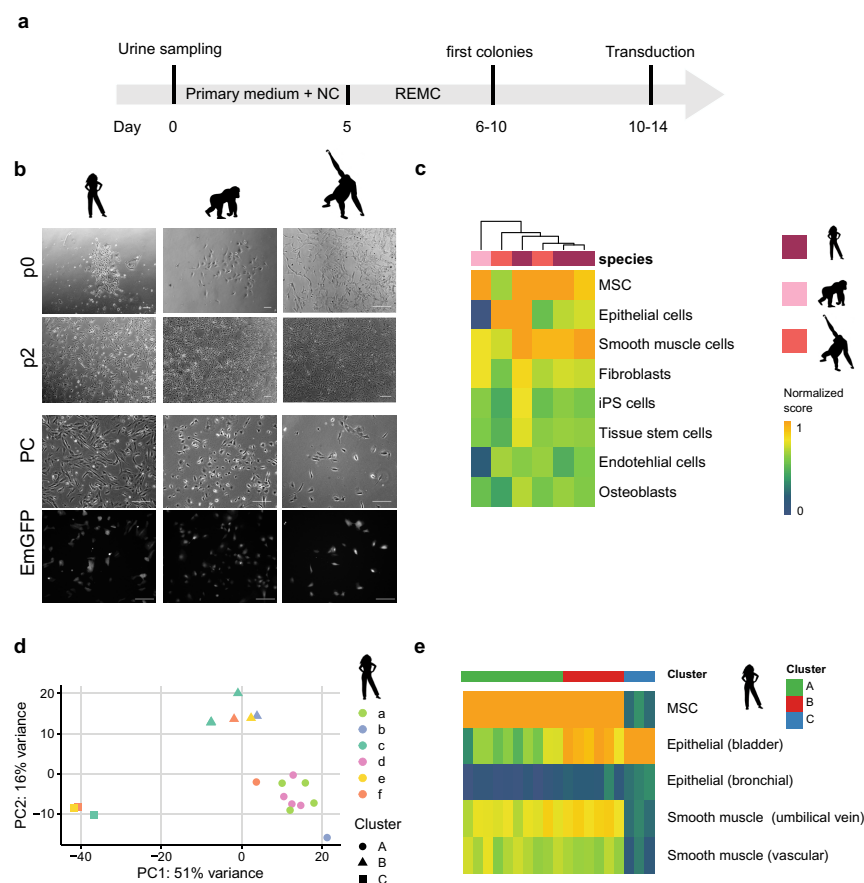
**Reprogramming human urinary cells is efficient when using suspension-Sendai Virus transduction.** Next, we investigated which integration-free overexpression strategy would be the most suitable to induce pluripotency in the isolated urine cells. To this end we compared transduction by a vector derived from the RNA-based Sendai Virus<sup>14,17</sup> in suspension<sup>10</sup>, to lipofection with episomal plasmids (Epi) derived from the Epstein Barr virus<sup>21,22</sup>. 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<sup>18,23</sup> 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 ( $\geq 5$  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<sup>24–26</sup>, 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<sup>27</sup> using the SingleR package<sup>28</sup>. 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<sup>27,28</sup> 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  $\mu$ m. **(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 Dörrenhaus et al.<sup>19</sup> 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 Dörrenhaus et al.<sup>19</sup> 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<sup>29</sup> 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<sup>15</sup>. 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 ( $\alpha$ -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)<sup>30</sup>. 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<sup>24–26</sup> to generate expression profiles with 19,000 genes detected on average.

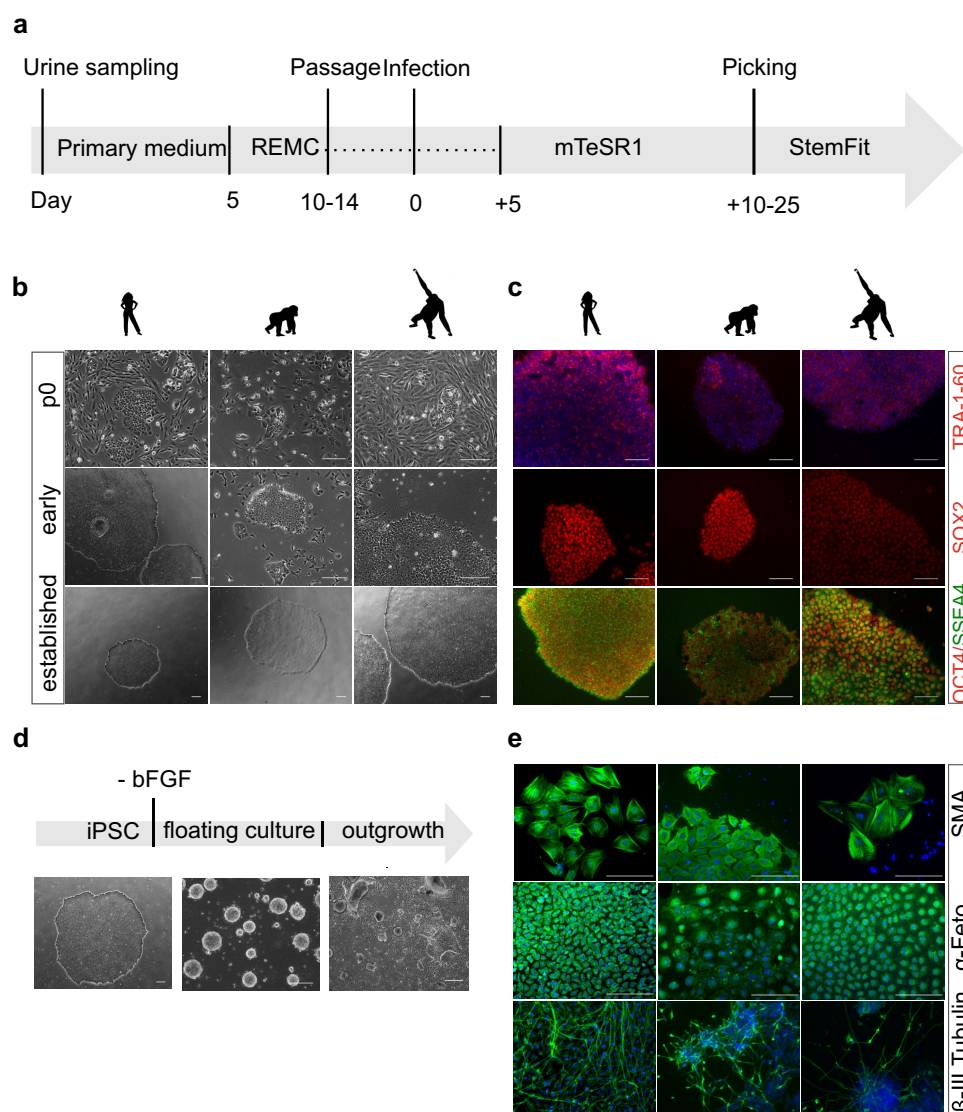
We classified the expression pattern of the iPSCs relative to the reference dataset of 38 cell types using SingleR as described for the urinary cells. ES cells or iPS cells are clearly the most similar cell type for all our iPS samples including the external PBMC-derived iPSC line (Fig. 5a). Principal 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

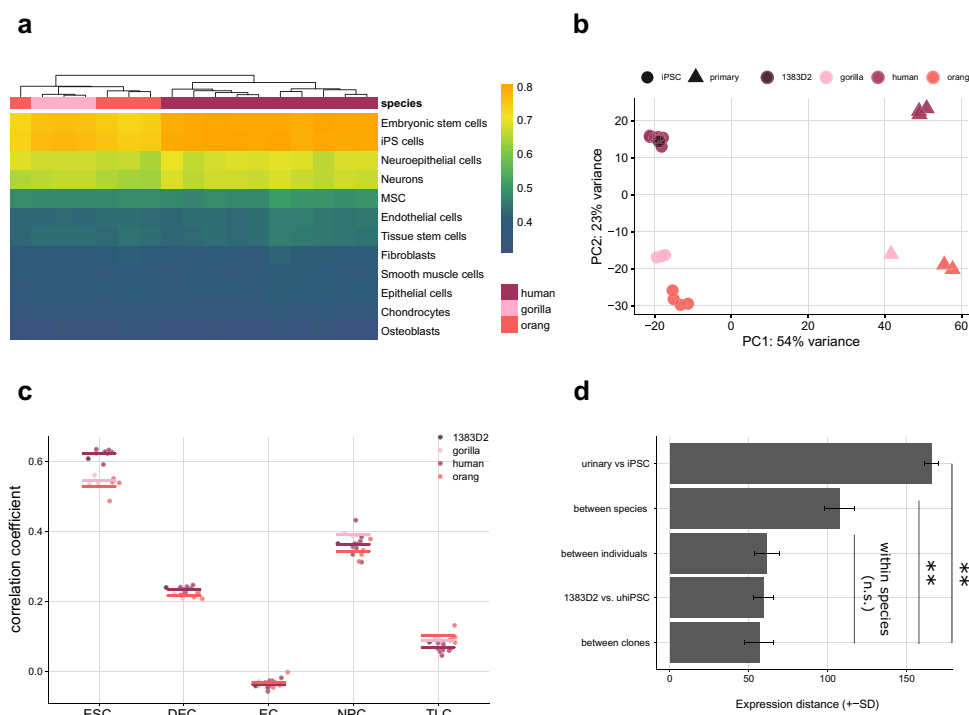
Here, we adapted a previously described protocol for human urine samples<sup>16</sup> to isolate proliferating cells from unsterile primate urine. We show that these urinary cells can be efficiently reprogrammed into integration-free and feeder-free iPSCs, which are closely comparable among each other 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<sup>31</sup> 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  $\mu$ m. **(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  $\mu$ m. **(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  $\mu$ m. **(e)** Immunofluorescence analyses of ectoderm ( $\beta$ -III Tubulin), mesoderm ( $\alpha$ -SMA) and endoderm markers ( $\alpha$ -Feto) after EB outgrowth. Nuclei were counterstained with DAPI. Scale bars represent 400  $\mu$ m. See also Supplementary Figure S7a.

www.nature.com/scientificreports/



**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 iPSC 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<sup>32</sup> for a recent overview). As these urine-derived stem cells (UDSCs) can be isolated non-invasively at low costs and reprogrammed efficiently<sup>16</sup>, they are increasingly used to generate iPSCs from patients (e.g.<sup>33–35</sup>). 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<sup>20</sup>, 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<sup>19</sup>, 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<sup>10</sup> 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<sup>7,36</sup>, 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<sup>37,38</sup>, 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<sup>11</sup> 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<sup>39</sup>, 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<sup>6</sup>. 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<sup>40</sup> it could be widely applicable. Additionally, iPSCs have been previously generated from many species, even avian species<sup>11</sup>, when using human reprogramming factors and culture conditions, albeit with over tenfold lower reprogramming efficiencies<sup>41,42</sup>. 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<sup>42</sup> 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<sup>43</sup>, and the optimization of feeder-free culture conditions for rhesus macaques and baboons<sup>42</sup>. 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<sup>44</sup>. 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<sup>8,9,45,46</sup> to infer human-specific changes more robustly<sup>47</sup>. 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<sup>23,48</sup>. 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<sup>49</sup>. This information is crucial, for example, when inferring the pathogenicity of genetic variants<sup>50</sup>. 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<sup>10</sup>. 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<sup>3</sup> to 30 × 10<sup>3</sup> 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<sup>5</sup> 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<sup>5</sup> cells were infected with CytoTune-EmGFP SeV. Infected cells were seeded on Geltrex (Thermo Fisher) coated 12-well-plates, routinely 10 × 10<sup>3</sup> and 25 × 10<sup>3</sup> 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<sup>51</sup>).

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  $9 \times 10^3$  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 iPSC cells and the corresponding primary cells and one to further characterize human UDCs derived from single colonies. For the first experiment one colony per clone corresponding to  $\sim 2 \times 10^4$  cells and  $2 \times 10^3$  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<sup>24-26</sup>, was used for library preparation<sup>24-26</sup>. 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<sup>52</sup> using STAR 2.6.0a<sup>53</sup> to generate expression profiles for barcoded UMI data. All samples were mapped to the human genome (hg38).

www.nature.com/scientificreports/

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<sup>54</sup> 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/Singl-eR/>) 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<sup>28</sup>. The 200 most variable genes were used in the 'de' option of SingleR to compare the obtained expression profiles to<sup>55</sup> as well as HPCA<sup>27</sup>. 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 iPSCs 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 iPSCs. *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 mcSCR-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).

www.nature.com/scientificreports/

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. Alfoldi, 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](http://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 Information

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

Johanna Geuder<sup>1</sup>, Lucas E. Wange<sup>1</sup>, Aleksandar Janjic<sup>1</sup>, Jessica Radmer<sup>1</sup>, Philipp Janssen<sup>1</sup>, Johannes W. Bagnoli<sup>1</sup>, Stefan Müller<sup>2</sup>, Artur Kaul<sup>3</sup>, Mari Ohnuki<sup>1\*</sup>, Wolfgang Enard<sup>1\*</sup>

<sup>1</sup>Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Großhaderner Straße 2, 82152 Martinsried, Germany

<sup>2</sup>Institute of Human Genetics, Munich University Hospital, Ludwig-Maximilians-University Munich, 80336 Munich, Germany

<sup>3</sup>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

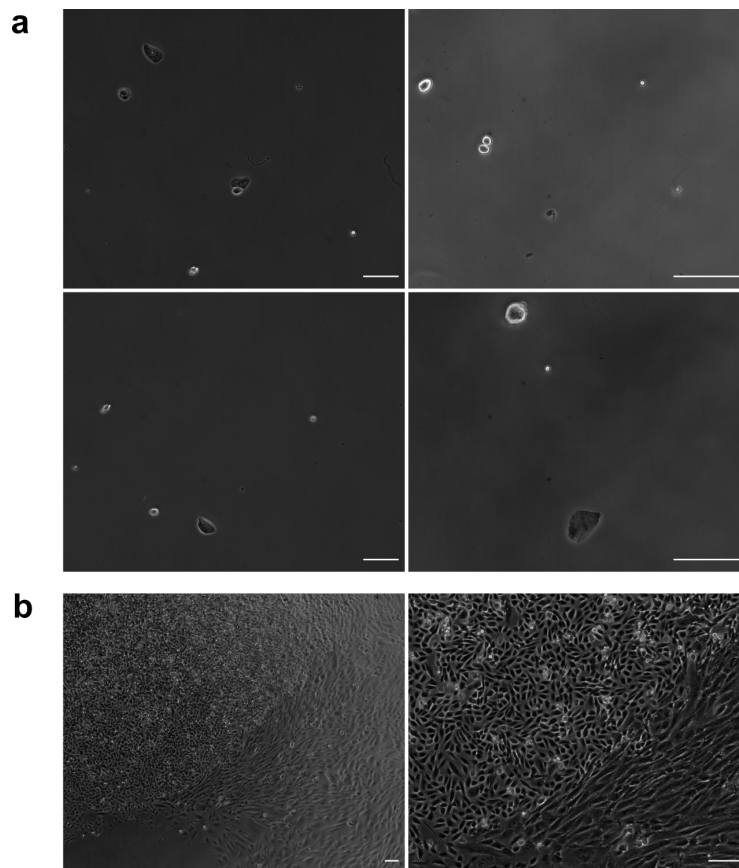
Großhaderner Str. 2

82152 Martinsried, Germany

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

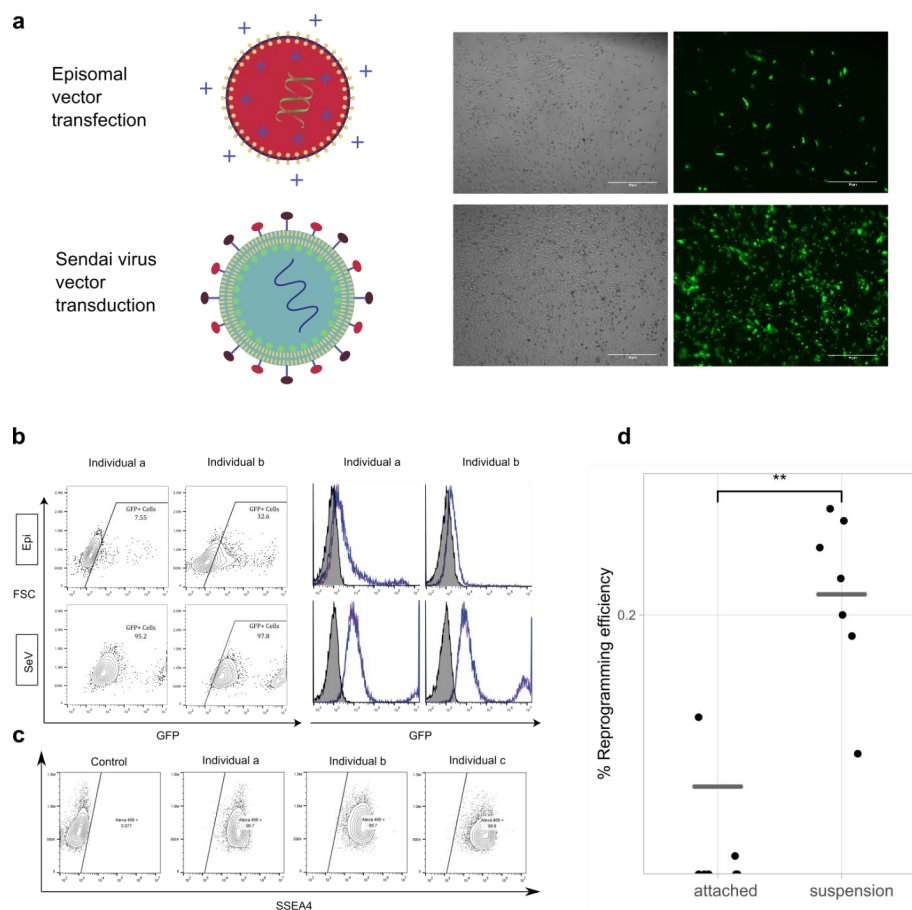
Fax: +49 (0)89 / 2180 - 74 331

E-Mail: [enard@bio.lmu.de](mailto:enard@bio.lmu.de), [ohnuki@biologie.uni-muenchen.de](mailto:ohnuki@biologie.uni-muenchen.de)



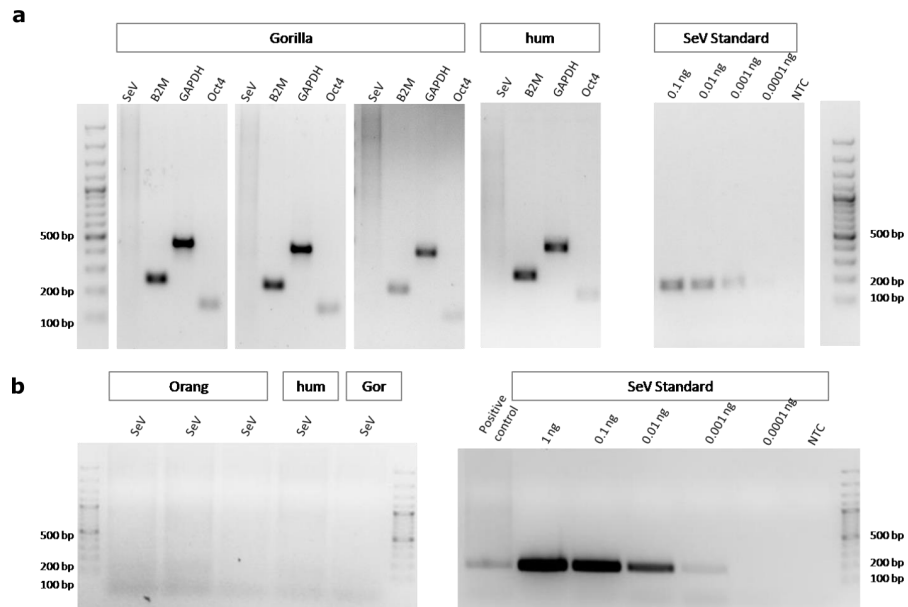
**Figure S1. Cell types found in human urine samples**

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**

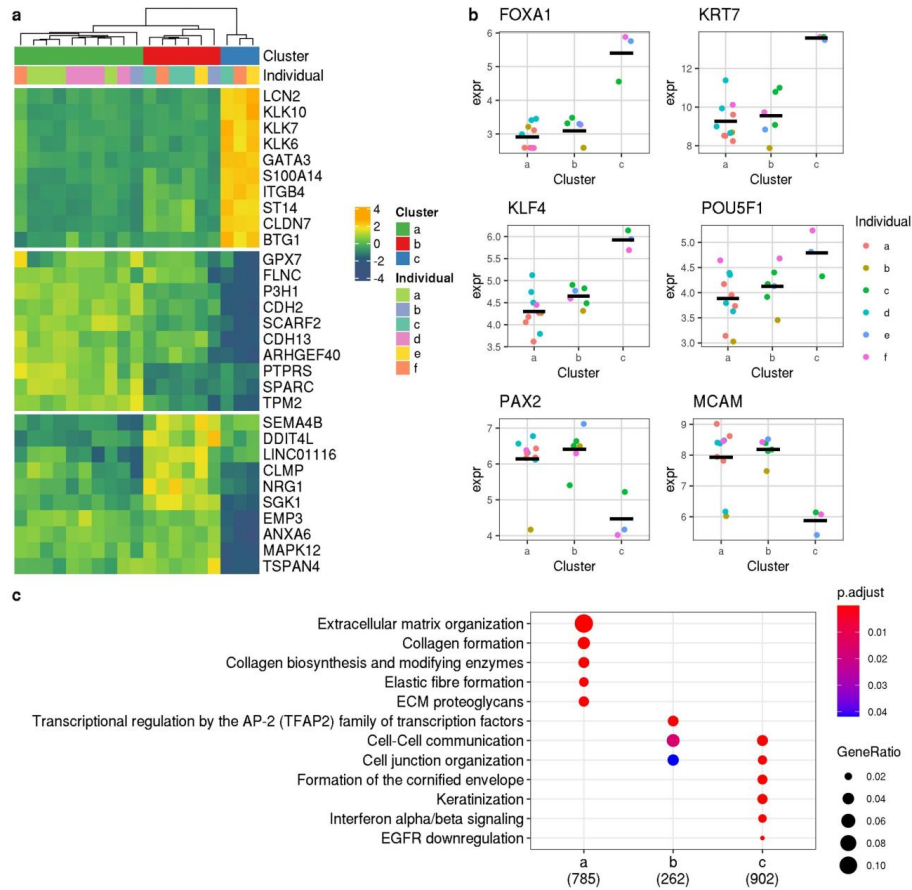
(a) GFP expression of urinary cells transfected with pcxle-EGFP episomal plasmids or CytoTune EmGFP transduced after 5 days (b) FACS analysis of GFP expressing cells 5 days post transfection/transduction (c) SSEA4 expression of urinary cells (d) Reprogramming efficiency comparison between attached and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



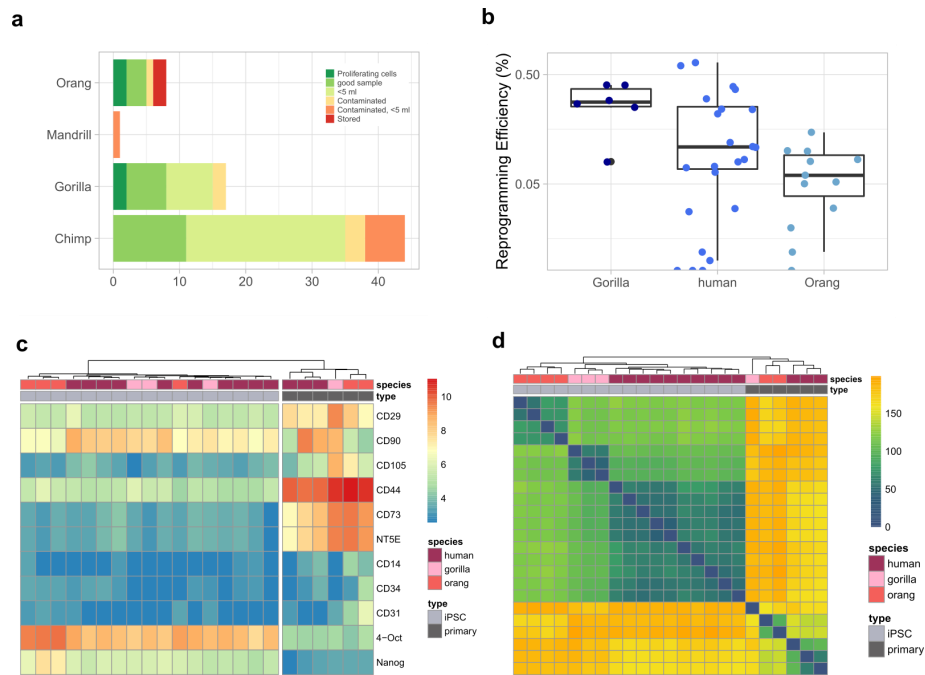
**Figure S3. SeV absence verification of primate iPSC lines**

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.



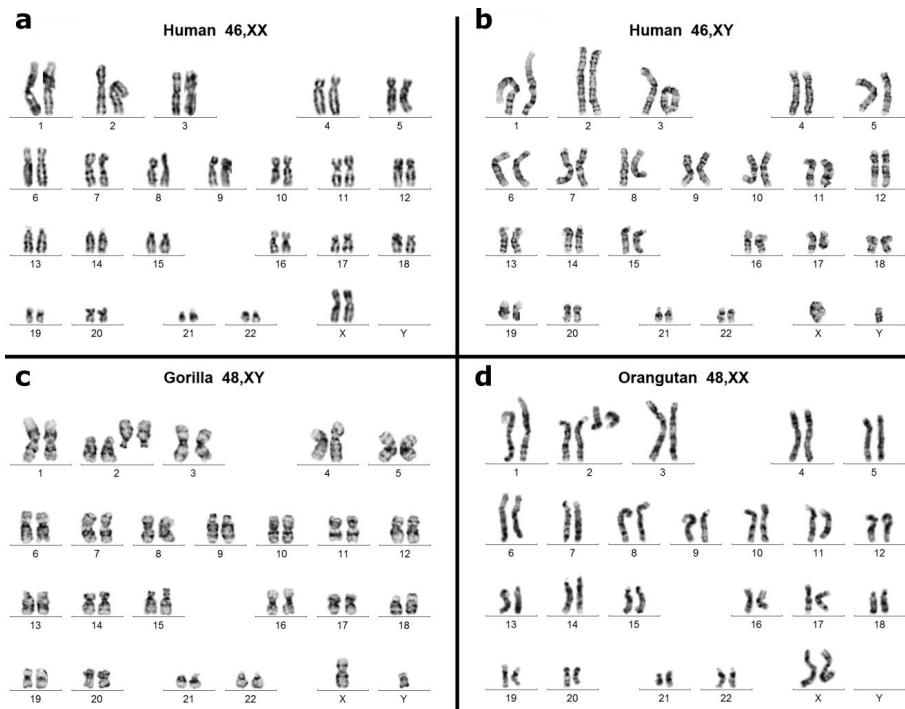


**Figure S4. Characterization of human UDSCs originating from single colonies**  
Expression profiles of single colonies from human urine samples were subjected to further analysis. (a) Heatmap of top differentially expressed genes between the clusters. (b) Marker gene expression of different cell clusters. 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.



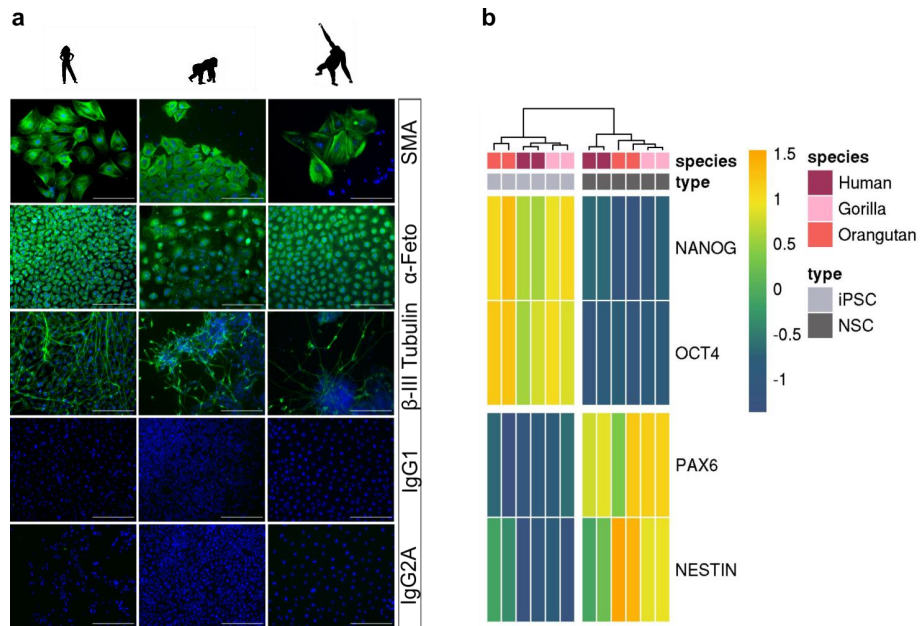
**Figure S5. UDSCs and corresponding iPSC characteristics**

(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 ( $\beta$ -III Tubulin), mesoderm ( $\alpha$ -SMA) and endoderm markers ( $\alpha$ -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 400 $\mu$ m.

(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.

## 5.5 TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals

Kliesmete, Zane and **Wange, Lucas E**; Vieth, Beate;Esgleas, Miriam; Radmer, Jessica; Hülsmann, Matthias; Geuder, Johanna; Richter, Daniel; Ohnuki, Mari; Götz, Magdalena; Hellmann, Ines; Enard, Wolfgang

"TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals" (2021)

*bioRxiv*

doi: 10.1038/s41467-019-12266-7

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

<https://www.biorxiv.org/content/10.1101/2021.02.05.429919v2.full>

### Abstract

Genomes can be seen as notebooks of evolution that contain unique information on successful genetic experiments. This allows to identify conserved genomic sequences and is very useful e.g. for finding disease-associated variants. Additional information from genome comparisons across species can be leveraged when considering phenotypic variance across species. Here, we exemplify such a cross-species association study for the gene *TRNP1* that is important for mammalian brain development. We find that the rate of TRNP1 protein evolution is highly correlated with the rate of cortical folding across mammals and that TRNP1 proteins from species with more cortical folding induce higher proliferation rates in neural stem cells. Furthermore, we identify a regulatory element in *TRNP1* whose activity correlates with cortical folding in Old World Monkeys and Apes. Our analyses indicate that coding and regulatory changes in *TRNP1* have modulated its activity to adjust cortical folding during mammalian evolution and provide a blueprint for cross-species association studies.

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

## TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals

Zane Kliesmete<sup>1,§</sup>, Lucas E. Wange<sup>1,§</sup>, Beate Vieth<sup>1</sup>, Miriam Esgleas<sup>2,3</sup>, Jessica Radmer<sup>1</sup>, Matthias Hülsmann<sup>1,4</sup>, Johanna Geuder<sup>1</sup>, Daniel Richter<sup>1</sup>, Mari Ohnuki<sup>1</sup>, Magdalena Götz<sup>2,3,5</sup>, Ines Hellmann<sup>1,\*</sup>,§, Wolfgang Enard<sup>1,\*</sup>,§

<sup>1</sup> Anthropology and Human Genomics, Department of Biology II, Ludwig-Maximilians Universität, Munich, Germany

<sup>2</sup> Department of Physiological Genomics, BioMedical Center - BMC, Ludwig-Maximilians Universität, Munich, Germany

<sup>3</sup> Institute for Stem Cell Research, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

<sup>4</sup> current address: Department of Environmental Microbiology, Eawag, 8600 Dübendorf, Switzerland & Department of Environmental Systems Science, ETH Zurich, 8092 Zürich, Switzerland

<sup>5</sup> SYNERGY, Excellence Cluster of Systems Neurology, BioMedical Center (BMC), Ludwig-Maximilians-Universität München, Planegg/Munich, Germany

§ equal author contribution

\* correspondence [hellmann@bio.lmu.de](mailto:hellmann@bio.lmu.de), [enard@bio.lmu.de](mailto:enard@bio.lmu.de)

### Abstract

Genomes can be seen as notebooks of evolution that contain unique information on successful genetic experiments<sup>1</sup>. This allows to identify conserved genomic sequences<sup>2</sup> and is very useful e.g. for finding disease-associated variants<sup>3</sup>. Additional information from genome comparisons across species can be leveraged when considering phenotypic variance across species. Here, we exemplify such a cross-species association study for the gene *TRNP1* that is important for mammalian brain development. We find that the rate of *TRNP1* protein evolution is highly correlated with the rate of cortical folding across mammals and that *TRNP1* proteins from species with more cortical folding induce higher proliferation rates in neural stem cells. Furthermore, we identify a regulatory element in *TRNP1* whose activity correlates with cortical folding in Old World Monkeys and Apes. Our analyses indicate that coding and regulatory changes in *TRNP1* have modulated its activity to adjust cortical folding during mammalian evolution and provide a blueprint for cross-species association studies.

Investigating mechanisms of molecular and cellular processes in model organisms by artificial genetic mutations is a central part of biological research. Investigating the evolution of organismic traits by analysing natural genetic mutations is another. These two areas increased their overlap due to the availability of genetic information from many organisms enabled by DNA reading technology and - more recently - also by testing genetic variants from many organisms enabled by DNA writing technology.

Here, we use these newly available resources to better understand which structural or regulatory molecular changes are necessary for a primate or mammalian brain to increase its size and/ or folding<sup>4-8</sup>. Several genes have been connected to the evolution of brain size by comparing the functional consequences of orthologues between pairs of species as for example human and chimpanzee<sup>4,8-10</sup>. While mechanistically convincing, it is unclear whether the proposed evolutionary link can be



bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

generalised. On the other side, approaches that correlate sequence changes with brain size changes across a larger phylogeny often lack mechanistic evidence<sup>11</sup>. Hence, a combination of mechanistic and comparative genetic approaches would reveal functional consequences of natural genetic variants and would leverage unique information to better understand the mechanisms of brain evolution as well as brain development. For example, TRNP1 promotes proliferation and NSC self-renewal in murine and ferret NSCs and its knock-down or dominant-negative forms promote the generation of cells leading to cortical folding. Importantly, regulation of its expression in a block-wise manner is critical for folding<sup>6,12</sup> and its N- and C-terminal intrinsically disordered domains are crucial for its function in regulating several nuclear compartments and M-phase length<sup>13</sup>. Thus, there is strong evidence that the regulation of a single gene, *TRNP1*, is necessary and sufficient to induce cortical folds in ferrets and mice. But it is entirely unclear whether the evolutionary diversity in cortical folding and brain size across mammals is mediated by structural and regulatory evolution of TRNP1. To answer this, we investigate genetic differences in TRNP1 coding as well as regulatory sequences and correlate them with brain phenotypes across the mammalian phylogeny.

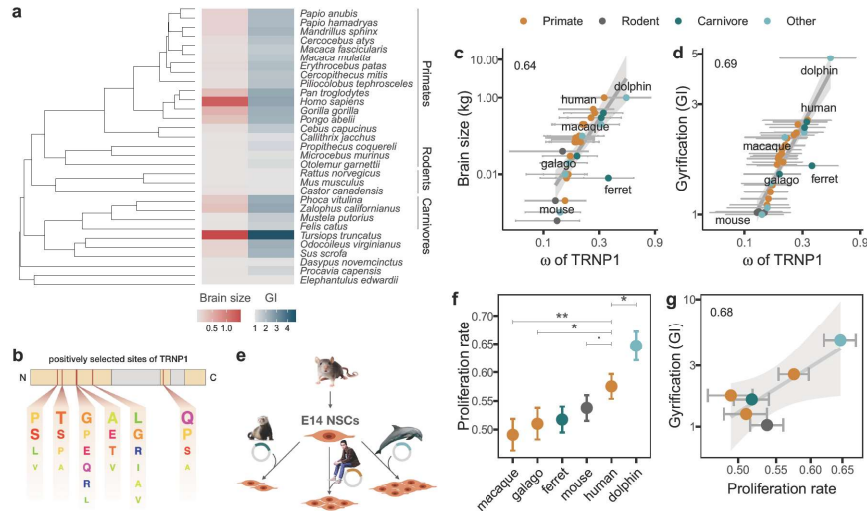
### Co-evolution between TRNP1 protein and cortical folding

We experimentally and computationally collected<sup>14</sup> and aligned<sup>15</sup> 45 mammalian TRNP1 coding sequences, including for example dolphin, elephant and 18 primates (97.4% completeness, Extended Data Fig. 1a). Using this large multiple alignment, we find that the best fitting evolutionary model includes that 8.2% of the codons show signs of recurrent positive selection<sup>16</sup> (i.e.  $\omega > 1$ ,  $p$ -value < 0.001, Suppl. Table 8). Six codons with a selection signature could be pinpointed with high confidence (Suppl. Table 9) and five out of those six reside within the first intrinsically disordered region (IDR) and one in the second IDR of the protein (Fig. 1b, Extended Data Fig. 1b). The IDRs have been shown to mediate homotypic and heterotypic interactions of TRNP1 and the associated functions of phase separation, nuclear compartment size regulation and M-phase length regulation<sup>13</sup>. While this shows that TRNP1 evolves under positive selection, it is yet unclear whether this selection is linked to cortical folding and/or brain size.

Cortical folding is usually quantified as the gyrification index (GI), which is the ratio of the cortical surface over the perimeter of the brain surface: A GI= 1 indicates a completely smooth brain and a GI > 1 indicates increasing levels of cortical folding<sup>17</sup>. In addition to GI and brain size, we also consider body mass as a potential confounding variable. Larger animals often have smaller effective population sizes, which in turn reduces the efficiency of selection. Therefore, a correlation between  $\omega$  and body size could also be explained by relaxation of constraint instead of directional selection<sup>18</sup>. Estimates for these three traits and TRNP1 sequences were available for 31 species (Fig. 1a). In order to test whether the evolution of the TRNP1 protein coding sequences is linked to any of the three traits, we used Coevol<sup>18</sup>, a Bayesian MCMC method that jointly models the rates of substitutions and quantitative traits. The resulting covariance matrix of substitution rates (branch length  $\lambda_S$ ,  $\omega$ ) and the phenotypic traits then allows for a quantitative evaluation of a potential co-evolution using the posterior probability ( $pp$ ) of the correlations<sup>18</sup>. Considering the traits separately, we find that GI has the highest marginal correlation with  $\omega$  ( $r=0.62$ ,  $pp=0.95$ ), followed by brain size ( $r=0.5$ ,  $pp=0.89$ ), and body mass ( $r=0.44$ ,  $pp=0.85$ ) (Suppl. Table 10). To better disentangle their effects, we then simultaneously inferred their correlations (Fig. 1c, 1d, Extended Data Fig. 1c, Suppl. Table 11). GI remained the strongest and only significant marginal correlation ( $r=0.69$ ,  $pp=0.98$ ) and also the strongest partial correlation ( $r=0.47$ ,  $pp=0.87$ ) compared to brain size ( $r=0.27$ ,  $pp=0.75$ ) and body mass ( $r=0.035$ ,  $pp=0.51$ ). Hence, these results show that TRNP1 evolved under positive selection and that its rate of sequence evolution is linked strongest to the evolution of gyrification,

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

independent of the evolution of body mass. This indicates that TRNP1 evolved under directional selection because the degree of gyrification changed during mammalian evolution. 57 58



**Fig. 1. TRNP1 amino acid substitution rates and proliferative activity co-evolve with cortical folding in mammals**

**a**, Mammalian species for which brain size, GI measurements and TRNP1 coding sequences were available ( $n=31$ ). The majority of the included species are primates ( $n=18$ ). **b**, Scheme of the mouse TRNP1 protein (223 AAs) with intrinsically disordered regions (orange) and sites (red lines) subject to positive selection in mammals ( $\omega > 1$ ,  $pp > 0.95$ ; Extended Data Fig. 1b). The letter size of the depicted AAs represents the abundance of AAs at the positively selected sites. **c**, TRNP1 protein substitution rates ( $\omega$ ) correlate non-significantly with brain size ( $r = 0.64$ ,  $pp=0.89$ ). Grey horizontal lines represent 95% confidence intervals of  $\omega$ . **d**,  $\omega$  significantly correlates with GI ( $r = 0.69$ ,  $pp=0.98$ ). **e**, Six different TRNP1 orthologues were transfected into neural stem cells (NSCs) isolated from cerebral cortices of 14 day old mouse embryos and proliferation rates were assessed after 48 h using Ki67-immunostaining as proliferation marker in 7-12 independent biological replicates. **f**, Proliferation rate estimates according to TRNP1 orthologues with bars indicating standard errors of logistic regression and asterisks indicating the significance of pairwise comparisons (Tukey test,  $p$ -value:  $< 0.1$ ,  $* < 0.01$ ,  $** < 0.001$ ). **g**, Proliferation rates are a significant predictor for GI in the respective species (PGLS LRT:  $\chi^2=6.76$ ,  $df=1$ ,  $p$ -value  $< 0.01$ ;  $\beta=3.91 \pm 1.355$ ,  $R^2 = 0.68$ ,  $n=6$ ). Error bars indicate standard errors.

Next, we explored functional properties of TRNP1 that could be affected by these evolutionary changes. As previous studies had shown that TRNP1 transfection increases the proliferation of neural stem cells (NSCs)<sup>12,13</sup>, we compared this property among six TRNP1 orthologues, covering the observed range of GI and  $\omega$  (Fig. 1d). We quantified the proportion of transfected (GFP+) and proliferating (Ki67+) primary mouse NSCs from embryonic day 14 for each construct in 7-12 independent transfections. We confirmed that TRNP1 transfection does increase proliferation compared to a GFP-only control ( $p$ -value  $< 2 \times 10^{-16}$ ; Extended Data Fig. 1d). Remarkably, the proportion of proliferating cells was highest in cells transfected with dolphin TRNP1 followed by human, which was significantly higher than the two other primates, galago and macaque (Fig. 1f; Suppl. Tables 14,15). Notably, the effect of mouse TRNP1 (0.54) was slightly higher than expected given its  $\omega$  and GI, possibly caused by an increased oligomerisation with the endogenous mouse TRNP1<sup>13</sup>. Nevertheless, even when including the mouse TRNP1, the proliferative activity of TRNP1 is a significant predictor of the gyrification of its species of origin (Phylogenetic generalised least



bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

squares PGLS, Likelihood Ratio Test (LRT)  $p$ -value < 0.01,  $R^2 = 0.68$ ; Fig. 1g). These results provide strong evidence that the evolution of cortical folding is tightly linked to the evolution of the TRNP1 protein.

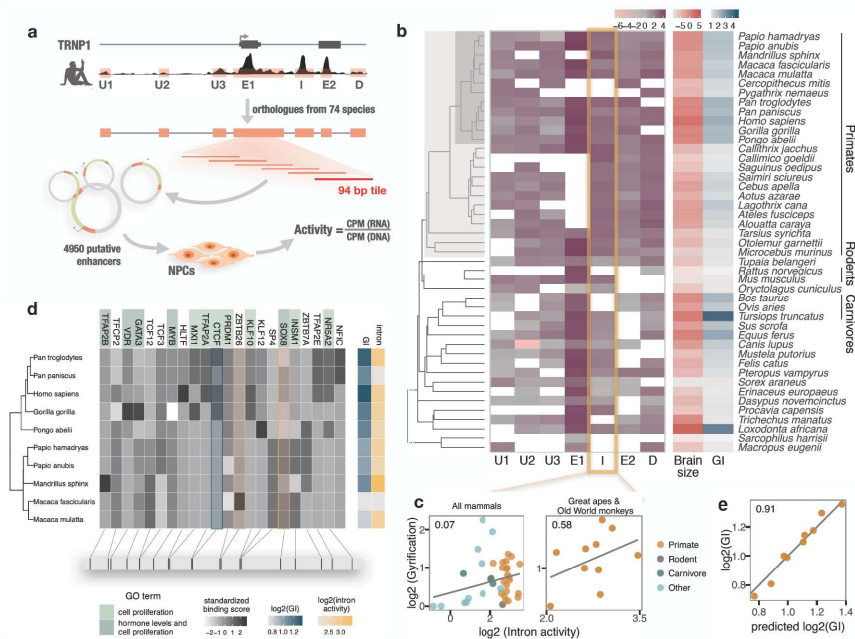
### Co-evolution of *TRNP1* regulation and cortical folding

We next investigated whether changes in *TRNP1* regulation may also be associated with the evolution of cortical folding and/or brain size by analyzing co-variation in the activity of *TRNP1* associated cis-regulatory elements (CREs). The recently developed massively parallel reporter assays (MPRAs) allow to measure regulatory activity for thousands of sequences at the same time. To this end, a library of putative regulatory sequences is cloned into a reporter vector and their activity is quantified simultaneously by the expression levels of element-specific barcodes<sup>19</sup>.

To identify putative CREs of *TRNP1*, we used DNase Hypersensitive Sites (DHS) from human fetal brain<sup>20</sup> and found three upstream CREs, the promoter-including exon 1, an intron CRE, one CRE overlapping the second exon, and one downstream CRE (Fig. 2a). The orthologous regions of the first five CREs were also identified as open chromatin in fetal brains of mice<sup>20</sup> (Extended Data Fig. 3), suggesting that those regions are likely to be CREs also in other mammals. We obtained additional orthologous sequences to the human CRE sequences either from genome databases or by sequencing yielding a total of 351 putative CREs in a panel of 75 mammalian species (Fig. 2b; Extended Data Fig. 3).

Due to limitations in the length of oligonucleotide synthesis, we cut each orthologous putative CRE into 94 base pairs highly overlapping fragments, resulting in 4950 sequence tiles, each synthesised together with a unique barcode sequence. From those, we successfully constructed a complex and unbiased (Extended Data Fig. 2a, 2b) lentiviral plasmid library containing at least 4251 (86%) CRE sequence tiles. Next, we stably transduced this library into neural progenitor cells (NPCs) derived from two humans and one macaque<sup>21</sup>. We calculated the activity per CRE sequence tile as the read-normalised reporter gene expression over the read-normalised input plasmid DNA (Fig. 2a). Finally, we use the per-tile activities (Extended Data Fig. 2c) to reconstruct the activities of the putative CREs from the 75 species. To this end, we summed all tile sequence activities for a given CRE while correcting for the built-in sequence overlap (Fig. 2b, Methods). CRE activities correlate well across cell lines and species (Pearson's  $r$  0.85-0.88; Extended Data Fig. 2d). The CREs covering exon1, the intron and downstream of *TRNP1* show the highest total activity across species and the upstream regions the lowest (Extended Data Fig. 4a). Next, we tested whether CRE activity can explain part of the variance in either brain size or GI across the 45 of the 75 mammalian species for which these phenotypes were available. None of the CREs showed any association with brain size (PGLS, LRT  $p$ -value > 0.1). In contrast, we found that the CRE activity of the intron sequence had a slight positive association with gyrification (PGLS, LRT  $p$ -value < 0.1, Fig. 2c left, Suppl. Table 16). These associations are much weaker than those of the TRNP1 protein evolution analysed above. Part of the reason might be that CREs have a much higher evolutionary turnover rate than coding sequences<sup>22,23</sup>. This also results in shorter orthologous sequences in species more distantly related to humans that defined the open chromatin regions (Extended Data Fig. 3). Therefore, we restricted our analysis to the catarrhine clade that encompasses Old World Monkeys, great apes and humans. Here, the association between intron CRE activity and GI becomes considerably stronger (PGLS, LRT  $p$ -value < 0.003, Fig. 2c right; Suppl. Table 17). Moreover, the intron CRE activity-GI association was consistently detectable across all three cell lines including the macaque NPCs (Suppl. Table 17).

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



**Fig. 2. Activity of a cis-regulatory element (CRE) of *TRNP1* correlates with cortical folding in catarrhines**

**a**, Experimental setup of the MPRA assay. Regulatory activity of 7 putative *TRNP1* CREs from 75 species were assayed in neural progenitor cells (NPC) derived from human and macaque induced pluripotent stem cells using lentiviral transduction. **b**, Log-transformed total regulatory activity per CRE in NPCs across species with available brain size and GI measurements ( $n=45$ ). **c**, Regulatory activity of the intron CRE is moderately associated with gyrification across mammals (PGLS, LRT  $p$ -value  $< 0.1$ ,  $R^2 = 0.07$ ,  $n=37$ ) and strongest across great apes and Old World Monkeys, i.e. catarrhines (PGLS, LRT  $p$ -value  $< 0.003$ ,  $R^2 = 0.58$ ,  $n=10$ ). **d**, Variation in binding scores of 22 transcription factors (TFs) across catarrhines. Shown are all TFs that are expressed in NPCs and have their binding motif enriched (motif weight  $\geq 1$ ) in the intron CRE sequence of catarrhines. Heatmaps indicate standardised binding scores (grey), GI values (blue) and intron CRE activities (yellow) from the respective species. TF background color indicates gene ontology assignment of the TFs to the 2 most significantly enriched biological processes (Fisher's  $p$ -value  $< 0.05$ ). The bottom panel indicates the spatial position of the top binding site (motif score  $> 3$ ) for each TF on the human sequence. Binding scores of 3 TFs (CTCF, ZBTB26, SOX8) are predictive for intron CRE activity, whereas only CTCF binding shows an association with the GI (PGLS, LRT  $p$ -value  $< 0.05$ ). **e**, A model combining *TRNP1* protein evolution rates ( $\omega$ ) and intron activity as predictors can explain GI across OWMs and great apes significantly better than  $\omega$  alone (PGLS, LRT  $p$ -value  $< 0.003$ ,  $R^2 = 0.91$ ,  $n=9$ ), indicating an additive, non-redundant effect of the *TRNP1* regulatory and structural evolution on the gyrification across these primate species.

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Reasoning that changes in CRE activities will most likely be mediated by their interactions with transcription factors (TF), we analysed the sequence evolution of putative TF binding sites. Among the 392 TFs that are expressed in our NPC lines, we identified 22 with an excess of binding sites<sup>24</sup> within the catarrhine intron CRE sequences (Fig. 2d, Extended Data Fig. 4b). In agreement with TRNP1 itself being involved in the regulation of cell proliferation<sup>12,13,25</sup>, these 22 TFs are predominantly involved in biological processes such as regulation of cell population proliferation and regulation of hormone levels (Fig. 2d; Suppl. Table 18). To further prioritise the 22 TFs, we used motif binding scores of these TFs for each of the 10 catarrhine intron CRE sequences to predict the observed intron activity in the MPRA assay and to predict the GI of the respective species. Out of the 22 TFs that are expressed and have TFBS enrichment within the intron CRE, the inter-species variation in motif binding scores of only 3 TFs (CTCF, ZBTB26, SOX8) is predictive for intron activity and only CTCF binding scores are predictive for GI (PGLS, LRT  $p$ -value < 0.05, Extended Data Fig. 4d, 4e). In summary, we find evidence that a higher activity of the intron CRE is correlated with gyrification in catarrhines, indicating that also regulatory changes in *TRNP1* contributed to the evolution of gyrification. To gauge the combined effects of structural and regulatory changes in TRNP1 to gyrification, we combined the standardised values of the estimated protein evolution rates of TRNP1 ( $\omega$ ) and intron CRE activity across catarrhines within the same PGLS framework (Fig. 2e). Although  $\omega$  of TRNP1 alone could already explain 75.5% of the variance in gyrification across these species, adding intron CRE activity significantly improved the model (PGLS, LRT  $p$ -value < 0.003, Suppl. Table 19), explaining in total 91% of the variance in GI across catarrhines. This suggests that in addition to the changes in the coding region, changes in regulatory regions of *TRNP1* also contribute to evolving more gyrified brains.

## Discussion

Here, we have shown that the rate of protein evolution of TRNP1 correlates with gyrification in mammals and that the activity of a regulatory element of *TRNP1* co-evolves with gyrification in catarrhines. Additionally, we have shown that also the proliferative activity of TRNP1 varies across mammals, as it would be predicted if TRNP1 was indeed the target for positive selection for a more gyrified brain. Hence, while previous experimental studies have speculated that TRNP1 could be important for the evolution of gyrification<sup>26</sup>, our analyses provide evidence that this is indeed the case.

Of note, the effect of structural changes appears stronger than the effect of regulatory changes. This is contrary to the notion that regulatory changes should be the more likely targets of selection as they are more cell-type specific<sup>27</sup> (but see also<sup>28</sup>). However, measures of regulatory activity are inherently less precise than counting amino acid changes, which will necessarily deflate the estimated association strength<sup>22,23</sup>. In any case, our analysis shows that evolution combined both regulatory and structural evolution to modulate and fine tune TRNP1 activity.

Moreover, our analyses generate specific hypotheses about the molecular mechanisms used to tune gyrification. They strongly suggest that an increased gyrification goes along with an increased proliferation activity in NSCs and suggests that amino acid changes in the disordered regions are responsible for this. Furthermore, we find that CTCF binding potential of the intron CRE is correlated with gyrification in catarrhines. This indicates a role for CTCF in regulating gyrification, in line with its regulatory role for several developmental processes<sup>29</sup>.

Finally, we think that our approach could serve as a blueprint to leverage the unique information stored in the evolutionary diversity among species. The fundamental principle of correlating genetic variants with phenotypes (GWAS) is a well established approach within populations and thus



bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

we believe that cross-species association studies (CSAS) will prove instrumental to understand  
 complex phenotypes. Genome sequences for essentially all vertebrates and eukaryotes are becoming  
 available<sup>2,30,31</sup>, making the availability of phenotype information the only limit to cross-species  
 association studies. On a molecular and cellular level, phenotyping of induced pluripotent stem cells  
 and their derivatives across many species would boost this approach<sup>32,33</sup>, further helping to tap the  
 potential of life's diversity to understand molecular mechanisms.

## References

1. Wright, S. H. Lander celebrates genome milestone in heavily attended talk. Accessed: 2020-5-22. <http://news.mit.edu/2001/lander-0228> (2001).
2. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. en. *Nature* **587**, 240–245 (2020).
3. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
4. Enard, W. Comparative genomics of brain size evolution. *Frontiers in human neuroscience* **8**, 345 (2014).
5. Borrell, V. & Calegari, F. Mechanisms of brain evolution: regulation of neural progenitor cell diversity and cell cycle length. *Neuroscience research* **86**, 14–24 (2014).
6. Borrell, V. & Götz, M. Role of radial glial cells in cerebral cortex folding. en. *Curr. Opin. Neurobiol.* **27**, 39–46 (2014).
7. Lewitus, E. *et al.* An adaptive threshold in mammalian neocortical evolution. *PLoS biology* **12**, e1002000 (2014).
8. Llinares-Benadero, C. & Borrell, V. Deconstructing cortical folding: genetic, cellular and mechanical determinants. *Nature Reviews Neuroscience* **20**, 161–176 (2019).
9. Heide, M. *et al.* Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. en. *Science* (2020).
10. Johnson, M. B. *et al.* Aspm knockout ferret reveals an evolutionary mechanism governing cerebral cortical size. en. *Nature* **556**, 370–375 (2018).
11. Montgomery, S. H. *et al.* Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. *Molecular biology and evolution* **28**, 625–638 (2010).
12. Stahl, R. *et al.* Trnp1 Regulates Expansion and Folding of the Mammalian Cerebral Cortex by Control of Radial Glial Fate. *Cell* **153**, 535–549. ISSN: 0092-8674 (2013).
13. Esgleas, M. *et al.* Trnp1 organizes diverse nuclear membrane-less compartments in neural stem cells. *The EMBO journal* **39**, e103373 (2020).
14. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421 (2009).
15. Löytynoja, A. in *Multiple sequence alignment methods* 155–170 (Springer, 2014).
16. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13**, 555–556 (1997).
17. Zilles, K. *et al.* Gyrfication in the cerebral cortex of primates. *Brain, Behavior and Evolution* **34**, 143–150 (1989).
18. Lartillot, N. & Poujol, R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution* **28**, 729–744 (2010).
19. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. en. *Genomics* **106**, 159–164 (2015).
20. Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature biotechnology* **28**, 1045 (2010).
21. Geuder, J. *et al.* A non-invasive method to generate induced pluripotent stem cells from primate urine. *bioRxiv*. eprint: <https://www.biorxiv.org/content/early/2020/08/12/2020.08.12.247619> (2020).
22. Danko, C. G. *et al.* Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. en. *Nat Ecol Evol* **2**, 537–548 (2018).
23. Berthelot, C. *et al.* Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. en. *Nat Ecol Evol* **2**, 152–163 (2018).

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

24. Frith, M. C. *et al.* Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic acids research* **31**, 3666–3668 (2003). 215  
216
25. Volpe, M. *et al.* trnp: A conserved mammalian gene encoding a nuclear protein that accelerates cell-cycle progression. en. *DNA Cell Biol.* **25**, 331–339 (2006). 217  
218
26. Martínez-Martínez, M. Á. *et al.* A restricted period for formation of outer subventricular zone defined by Cdh1 and Trnp1 levels. *Nature communications* **7**, 11812 (2016). 219  
220
27. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008). 221  
222
28. Hoekstra, H. E. & Coyne, J. A. The locus of evolution: evo devo and the genetics of adaptation. en. *Evolution* **61**, 995–1016 (2007). 223  
224
29. Arzate-Mejía, R. G. *et al.* Developing in 3D: the role of CTCF in cell differentiation. *Development* **145**, dev137729 (2018). 225  
226
30. Koepfli, K.-P. *et al.* The Genome 10K Project: a way forward. en. *Annu Rev Anim Biosci* **3**, 57–111 (2015). 227  
228
31. Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. en. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4325–4333 (2018). 229  
230
32. Enard, W. Functional primate genomics-leveraging the medical potential. *J. Mol. Med.* **90**, 471–480 (2012). 231  
232
33. Housman, G. & Gilad, Y. Prime time for primate functional genomics. en. *Curr. Opin. Genet. Dev.* **62**, 1–7 (2020). 233  
234

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

## Methods

### Sequencing of *TRNP1* for primate species

**Identification of cis-regulatory elements of *TRNP1*.** DNase hypersensitive sites in the proximity to *TRNP1* (25 kb upstream, 3 kb downstream) were identified in human fetal brain and mouse embryonic brain DNase-seq data sets downloaded from NCBI's Sequence Read Archive (Suppl. Table 2). Reads were mapped to human genome version hg19 and mouse genome version mm10 using NextGenMap with default parameters (NGM; v. 0.0.1)<sup>34</sup>. Peaks were identified with Hotspot v.4.0.0 using default parameters<sup>35</sup>. Overlapping peaks were merged, and the union per species was taken as putative cis-regulatory elements (CREs) of *TRNP1* (Suppl. Tables 3 & 4). The orthologous regions of human *TRNP1* DNase peaks in 49 mammalian species were identified with reciprocal best hit using BLAT (v. 35x1)<sup>36</sup>. Firstly, sequences of human *TRNP1* DNase peaks were extended by 50 bases down and upstream of the peak. Then, the best matching sequence per peak region were identified with BLAT using the following settings: -t=DNA -q=DNA -stepSize=5 -repMatch=2253 -minScore=0 -minIdentity=0 -extendThroughN. These sequences were aligned back to hg19 using the same settings as above. The resulting best matching hits were considered reciprocal best hits if they fell into the original human *TRNP1* CREs.

**Cross-species primer design for sequencing.** We sequenced *TRNP1* coding sequences in 6 primates for which reference genome assemblies were either unavailable or very sparse (see Suppl. Table 1). We used NCBI's tool Primer Blast<sup>37</sup> with the human *TRNP1* gene locus as a reference. Primer specificity was confirmed using the predicted templates in 12 other primate species available in Primer Blast. Primer pair 1 was used for sequencing library generation as it reliably worked for all 6 resequenced primate species (Suppl. Table 20).

In order to obtain *TRNP1* CREs for the other primate species, we designed primers using primux<sup>38</sup> based on the species with the best genome assemblies and were subsequently tested in closely related species in multiplexed PCR reactions. The species used as reference for primer design were *H.sapiens* (hg19), *P.paniscus* (PanPan1), *P.troglodytes* (panTro4), *G.gorilla* (gorGor3), *P.abelii* (ponAbe2), *N.leucogenys* (nomLeu3), *M.fascicularis* (MacFas), *M.mulatta* (rheMac3), *P.anubis* (papAnu2), *P.hamadryas* (papHam1), *C.sabaeus* (ChlSab), *C.jacchus* (calJac3) and *S.boliviensis* (saiBol1). A detailed list of designed primer pairs per CRE and reference genome can be found in Suppl. Table 21 and final pools of multiplexed primers per CRE and species can be found in Suppl. Table 22.

**Sequencing of target regions for primate species.** Primate gDNAs were obtained from Deutsches Primaten Zentrum, DKFZ and MPI Leipzig (see Suppl. Table 5). Depending on concentration, gDNAs were whole genome amplified prior to sequencing library preparation using GenomiPhi V2 Amplification Kit (Sigma). After amplification, gDNAs were cleaned up using SPRI beads (CleaNA). Both *TRNP1* coding regions and CREs were resequenced using a similar approach that included a touchdown PCR to amplify the target region followed by a ligation and Nextera XT library construction. The main difference between the two being the polymerase and the exact touchdown PCR conditions. For the CRE resequencing Q5 High-Fidelity DNA Polymerase (New England Biolabs) was used, while KAPA HiFi Polymerase (Roche) was used for the coding region. Both with their respective High GC buffer to account for the high GC content of the template. PCRs were performed as a 25  $\mu$ l reaction on varying amounts of Template DNA (usually 20 ng) and 0.05  $\mu$ M - 0.4  $\mu$ M per primer depending on the degree of multiplexing. The thermo cycling conditions were according to manufacturers instructions for the respective polymerase. While the CREs were amplified for 20 cycles in a touchdown PCR followed by 20 cycles of standard PCR, the coding



bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

region was amplified only in a touchdown PCR for 30 cycles. In the touchdown phase the annealing temperature was gradually decreased from 70 °C in the first cycle to 60 °C in the last cycle. Final elongation times were altered according to expected product length.

Blunt end ligation of PCR fragments and phosphorylation were done in the same reaction in T4 DNA ligation buffer with 2.5 U of T4 DNA ligase, 10 U of T4 Polynucleotide Kinase (Thermo Fisher Scientific) and up to 0.5 µg of input at 16°C overnight. The resulting products were SPRI bead purified. A Nextera XT DNA sample preparation kit (Illumina) was used to perform the library preparation from ligated PCR products according to the manufacturer’s protocol with the following modifications: the initial 55°C tagmentation time was increased from 5 to 10 minutes and the whole reaction was downscaled to 1/5th. Unique i5 and i7 primers were used for each library. Fragment size distributions were determined using capillary gel electrophoresis (Agilent Bioanalyzer 2100, DNA HS Kit) and libraries were pooled in equimolar ratios. For the missing CDS, sequencing was performed as 250 bases paired end with dual indexing on an Illumina MiSeq and the CRE libraries libraries were sequenced 50bp paired end on an Illumina HiSeq 1500.

**Assembly of sequenced regions.** Reads were demultiplexed using deML<sup>39</sup>. The resulting sequences per species were subsequently trimmed to remove PCR-handles using cutadapt (version 1.6)<sup>40</sup>. For sequence reconstruction, Trinity (version 2.0.6) in reference-guided mode was used<sup>41</sup>. The reference here is defined as the mapping of sequences to the closest reference genome with NGM (version 0.0.1)<sup>34</sup>. Furthermore, read normalisation was enabled and a minimal contig length of 500 was set. The sequence identity of the assembled contigs was validated by BLAT<sup>36</sup> alignment to the closest reference *TRNP1* as well as to the human *TRNP1*. The assembled sequence with the highest similarity and expected length was selected per species.

***TRNP1* coding sequence retrieval and alignment.** Human TRNP1 protein sequence was retrieved from UniProt database<sup>42</sup> under accession number Q6NT89. We used the human TRNP1 in a `tblastn`<sup>43</sup> search of genomes from 45 species specified in Suppl. Table 1 (R-package rBLAST version 0.99.2). The following additional arguments were specified:

`-soft masking false` — Turn off applying filtering locations as soft masks  
`-seg no` — Turn off masking low complexity sequences.

PRANK<sup>44</sup> (version 150803) was used to align TRNP1 coding sequences, using the mammalian tree from Bininda et al.<sup>45</sup> to guide the multiple alignment. Alignment was done using the default settings, specifying one additional parameter:

`-translate` — additionally translate the aligned nucleotide sequences to a protein.

### Evolutionary sequence analysis

**Identification of sites under positive selection.** Program `codeml` from PAML software<sup>46</sup> (version 4.8) was used to infer whether a significant proportion of TRNP1 protein sites evolve under positive selection across the phylogeny of 45 species. Site models M8 and M7 were compared<sup>47</sup>, that allow  $\omega$  to vary among sites across the phylogenetic tree, but not between branches. M7 and M8 are nested with M8 allowing for sites under positive selection with  $\omega_s$ . Likelihood ratio test (LRT) was used to compare these models. Naive Empirical Bayes (NEB) analysis was used to identify the specific sites under positive selection ( $\Pr(\omega > 1) > 0.95$ ). Additional general model settings:

`-tree topology from`<sup>45</sup>

`-seqtype = 1` — codon model

`-clock = 0` — no molecular clock

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

-aaDist = 0 — equal AA distances 322  
 -CodonFreq = 2 — codon frequency: from the average nucleotide frequencies at the third codon 323  
 positions (F3X4). 324  
 325

**Inferring correlated evolution using Coevol.** Coevol<sup>48</sup> (version 1.4) was utilised to infer the 326  
 covariance between TRNP1 evolutionary rate  $\omega$  and three morphological traits (brain size, GI and 327  
 body mass) across species (Suppl. Table 7). `dsom` command was used to activate the codon model, 328  
 in which the two a priori independent variables are dS and  $\omega$ . For each model, the MCMC was 329  
 run for at least 10,000 cycles, using the first 1,000 as burn-in and two runs were performed to 330  
 examine global convergence. `tracecomp` was used to access the relative differences between runs ( $d = 2|\mu_1 - \mu_2|/(\sigma_1 + \sigma_2)$ ) where  $\mu$  are the means and  $\sigma$  are standard deviations of each parameter 331  
 for the two chains) and mixing diagnostics (effective sample size) between the runs. All parameters 332  
 have a relative difference  $< 0.1$  and effective size  $> 300$ , which indicates good convergence and 333  
 quantitatively reliable runs<sup>48</sup>. 334  
 335

We report posterior probabilities ( $pp$ ), the marginal and partial correlations of the full model (Suppl. 336  
 Table 11) and the separate models where including only either one of the three traits (Suppl. Table 10). 337  
 The posterior probabilities for a negative correlation are given by  $1 - pp$ . These were back-calculated 338  
 to make them directly comparable, independently of the correlation direction, i.e. higher  $pp$  means 339  
 more statistical support for the respective correlation. 340

#### Proliferation assay 341

**Plasmids.** The six *TRNP1* orthologous sequences containing the restriction sites BamHI and 342  
 XhoI were synthesized by GeneScript ([www.genscript.com](http://www.genscript.com)). All plasmids for expression were first 343  
 cloned into a pENTR1a gateway plasmid described in Stahl et al., 2013<sup>49</sup> and then into a Gateway 344  
 (Invitrogen) form of pCAG-GFP (kind gift of Paolo Malatesta). The gateway LR-reaction system 345  
 was used to then sub-clone the different TRNP1 forms into the pCAG destination vectors. 346

**Primary cerebral cortex cultures and transfection.** Cerebral cortices were dissected removing 347  
 the ganglionic eminence, the olfactory bulb, the hippocampal anlage and the meninges and cells 348  
 were mechanically dissociated with a fire polish Pasteur pipette. Cells were then seeded onto poly-D- 349  
 lysine (PDL)-coated glass coverslips in DMEM-GlutaMAX with 10% FCS (Life Technologies) and 350  
 cultured at 37°C in a 5% CO<sub>2</sub> incubator. Plasmids were transfected with Lipofectamine 2000 (Life 351  
 technologies) according to manufacturer's instruction 2h after seeding the cells onto PDL coated 352  
 coverslips. One day later cells were washed with phosphate buffered saline (PBS) and then fixed in 353  
 4% Paraformaldehyde (PFA) in PBS and processed for immunostaining. 354

**Immunostaining.** Cells plated on poly-D-lysine coated glass coverslips were blocked with 2% BSA, 355  
 0.5% Triton-X (in PBS) for 1 hour prior to immunostaining. Primary antibodies (GFP and Ki67) 356  
 were applied in blocking solution overnight at 4°C. Fluorescent secondary antibodies were applied in 357  
 blocking solution for 1 hour at room temperature. DAPI (4',6-Diamidin-2-phenylindol, Sigma) was 358  
 used to visualize nuclei. Stained cells were mounted in Aqua Polymount (Polysciences). All secondary 359  
 antibodies were purchased from Life Technologies. Images were taken using an epifluorescence 360  
 microscope (Zeiss, Axio ImagerM2) equipped with a 20X/ 0.8 N.A and 63X/1.25 N.A. oil immersion 361  
 objectives. Post image processing with regard to brightness and contrast was carried out where 362  
 appropriate to improve visualization, in a pairwise manner. 363



bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

**Proliferation rate calculation using logistic regression.** The proportion of successfully transfected cells that proliferate under each condition (Ki67-positive/GFP-positive) was modelled using logistic regression (R-package `stats` (version 4.0.3), `glm` function) with logit link function  $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ , for  $0 \leq p \leq 1$ , where  $p$  is the probability of success. The absolute number of GFP-positive cells were added as weights. Model selection was done using LRT within `anova` function from `stats`. Adding the donor mouse as a batch improved the models (Suppl. Tables 12, 13). To back-calculate the absolute proliferation probability (i.e., rate) under each condition, intercept of the respective model was set to zero and the inverse logit function  $\frac{e^{\beta_i X_i}}{1+e^{\beta_i X_i}}$  was used, where  $i$  indicates condition (Suppl. Table 14). Two-sided multiple comparisons of means between the conditions of interest were performed using `glht` function (Tukey test, user-defined contrasts) from R package `multcomp` (version 1.4-13) (Suppl. Table 15).

**Phylogenetic modeling of proliferation rates using generalized least squares (PGLS).** The association between the induced proliferation rates for each TRNP1 orthologue and the GI of the respective species was analysed using generalised least squares (R-package `nlme`, version 3.1-143), while correcting for the expected correlation structure due to phylogenetic relation between the species. The expected correlation matrix for the continuous trait was generated using a Brownian motion<sup>50,18</sup> (`ape` (version 5.4), using function `corBrownian`) on the mammalian phylogeny from Bininda et al. (2007)<sup>45</sup> adding the missing species (Fig 1a). The full model was compared to a null model using the likelihood ratio test (LRT). Residual  $R^2$  values were calculated using `R2.resid` function from R package `RR2` (version 1.0.2).

#### Massively Parallel Report Assay (MPRA)

**MPRA library design.** *TRNP1* CRE sequences identified in human fetal brain, mouse embryonic brain as well as orthologous regions in 73 mammalian species were considered for the Massively Parallel Reporter Assay (MPRA). In total 351 sequences were included where a sliding window per sequence entry was applied moving by 40 bases for the sequences that are longer than 94 bases, resulting in 4,950 oligonucleotide sequences which are flanked upstream by a first primer site (ACTGGCCGCTTCACTG), downstream by a KpnI/XbaI restriction cut site (GGTACCTCTAGA), a 10 base long barcode sequence as well as a second primer site (AGATCGGAAGAGCGTCG). Barcode tag sequences were specifically designed so that they contain all four nucleotides at least once, do not contain stretches of four identical nucleotides, do not contain microRNA seed sequences (retrieved from microRNA Bioconductor R package version 1.28.0) and do not contain restriction cut site sequences for KpnI nor XbaI (5'-GGTACG-3', 3'-CCATGG-5').

**MPRA plasmid library construction.** We modified the original protocol by Melnikov et al.<sup>52</sup>: We used a lentiviral delivery system as previously described<sup>53</sup> and introduced green fluorescent protein (GFP) instead of nano luciferase. All DNA purification and clean up steps were performed using SPRI beads unless stated otherwise, all plasmid DNA isolations were done using the standard protocol of a column-based kit (PureYield Plasmid Midiprep System, Promega). Primer sequences and plasmids used in the MPRA can be found in Suppl. Table 23 and 24 respectively. The *TRNP1* enhancer oligonucleotide library was synthesised on an oligo array by Custom Array. In a first step the single stranded oligos were double stranded and restriction sites flanking these oligos were introduced via PCR and subsequently used for directional cloning. Emulsion PCR using the commercially available Micellula DNA Emulsion & Purification Kit (roboklon) was performed in this step to avoid loss of individual variants and ensure unbiased amplification. Restriction digest using SfiI (New England Biolabs) and cloning of the variant library into the pMPRA1 plasmid (Addgene, #49349)

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

was performed according to the original protocol. To ensure maximum complexity this first step was carried out twice and the initial emulsion PCR was performed in quadruplicates each time. Transformation of the ligation products was performed in triplicates where one fourth of the ligation product was used to transform 50  $\mu$ l of chemically competent *E. coli* (5-alpha High Efficiency, New England Biolabs). Next a constant sequence, transcribed under the influence of the library, needed to be introduced into the generated plasmid pool. In the original publication a nano luciferase with a minimal promoter is introduced however we decided to use a GFP reporter here. To this end we used pNL3.1 and replaced the nano luciferase with an EGFP ORF using Gibson assembly. The resulting plasmid carried the same restriction sites as the ones used in the original publication and hence cloning was performed as described previously<sup>52</sup>. Electroporation into electro competent *E. coli* (10-beta, New England Biolabs) was performed to maximise transformation efficiency. All transformations were carried out in triplicates that were pooled and grown in 150 ml liquid cultures. For the final cloning step, the enhancer library including GFP and the minimal promoter were inserted into a lentiviral backbone (pMPRALenti1, Addgene #61600). Both plasmids were digested with SfiI (New England Biolabs) to allow for directional cloning of the whole construct. The lentiviral backbone pMPRALenti1 was treated similarly with the addition of shrimp alkaline phosphatase (rSAP, New England Biolabs). Ligation was performed with a 3:1 molar ratio of backbone to insert using 1 U of T4 DNA Ligase (Thermo Fisher Scientific) and 1 mM ATP for 3 h at 20°C. Ligation reactions were cleaned up and used to transform electrocompetent *E. coli* (10-beta, New England Biolabs). All transformations were pooled and used to inoculate a 200 ml bacterial culture and plasmids were isolated as before.

**MPRA lentiviral particle production.** Lentiviral particles were produced according to standard methods in HEK 293T cells<sup>54</sup>. The MPRA library was co-transfected with third generation lentiviral plasmids carrying the env, rev, gag and pol genes (pMDLg/pRRE, pRSV-Rev; Addgene #12251, #12253) as well as the VSV glycoprotein (pMD2.G, Addgene #12259) using Lipofectamine 3000. The lentiviral particle containing supernatant was harvested 48 hrs post transfection and filtered using 0.45  $\mu$ m PES syringe filters. Viral titer was determined by infecting Neuro-2A cells (ATCC CCL-131) and counting GFP positive cells. To this end, N2A cells were infected with a 50/50 volume ratio of viral supernatant to cell suspension with addition of 8  $\mu$ g/ml Polybrene. Cells were exposed to the lentiviral particles for 24 hrs until medium was exchanged. After additional 48 hrs, infected cells were positively selected using Blasticidin.

**Culture of neural progenitor cells.** Neural progenitor cells were cultured on Geltrex (Thermo Fisher Scientific) in DMEM F12 (Fisher scientific) supplemented with 2 mM GlutaMax-I (Fisher Scientific), 20 ng/ $\mu$ l bFGF (Peprotech), 20 ng/ $\mu$ l hEGF (Miltenyi Biotec), 2% B-27 Supplement (50X) minus Vitamin A (Gibco), 1% N2 Supplement 100X (Gibco), 200 $\mu$ M L-Ascorbic acid 2-phosphate (Sigma) and 100  $\mu$ g/ml penicillin-streptomycin with medium change every second day. For passaging, NPCs were washed with PBS and then incubated with TrypLE Select (Thermo Fisher Scientific) for 5 min at 37°C. Culture medium was added and cells were centrifuged at 200  $\times$  g for 5 min. Supernatant was replaced by fresh culture medium and cells were transferred to a new Geltrex coated dish. The cells were split every two to three days in a ratio of 1:3.

**MPRA lentiviral transduction.** The transduction of the MPRA library was performed in triplicates on two *Homo sapiens* and one *Macaca fascicularis* NPC lines<sup>55</sup> (see Suppl. Table 6).  $2.5 \times 10^5$  NPCs per line and replicate were dissociated, dissolved in 500  $\mu$ l cell culture medium containing 8  $\mu$ g/ml Polybrene and incubated with virus at MOI 12.7 for 1 h at 37°C in suspension<sup>56</sup>.



bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Thereafter cells were seeded on Geltrex and cultured as described above. Virus containing medium was replaced the next day and cells were cultured for additional 24 hrs. Cells were collected, lysed in 100  $\mu$ l TRI reagent and frozen at -80°C.

**MPRA sequencing library generation.** As input control for RNA expression, DNA amplicon libraries were constructed using 100 - 500 pg plasmid DNA. Library preparation was performed in two subsequent PCRs. A first PCR termed Adapter PCR introduced the 5' transposase mosaic end, this was used in the second PCR (Index PCR) to add a library specific index sequence and Illumina Flow cell adapters. The Adapter PCR was performed in triplicates using DreamTaq polymerase (Thermo Fisher Scientific). PCR products were cleaned up using SPRI beads (1/1 ratio) and quantified. 1-5 ng were subjected to the Index PCR using Q5 polymerase. After cleaning up libraries using SPRI beads (2/1 ratio), amplified DNA was quantified and quality control was performed using capillary gel electrophoresis (Agilent Bioanalyzer 2100). Total RNA from NPCs was extracted using the Direct-zol RNA Microprep Kit (Zymo Research). 500ng of RNA were subjected to reverse transcription using Maxima H Minus RT (Thermo Fisher Scientific) with Oligo-dT primers. 50 ng of cDNA were used for library preparation as described for plasmid DNA, with the alteration that Q5 DNA polymerase was used in both PCRs. 15-20 ng of the Adapter PCR product were subjected to the second library PCR and further treated as described for plasmid libraries. Plasmid and cDNA libraries were pooled and quality was evaluated using capillary gel electrophoresis (Agilent Bioanalyzer 2100). Sequencing was performed on an Illumina HiSeq 1500 instrument using a single-index, 50bp, paired-end protocol.

**MPRA data processing and analysis.** MPRA reads were demultiplexed with deML<sup>39</sup> using i5 and i7 adapter indices from Illumina. Next, we removed barcodes with low sequence quality, requiring a minimum Phred quality score of 10 for all bases of the barcode (zUMIs, fqfilter.pl script<sup>57</sup>). Furthermore, we removed reads that had mismatches to the constant region (the first 20 bases of the GFP sequence TCTAGAGTCGCGGCCTACT). The remaining reads that matched one of the known CRE-tile barcodes were tallied up resulting in a count table. Next, we filtered out CRE tiles that had been detected in only one of the 3 input plasmid library replicates (4202/4950). Counts per million (CPM) were calculated per CRE tile per library (median counts:  $\sim$  900k range: 590k-1,050k). Macaque replicate 3 was excluded due its unusually low correlation with the other samples (Pearson's  $r$ :  $\bar{r} - 1.5 \times \sigma_r$ ). The final regulatory activity for each CRE tile per cell line was calculated as:

$$a_i = \frac{\text{median}(CPM_i)}{\text{median}(CPM_i)_p}, \quad (1)$$

where  $a$  is regulatory activity,  $i$  indicates CRE tile and  $p$  is the input plasmid library. Median was calculated across the replicates from each cell line.

Given that each tile was overlapping with two other tiles upstream and two downstream, we calculated the total regulatory activity per CRE region in a coverage-sensitive manner, i.e. for each position in the original sequence mean per-bp-activity across the detected tiles covering it was calculated. The final CRE region activity is the sum across all base positions.

$$a_r = \sum_{b=1}^k \frac{1}{n} \sum_{i=1}^n \frac{a_i}{l_i}, \quad (2)$$

where  $a_r$  is regulatory activity of CRE region  $r$ ,  $b = 1, \dots, k$  is the base position of region  $r$ ,  $i, \dots, n$  are tiles overlapping the position  $b$ ,  $a_i$  is tile activity from equation 1 and  $l_i$  is tile length. CRE activity and brain phenotypes were associated with one another using PGLS analysis (see above). The number of species varied for each phenotype-CRE pair (brain size: min. 37 for exon1, max.

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

48 for intron and downstream regions; GI: min. 32 for exon2, max. 37 for intron), therefore the activity of each of the seven CRE regions was used separately to predict either GI or brain size of the respective species.

#### Combining protein evolution rates and intron activity to predict GI across catharrines

PGLS model fits were compared either including only  $\omega$  of TRNP1 protein from Coevol<sup>48</sup> or including  $\omega$  and intron CRE activity as predictors. For this, the standardised values of either measurement were used, calculated as  $\frac{x_i - \bar{x}}{\sigma}$ , where  $x_i$  is each observed value,  $\bar{x}$  is the mean and  $\sigma$  is the standard deviation.

#### Transcription Factor analysis

**RNA-seq library generation** RNA sequencing was performed using the prime-seq method, a bulk derivative of the single cell RNA-seq method SCRIB-seq<sup>58</sup>. The major features of this method are early pooling enabled by the introduction of a cell barcode and a unique molecular identifier (UMI) in the reverse transcription reaction followed by full length cDNA amplification and enrichment of 3 prime ends in the library preparation. The full prime-seq protocol including primer sequences can be found at protocols.io (<https://www.protocols.io/view/prime-seq-s9veh66>). Here we used 10 ng of the isolated RNA from the MPRA experiment and subjected it to the prime-seq protocol with minor modifications. As sequencing of the MPRA transcripts contained in the RNA of the infected NPCs, may lead to problems in sequencing due to a duplicated read start, we determined the amount of contamination caused by MPRA transcripts in the transcriptome library. Using an additional primer (Suppl. Table 23) in the pre-amplification which generates a small MPRA amplicon, followed by a size selection we found the contamination to be negligible and proceeded with the standard prime-seq protocol. After reverse transcription all samples were pooled, the pool was cleaned up, Exonuclease 1 (Thermo Fisher Scientific) digested and finally subjected to cDNA amplification using Kapa HiFi polymerase (Roche). Nextera XT (Illumina) library preparation was performed in triplicates of 0.8 ng of amplified cDNA each. Instead of an i5 index primer a custom 3 prime enrichment primer was added to the Library PCR reaction and annealing temperature was increased to 62°C. The replicates of the sequencing library were pooled and size selected (300 -900 bp) using an 2% agarose gel followed by gel extraction to ensure optimal sequencing quality. Finally the size distribution and molarity of the library was measured using capillary gel electrophoresis (Agilent Bioanalyzer 2100). Sequencing was performed on an Illumina HiSeq 1500 instrument with an unbalanced paired end layout, where read 1 was 16 base pair long and read 2 was 50 base pair long, additionally an 8 base pair index read was performed.

**RNA-seq data processing** Bulk RNA-seq data was generated from the same 9 samples (3 cell lines, 3 biological replicates each) that were transduced and assayed in the MPRA. This was done to detect which TFs are expressed in the assayed cell lines and might be responsible for the observed intron CRE activity. Raw read fastq files were pre-processed using zUMIs<sup>57</sup> together with STAR<sup>59</sup> to generate expression count tables for barcoded UMI data. Reads were mapped to human reference genome (hg38, Ensembl annotation GRCh38.84). Further filtering was applied keeping genes that were detected in at least 7/9 samples and had on average more than 7 counts. For further analysis, we used normalised and variance stabilised expression estimates as provided by DESeq2<sup>60</sup>.

**TFBS motif analysis on the intron CRE sequence** TF Position Frequency Matrices (PFM) were retrieved from JASPAR CORE 2020<sup>61</sup>, including only non-redundant vertebrate motifs (746 in

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

total). These were filtered for the expression in our NPC RNA-seq data, leaving 392 TFs with 462 motifs in total).

A Hidden Markov Model (HMM)-based program Cluster-Buster<sup>62</sup> (compiled on Jun 13 2019) was used to infer the enriched TF binding motifs on the intron sequence. First, the auxiliary program Cluster-Trainer was used to find the optimal gap parameter between motifs of the same cluster and to obtain weights for each TF based on their motif abundance per kb across catharrine intron CREs from 10 species with available GI measurements. Weights for each motif suggested by Cluster-Trainer were supplied to Cluster-Buster that we used to find clusters of regulatory binding sites and to infer the enrichment score for each motif on each intron sequence. The program was run with the following parameters:

-g3 — gap parameter suggested by Cluster-Trainer  
 -c5 — cluster score threshold  
 -m3 — motif score threshold.

To identify the most likely regulators of *TRNP1* that bind to its intron sequence and might influence the evolution of gyrification, we filtered for the motifs that were most abundant across the intron sequences (Cluster-Trainer weights >1). These motifs were distinct from one another (mean pairwise distance 0.72, Extended Data Fig. 4c). Gene-set enrichment analysis contrasting the TFs with the highest binding potential with the other expressed TFs was conducted using the Bioconductor-package topGO<sup>63</sup> (version 2.40.0) (Suppl. Table 18).

PGLS model was applied as previously described, using Cluster-Buster binding scores across catharrine intron CRE sequences as predictors and predicting either intron activity or GI from the respective species. The relevance of the three TFs that were associated with intron activity was then tested using an additive model and comparing the model likelihoods with reduced models where either of these were dropped.

34. Sedlazeck, F. J. *et al.* NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *en. Bioinformatics* **29**, 2790–2791. ISSN: 1367-4803, 1367-4811 (2013).
35. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268. ISSN: 1061-4036, 1546-1718 (2011).
36. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664. ISSN: 1088-9051, 1549-5469 (2002).
37. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *en. BMC Bioinformatics* **13**, 134. ISSN: 1471-2105 (2012).
38. Hysom, D. A. *et al.* Skip the alignment: degenerate, multiplex primer and probe design using K-mer matching instead of alignments. *en. PLoS One* **7**, e34560 (2012).
39. Renaud, G. *et al.* deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics* **31**, 770–772 (2015).
40. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *en. EMBnet.journal* **17**, 10–12. ISSN: 2226-6089, 2226-6089 (2011).
41. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652. ISSN: 1087-0156 (2011).
42. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **47**, D506–D515 (2019).
43. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421 (2009).
44. Löytynoja, A. in *Multiple sequence alignment methods* 155–170 (Springer, 2014).
45. Bininda-Emonds, O. R. *et al.* The delayed rise of present-day mammals. *Nature* **446**, 507 (2007).
46. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13**, 555–556 (1997).
47. Yang, Z. *et al.* Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449 (2000).



bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

48. Lartillot, N. & Poujol, R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution* **28**, 729–744 (2010). 582
49. Stahl, R. *et al.* Trnp1 Regulates Expansion and Folding of the Mammalian Cerebral Cortex by Control of Radial Glial Fate. *Cell* **153**, 535–549. ISSN: 0092-8674 (2013). 583
50. Felsenstein, J. Phylogenies and the comparative method. *The American Naturalist* **125**, 1–15 (1985). 584
51. Martins, E. P. & Hansen, T. F. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist* **149**, 646–667 (1997). 585
52. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. en. *Nat. Biotechnol.* **30**, 271–277 (2012). 586
53. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. en. *Genome Res.* **27**, 38–52 (2017). 587
54. Dull, T. *et al.* A third-generation lentivirus vector with a conditional packaging system. en. *J. Virol.* **72**, 8463–8471 (1998). 588
55. Gender, J. *et al.* A non-invasive method to generate induced pluripotent stem cells from primate urine. *bioRxiv*. eprint: <https://www.biorxiv.org/content/early/2020/08/12/2020.08.12.247619> (2020). 589
56. Nakai, R. *et al.* Derivation of induced pluripotent stem cells in Japanese macaque (*Macaca fuscata*). en. *Sci. Rep.* **8**, 12187 (2018). 590
57. Parekh, S. *et al.* zUMIs—a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, giy059 (2018). 591
58. Bagnoli, J. W. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCR-seq. *Nat. Commun.* **9** (2018). 592
59. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). 593
60. Love, M. I. *et al.* Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014). 594
61. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research* **48**, D87–D92 (2020). 595
62. Frith, M. C. *et al.* Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic acids research* **31**, 3666–3668 (2003). 596
63. Alexa, A. & Rahnenführer, J. Gene set enrichment analysis with topGO. *Bioconductor Improv* **27** (2009). 597
64. Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature biotechnology* **28**, 1045 (2010). 598

## Data Availability 618

The RNA-seq data used in this manuscript are publicly available at Array Express E-MTAB-9951. 619  
The MPRA data are available at Array Express under accession number E-MTAB-9952. Additional 620  
primate sequences for TRNP1 are available at GenBank (MW373535 - MW373709). 621

## Code Availability 622

A compendium containing processing scripts and detailed instructions to reproduce the analysis for this 623  
manuscript is available from the following GitHub repository: [https://github.com/Hellmann-Lab/](https://github.com/Hellmann-Lab/Co-evolution-TRNP1-and-GI) 624  
Co-evolution-TRNP1-and-GI. 625

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

### Author Contributions

626

M.G. proposed the project and W.E. and I.H. conceived the approaches of this study. B.V. designed all initial sequence acquisitions. L.W., M.H. D.R. and J.R. conducted the MPRA assay. M.E. designed and conducted the proliferation assay. J.R., J.G. and M.O. were responsible for all primate cell culture work. Z.K. collected, integrated and analysed all data. M.G. and M.E. provided expertise on Trnp1 function throughout the study. W.E. and I.H. supervised the work and provided guidance in data analysis. Z.K., I.H., and W.E. wrote the manuscript. All authors read, corrected and approved the final manuscript.

627

628

629

630

631

632

633

### Acknowledgements

634

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through LMUexcellent, SFB1243 (Subproject A14/A15 to W.E. and I.H., respectively), DFG grant HE 7669/1-1 (to I.H.) and the advanced ERC grant ChroNeuroRepair (to M.G.). We want to thank Christian Roos from the German Primate Center for providing genomic DNA from primates, project students Gunnar Kuut and Fatih Sarigoel for helping to generate TRNP1 orthologous sequences, Nikola Vuković for helping to establish the MPRA assay, Reza Rifat for helping with the proliferation assays and Christoph Neumayr for helping in data analysis.

635

636

637

638

639

640

641

## Supplementary Information

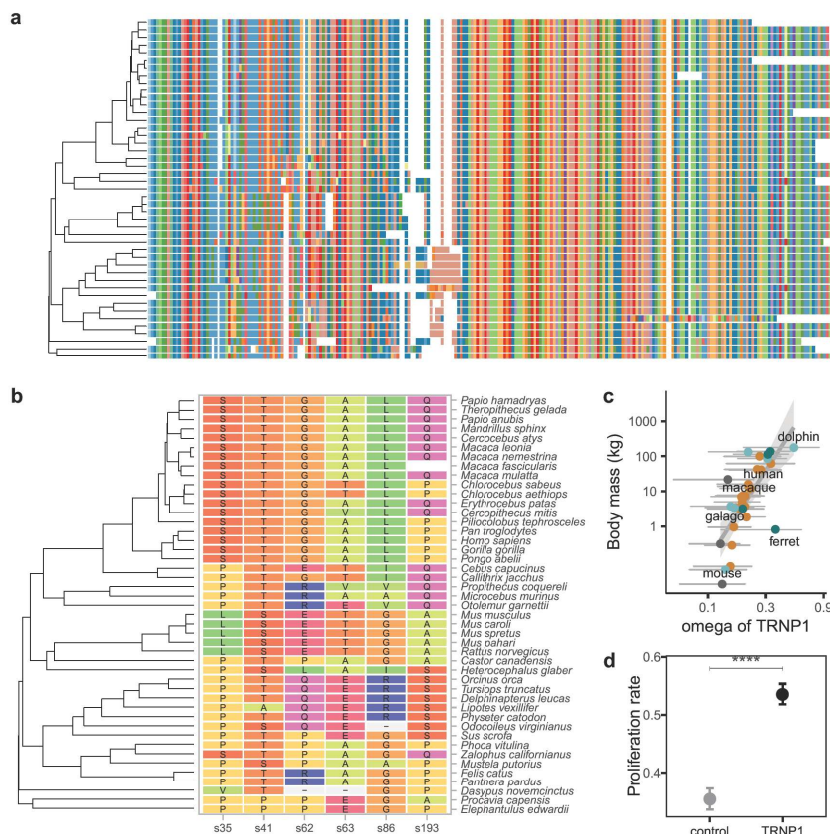


## 5.5 TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals 299

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

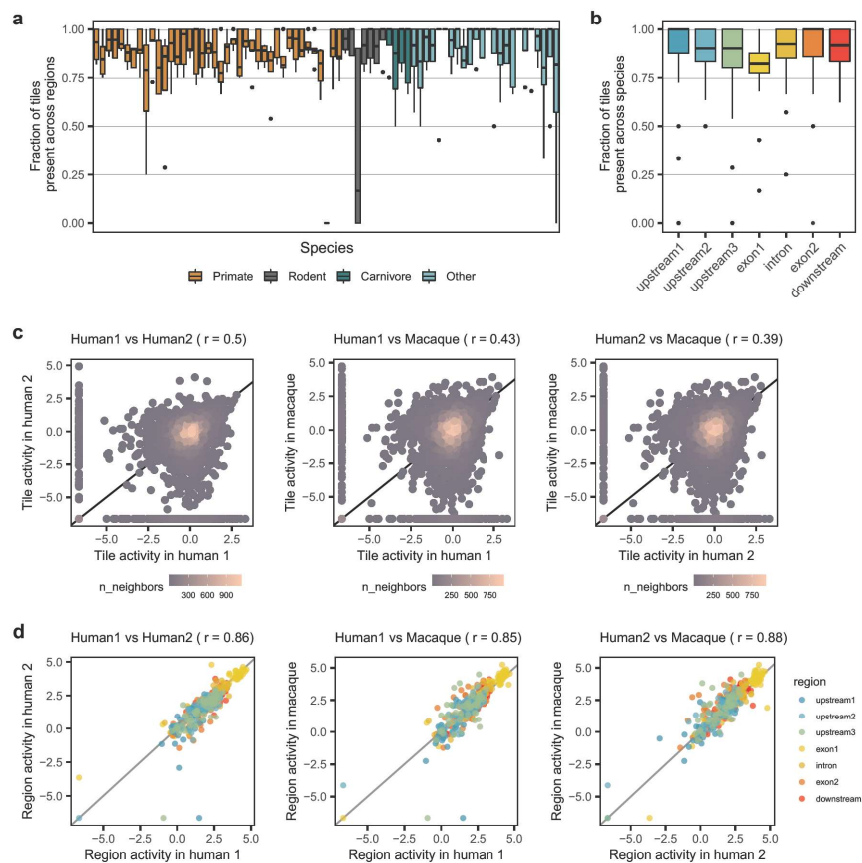
### Extended Data Figures

642



**Extended Data Fig. 1.** TRNP1 protein-coding sequence analysis. **a**, Multiple alignment of 45 TRNP1 coding sequences (97.4% completeness) using phylogeny-aware aligner PRANK[44]. The alignment is 744 bases long, which translates to 248 amino acids (AAs). For comparison: human TRNP1 coding sequence is 227 AA long, whereas mouse - 223 AAs. **b**, Sites under positive selection across the phylogenetic tree according to PAML[46] M8 model (in total 8.2% of sites with  $\omega > 1$ , LRT,  $p$ -value < 0.001). The depicted sites had a posterior probability  $\Pr(\omega > 1) > 0.95$  according to Naive Empirical Bayes analysis. Colours of the amino acids indicate their relatedness in biochemical properties. Sites with light-grey background and a dash indicate gaps/indels, while a white bar indicates one missing AA. **c**,  $\omega$  and body mass correlate moderately across mammal species ( $\omega \sim \text{BM}$ :  $r=0.55$ ,  $pp=0.9$ ). **d**, The overall effect of TRNP1 on proliferation rates in primary mouse NSCs. Proliferation induced by all 6 TRNP1 orthologues combined was compared to the control transfected with a plasmid carrying only GFP, but no TRNP1 coding sequence. TRNP1 presence in NSCs significantly increases the proliferation rates (TRNP1:  $0.54 (\pm 0.018)$ , control:  $0.35 (\pm 0.019)$ , Tukey test  $p$ -value <  $2e - 16$ ,  $df=68$ ).  $n=7$  for galago, macaque and dolphin,  $n=12$  for mouse, ferret, human and GFP-control.

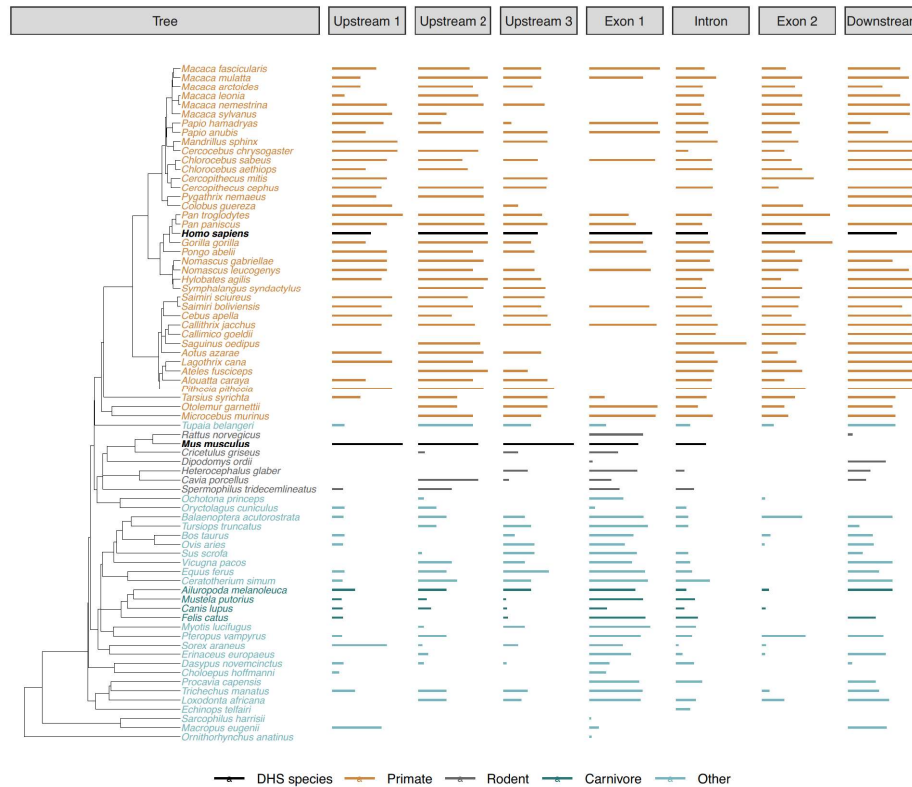
bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



**Extended Data Fig. 2.** Analysis of massively parallel reporter assay (MPRA) data. **a**, Fraction of the detected CRE tiles in the plasmid library per species across regions. The detection rates are unbiased and uniformly distributed across species and clades with only one extreme outlier *Dipodomys ordii*. This is mainly due to the fact that out of 3 total orthologous regions identified in the genome of this species, upstream 3 consisted of only 1 - uncaptured - tile. **b**, Fraction of the detected CRE tiles in the plasmid library per region across species. **a,b** Each box represents the median and first and third quartiles 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. Individual points indicate outliers. **c**, Pairwise correlation of the log<sub>2</sub>-transformed CRE tile activity between the three transduced cell lines: human 1, human 2 and macaque. Pearson's  $r$  is specified in the brackets of figure titles. **d**, Pairwise correlation of the log<sub>2</sub>-transformed summarized activity per CRE region between cell lines. Pearson's  $r$  is specified in the brackets of figure titles.

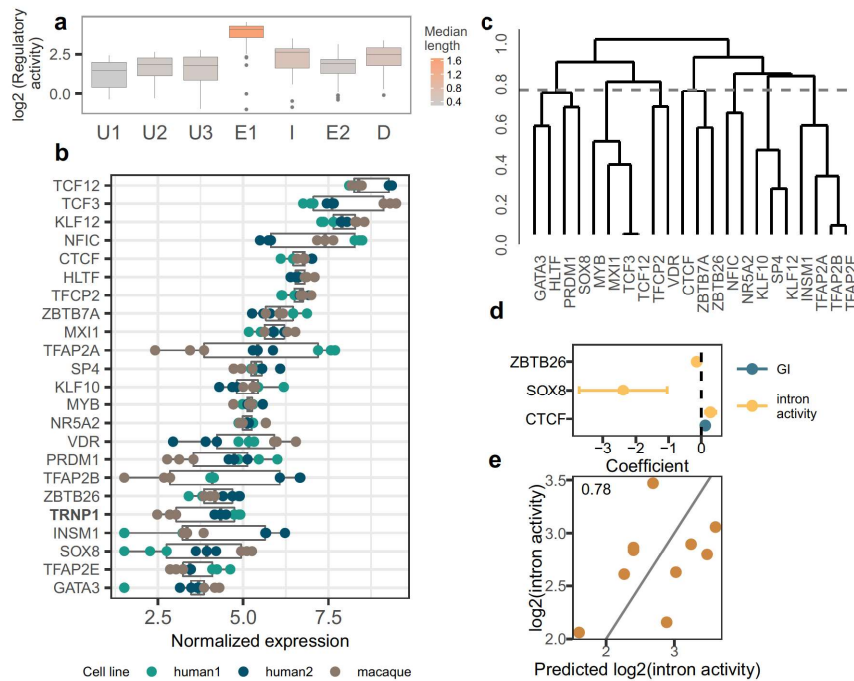
## 5.5 TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals 301

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).



**Extended Data Fig. 3.** Length of the covered CRE sequences in the MPRA library across the tree. Species for which the regions were inferred based on DNase-Hypersensitive Sites (DHS) from embryonic brain [64] are marked in bold and black: human (*Homo sapiens*) and mouse (*Mus musculus*). These species do not show extreme differences in length compared to others (human: 5/7, mouse: 3/5 regions within the 10% and 90% quantiles). The orthologous CRE sequence length differs strongly between primate and non-primate species, being in average 1.8 to 2.8 times longer in the primate species than in the other mammals.

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



**Extended Data Fig. 4.** Intron-binding transcription factor analysis. **a**, Total activity per CRE across species. Exon1 (E1), intron (I) and the downstream (D) regions are more active and longer than other regions. Each box represents the median and first and third quartiles 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. Individual points indicate outliers. **b**, Variance-stabilized expression of the 22 transcription factors with enriched binding sites on the intron CRE region. *TRNP1* could be consistently detected in all replicates, meaning that the TFs inducing its expression are present in this cellular system. **c**, Hierarchical clustering (average linkage) of TF Position Frequency matrices retrieved from JASPAR2020[61] for the 22 intron-enriched TFs. Dashed grey line indicates the mean pairwise binding motif distance of 0.72. **d**, Coefficients of the candidate TFs (PGLS, LRT  $p$ -value < 0.05) predicting either intron activity or GI using TF binding score for the intron CRE sequence. **e**, Predicted intron activity using the additive model combining the three predictor TF binding scores compared to the observed intron CRE activity across the catarrhine species ( $R^2=0.78$ ,  $n=10$ ). Dropping of either predictor TF was not supported by the model (PGLS, LRT  $p$ -value < 0.05).

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-ND 4.0 International license](#).

## Supplementary Information

### TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals

Zane Kliesmete<sup>1,§</sup>, Lucas E. Wange<sup>1,§</sup>, Beate Vieth<sup>1</sup>, Miriam Esgleas<sup>2,3</sup>, Jessica Radmer<sup>1</sup>, Matthias Hülsmann<sup>1,4</sup>, Johanna Geuder<sup>1</sup>, Daniel Richter<sup>1</sup>, Mari Ohnuki<sup>1</sup>, Magdalena Götz<sup>2,3,5</sup>, Ines Hellmann<sup>1,\*§</sup>, Wolfgang Enard<sup>1,\*§</sup>

<sup>1</sup> Anthropology and Human Genomics, Department of Biology II, Ludwig-Maximilians Universität, Munich, Germany

<sup>2</sup> Department of Physiological Genomics, BioMedical Center - BMC, Ludwig-Maximilians Universität, Munich, Germany

<sup>3</sup> Institute for Stem Cell Research, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

<sup>4</sup> current address: Department of Environmental Microbiology, Eawag, 8600 Dübendorf, Switzerland & Department of Environmental Systems Science, ETH Zurich, 8092 Zürich, Switzerland

<sup>5</sup> SYNERGY, Excellence Cluster of Systems Neurology, BioMedical Center (BMC), Ludwig-Maximilians-Universität München, Planegg/Munich, Germany

§ equal author contribution

\* correspondence [hellmann@bio.lmu.de](mailto:hellmann@bio.lmu.de), [enard@bio.lmu.de](mailto:enard@bio.lmu.de)



bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

**Supplementary Table 1.** Genome sources of the TRNP1 protein coding sequences

Species	Genome Source	Genome Assembly
1 Callithrix jacchus	NCBI	ASM275486v1
2 Castor canadensis	NCBI	C.can genome v1.0
3 Cebus capucinus	Ensembl41	Cebus_imitator-1.0
4 Cercocebus atys	Ensembl41	Caty_1.0
5 Cercopithecus mitis	targeted resequencing, Enard Lab	-
6 Chlorocebus aethiops	targeted resequencing, Enard Lab	-
7 Chlorocebus sabeus	NCBI	chlSab2
8 Dasypus novemcinctus	NCBI	dasNov3
9 Delphinapterus leucas	NCBI	ASM228892v3
10 Elephantulus edwardii	NCBI	EleEdw1.0
11 Erythrocebus patas	NCBI	EryPat_v1_LBIUU
12 Felis catus	Ensembl41	Felis_catus_9.0
13 Gorilla gorilla	NCBI	Susie3
14 Heterocephalus glaber	NCBI	hetGla2
15 Homo sapiens	targeted resequencing, Enard Lab	-
16 Lipotes vexillifer	NCBI	Lipotes_vexillifer_v1
17 Macaca fascicularis	Ensembl41	Macaca_fascicularis_5.0
18 Macaca leonia	targeted resequencing, Enard Lab	-
19 Macaca mulatta	NCBI	rheMac8
20 Macaca nemestrina	Ensembl41	Mnem_1.0
21 Mandrillus sphinx	targeted resequencing, Enard Lab	-
22 Microcebus murinus	NCBI	micMur2
23 Mus caroli	Ensembl41	CAROLLEIJ_v1.1
24 Mus musculus	NCBI	mm10
25 Mus pahari	Ensembl41	PAHARLEIJ_v1.1
26 Mus spretus	Ensembl41	SPRET_EiJ_v1
27 Mustela putorius	cDNA, Goetz Lab	-
28 Odocoileus virginianus	NCBI	Ovir.te.1.0
29 Orcinus orca	NCBI	Oorc.1.1
30 Otolemur garnettii	NCBI	otoGar3
31 Pan troglodytes	NCBI	panTro6
32 Panthera pardus	Ensembl41	PanPar1.0
33 Papio anubis	targeted resequencing, Enard Lab	-
34 Papio hamadryas	NCBI	papHam1
35 Phoca vitulina	NCBI	GSC_HSeal_1.0
36 Physter catodon	NCBI	Physter_macrocephalus-2.0.2
37 Ptilocolobus tephrosceles	NCBI	ASM277652v1
38 Pongo abelii	NCBI	ponAbe3
39 Procavia capensis	NCBI	proCap1
40 Propithecus coquereli	Ensembl41	Pcoq_1.0
41 Rattus norvegicus	NCBI	rn6
42 Sus scrofa	NCBI	susScr3
43 Theropithecus gelada	NCBI	Tgel_1.0
44 Tursiops truncatus	NCBI	Tur_tru v1
45 Zalophus californianus	NCBI	zalCal2.2

## 5.5 TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals

305

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

**Supplementary Table 2.** DNase-seq experiments

	SRX	Species	Tissue	Stage	GEO Accession
1	SRX027085	Human	Fetal Brain	Day 85	GSM595922
2	SRX027086	Human	Fetal Brain	Day 85	GSM595923
3	SRX027089	Human	Fetal Brain	Day 96	GSM595926
4	SRX027091	Human	Fetal Brain	Day 96	GSM595928
5	SRX121276	Human	Fetal Brain	Day 101	GSM878650
6	SRX121277	Human	Fetal Brain	Day 104	GSM878651
7	SRX201815	Human	Fetal Brain	Day 105	GSM1027328
8	SRX121278	Human	Fetal Brain	Day 109	GSM878652
9	SRX040380	Human	Fetal Brain	Day 112	GSM665804
10	SRX027083	Human	Fetal Brain	Day 117	GSM595920
11	SRX026914	Human	Fetal Brain	Day 122	GSM530651
12	SRX027076	Human	Fetal Brain	Day 122	GSM595913
13	SRX062364	Human	Fetal Brain	Day 122	GSM723021
14	SRX040395	Human	Fetal Brain	Day 142	GSM665819
15	SRX188655	Mouse	Fetal Brain	E 14.5	GSM1003828
16	SRX191055	Mouse	Fetal Brain	E 14.5	GSM1014197
17	SRX191042	Mouse	Fetal Brain	E 18.5	GSM1014184

**Supplementary Table 3.** Human *TRNP1* DNase hypersensitive sites

	Chromosome	Start	End	ID
1	chr1	27293479	27293766	upstream1
2	chr1	27310581	27310877	upstream2
3	chr1	27318087	27318439	upstream3
4	chr1	27319449	27321900	promexon1
5	chr1	27323922	27324667	intron
6	chr1	27327174	27327461	exon2
7	chr1	27328171	27328804	downstream

**Supplementary Table 4.** Mouse *Trnp1* DNase hypersensitive sites

	Chromosome	Start	End	ID
1	chr4	133494338	133494835	intron
2	chr4	133495742	133496109	unique1
3	chr4	133497135	133498824	promexon1
4	chr4	133504292	133504667	unique2
5	chr4	133504895	133505292	unique3
6	chr4	133508796	133509525	upstream3
7	chr4	133511090	133511417	upstream2
8	chr4	133512990	133513416	upstream1

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

**Supplementary Table 5.** gDNA samples

	Family	Species	Source
1	Apes	<i>Homo sapiens</i>	DKFZ
2	Apes	<i>Pan troglodytes</i>	MPI Leipzig
3	Apes	<i>Pan paniscus</i>	MPI Leipzig
4	Apes	<i>Gorilla gorilla</i>	MPI Leipzig
5	Apes	<i>Pongo abelii</i>	MPI Leipzig
6	Apes	<i>Symphalangus syndactylus</i>	MPI Leipzig
7	Apes	<i>Nomascus gabriellae</i>	MPI Leipzig
8	Apes	<i>Hylobates agilis</i>	MPI Leipzig
9	Old World Monkeys	<i>Papio anubis</i>	Deutsches Primatenzentrum Goettingen
10	Old World Monkeys	<i>Papio hamadryas</i>	MPI Leipzig
11	Old World Monkeys	<i>Mandrillus sphinx</i>	MPI Leipzig
12	Old World Monkeys	<i>Cercocebus chrysogaster</i>	Deutsches Primatenzentrum Goettingen
13	Old World Monkeys	<i>Macaca mulatta</i>	MPI Leipzig
14	Old World Monkeys	<i>Macaca arctoides</i>	Deutsches Primatenzentrum Goettingen
15	Old World Monkeys	<i>Macaca leonia</i>	Deutsches Primatenzentrum Goettingen
16	Old World Monkeys	<i>Macaca sylvanus</i>	Deutsches Primatenzentrum Goettingen
17	Old World Monkeys	<i>Macaca nemestrina</i>	MPI Leipzig
18	Old World Monkeys	<i>Cercopithecus cephus</i>	Deutsches Primatenzentrum Goettingen
19	Old World Monkeys	<i>Cercopithecus mitis</i>	Deutsches Primatenzentrum Goettingen
20	Old World Monkeys	<i>Chlorocebus aethiops</i>	Deutsches Primatenzentrum Goettingen
21	Old World Monkeys	<i>Colobus guereza</i>	Deutsches Primatenzentrum Goettingen
22	Old World Monkeys	<i>Semnopithecus entellus</i>	Deutsches Primatenzentrum Goettingen
23	Old World Monkeys	<i>Pygathrix nemaeus</i>	Deutsches Primatenzentrum Goettingen
24	New World Monkeys	<i>Alouatta caraya</i>	Deutsches Primatenzentrum Goettingen
25	New World Monkeys	<i>Lagothrix cana</i>	Deutsches Primatenzentrum Goettingen
26	New World Monkeys	<i>Ateles fusciceps</i>	Deutsches Primatenzentrum Goettingen
27	New World Monkeys	<i>Cebus apella</i>	MPI Leipzig
28	New World Monkeys	<i>Saimiri sciureus</i>	MPI Leipzig
29	New World Monkeys	<i>Aotus azarae</i>	MPI Leipzig
30	New World Monkeys	<i>Saguinus oedipus</i>	Deutsches Primatenzentrum Goettingen
31	New World Monkeys	<i>Callimico goeldii</i>	Deutsches Primatenzentrum Goettingen
32	New World Monkeys	<i>Callithrix jacchus</i>	Deutsches Primatenzentrum Goettingen
33	New World Monkeys	<i>Pithecia pithecia</i>	Deutsches Primatenzentrum Goettingen

**Supplementary Table 6.** Cell lines used for the MPRA

Name	Purpose	Species	Cell Line	Source
Neuro2a cells	Lentiviral titer determination	<i>Mus musculus</i>	N2A	ATCC
Human embryonic kidney cells	Production of lentiviral particles	<i>Homo sapiens</i>	HEK293T	ATCC
Human neural progenitor cells	used in MPRA	<i>Homo sapiens</i>	N4_29B5	Geuder et. al.
Macaca fascicularis neural progenitor cells	used in MPRA	<i>Macaca fascicularis</i>	N15_39B2	Geuder et. al.



## 5.5 TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals 307

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

**Supplementary Table 7.** Phenotype data and its source publications used in this study. In cases where there are multiple sources listed, mean across the individual measurements was calculated. For 11 species, the phenotype data of their close sister species (column "Original species") was used. For 3 additional species with only missing GI, this information was borrowed from the indicated sister species (column "GI source" in the brackets)

Species	Original species	Body mass(g)	Brain size(g)	EQ	Brain, body mass source	GI	GI source
Alouatta caraya		2955.00	45.60	1.80	[1]	1.47	[2] (A.seninculus)
Aotus azarae	A.trivirgatus	783.60	17.40	1.67	[2]	1.31	[2]
Ateles fusciceps		9026.50	113.60	2.12	[3]	1.68	[2] (A.paniscus)
Bos taurus		596666.67	462.00	0.52	[2]	2.53	[2]
Callimico goeldii		492.20	10.95	1.43	[2]	1.25	[2]
Callithrix jacchus		288.12	7.61	1.43	[2]	1.17	[2]
Canis lupus	C.latrans	10750.00	86.23	1.43	[2]	1.80	[2]
Castor canadensis		21750.00	41.17	0.43	[2]	1.02	[2]
Cebus apella		2589.00	71.30	3.07	[4],[5],[1]	1.60	[5]
Cebus capucinus		1879.00	70.21	3.75	[3],[6]	1.69	[7] (C.albifrons)
Cercocebus atys		3792.86	100.80	3.36	[6],[1]	1.84	[5]
Cercopithecus cephus		1915.00	57.50	3.03	[6]		
Cercopithecus mitis		5041.29	67.00	1.85	[2]	1.78	[2]
Chlorocebus aethiops		3452.67	64.13	2.28	[4],[1]		
Chlorocebus sabeus		3042.80	73.61	2.84	[6],[1]		
Colobus guereza		10281.25	83.90	1.43	[8]		
Dasyptus novemcinctus		3762.00	10.75	0.36	[2]	1.07	[2]
Delphinapterus leucas		636000.00	2083.00	2.24	[9]		
Elephantulus edwardii	E.fuscipes	57.00	1.33	0.74	[2]	1.00	[2]
Equus ferus	E.caballus	367000.00	712.00	1.11	[2]	2.80	[2]
Erinaceus europaeus		801.00	3.50	0.33	[2]	1.00	[2]
Erythrocebus patas		7421.40	105.65	2.25	[2]	1.91	[2]
Felis catus		3183.33	31.18	1.17	[2]	1.50	[2]
Gorilla gorilla		99648.40	477.44	1.78	[2]	2.26	[2]
Heterocephalus glaber		41.67	0.46	0.31	[10]		
Homo sapiens		61770.32	1300.00	6.68	[2]	2.56	[2]
Hylobates agilis		5528.75	88.10	2.28	[6]		
Lagothrix cana	L.lagothricha	5959.00	95.58	2.35	[2]	1.97	[2]
Lipotes vexillifer		180000.00	558.00	1.40	[9]		
Loxodonta africana		3775000.00	5253.56	1.72	[2]	3.84	[2]
Macaca arctoides		7630.00	100.70	2.10	[4]		
Macaca fascicularis		3109.45	66.93	2.55	[4],[6],[1]	1.65	[11]
Macaca leonia		2050.00	90.00	4.53	[1]		
Macaca mulatta		7102.83	89.22	1.95	[2]	1.75	[5]
Macaca nemestrina		4456.00	110.00	3.29	[12],[1]		
Macaca sylvanus		11200.00	87.70	1.42	[1]		
Macropus eugenii		6500.00	23.70	0.55	[2]	1.13	[2]
Mandrillus sphinx		12125.00	159.40	2.44	[2]	2.14	[2]
Microcebus murinus		71.83	1.85	0.88	[2]	1.10	[2]
Mus musculus		22.25	0.55	0.57	[2]	1.03	[2]
Mustela putorius		809.00	8.25	0.77	[2]	1.63	[13]
Odocoileus virginianus		87000.00	160.00	0.65	[2]	2.27	[2]
Orcinus orca		2049000.00	5617.00	2.76	[9]		
Oryctolagus cuniculus		2000.00	6.50	0.33	[2]	1.15	[2]
Otolemur garnettii	O.crassicaudatus	952.43	10.60	0.89	[2]	1.25	[2]
Ovis aries		63966.67	125.00	0.63	[2]	1.94	[2]
Pan paniscus		39700.00	329.70	2.28	[5],[8]	2.17	[5]
Pan troglodytes		41057.09	392.06	2.65	[2]	2.46	[2]
Papio anubis		13829.29	179.10	2.51	[4],[6],[8],[14]	2.00	[7]
Papio hamadryas		16000.00	182.00	2.31	[2]	1.99	[2]
Phoca vitulina		115000.00	273.75	0.93	[2]	2.38	[2]
Ptilocolobus tephroscelus	P.badius	7854.83	76.75	1.57	[2]	1.80	[2]
Pongo abelii	P.pygmaeus	42447.67	347.75	2.30	[2]	2.21	[2]
Procapra capensis		3420.00	19.17	0.68	[2]	1.37	[2]
Propithecus coquereli	P.verreauxi	3547.12	26.90	0.94	[2]	1.34	[2]
Pteropus vampyrus	P.giganteus	1038.75	9.00	0.71	[2]	1.25	[2]
Pygathrix nana		8481.67	84.83	1.65	[2]	1.64	[2]
Rattus norvegicus		314.67	2.41	0.43	[2]	1.02	[2]
Saguinus oedipus		368.83	9.80	1.56	[2]	1.20	[2]
Saimiri boliviensis		750.00	24.10	2.38	[8]		
Saimiri sciureus		680.60	22.98	2.42	[2]	1.55	[2]
Sarcophilus harrisii			15.00		[7]	1.33	[13]
Semnopithecus entellus		7010.00	111.50	2.46	[1]		
Sorex araneus		9.00	0.20	0.38	[2]	1.00	[2]
Sus scrofa		133600.00	137.65	0.42	[2]	2.16	[2]
Symphalangus syndactylus		12172.00	134.80	2.06	[6],[8]		
Tarsius syrichta		112.05	3.83	1.35	[2]	1.10	[2]
Theropithecus gelada		7710.00	130.00	2.69	[6]		
Trichechus manatus		797000.00	382.00	0.35	[2]	1.02	[2]
Tupaia belangeri	T.glis	173.33	3.03	0.80	[2]	1.06	[2]
Tursiops truncatus		177500.00	1489.00	3.77	[2]	4.76	[2]
Zalophus californianus		140000.00	363.00	1.08	[2]	2.52	[2]

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

**Supplementary Table 8.** Comparison between PAML[15] site models for TRNP1 protein-coding sequence using Likelihood Ratio Test (LRT). Maximum likelihood of M8 is significantly higher than that of M7, suggesting that a proportion ( $p_1=8.2\%$ ) of amino acid sites in TRNP1 evolve under positive selection

Model	Parameters	lnL	$2(\ln L(M8) - \ln L(M7))$	Df	$\chi^2$ p-value
<b>M8 (beta and <math>\omega</math>)</b>	$p_0(p_1 = 1 - p_0), p, q, \omega_s > 1$	-5438.53	17.34	2	< 0.001
<b>M7 (beta)</b>	$p, q$	-5447.20			

**Supplementary Table 9.** TRNP1 amino acid sites under positive selection across the phylogeny according to Naive Empirical Bayes analysis (\*: P>95%; \*\*: P>99%)

Alignment position	Pr( $\omega > 1$ )	post mean $\omega$
35	0.977*	1.11
41	0.994**	1.12
62	0.991**	1.12
63	0.971*	1.10
86	0.984*	1.12
193	0.967*	1.10

**Supplementary Table 10.** Pairwise correlations between the substitution rates of TRNP1 (dS: synonymous substitution rates,  $\omega$ : the ratio of the non-synonymous over the synonymous substitution rates) and the rate of change in either GI, brain size or body mass estimated separately across 31 mammalian species using Coevol[16]. Partial correlations are the maximally controlled correlations, controlling for all other included variables

Parameter 1	Parameter 2	Marginal Correlation (Posterior Probability)	Partial Correlation (Posterior Probability)
GI	$\omega$	0.62 (0.95)	0.745 (0.98)
$\omega$	dS	-0.044 (0.55)	0.291 (0.71)
GI	dS	-0.404 (0.97)	-0.411 (0.82)
brain mass	$\omega$	0.499 (0.89)	0.667 (0.96)
$\omega$	dS	-0.031 (0.52)	0.334 (0.74)
brain mass	dS	-0.524 (0.99)	-0.516 (0.88)
body mass	$\omega$	0.44 (0.85)	0.587 (0.92)
$\omega$	dS	0.008 (0.51)	0.27 (0.7)
body mass	dS	-0.436 (0.97)	-0.425 (0.86)

**Supplementary Table 11.** Pairwise correlations between the substitution rates of TRNP1 and the rate of change in the three morphological traits (GI, brain size, body mass) across 31 mammalian species, all estimated together in a joint framework using Coevol[16]

Parameter 1	Parameter 2	Marginal Correlation (Posterior Probability)	Partial Correlation (Posterior Probability)
GI	$\omega$	0.69 (0.98)	0.474 (0.87)
brain size	$\omega$	0.638 (0.93)	0.273 (0.75)
body mass	$\omega$	0.553 (0.90)	0.035 (0.51)
$\omega$	dS	-0.242 (0.74)	0.192 (0.66)
GI	dS	-0.41 (0.97)	0.016 (0.53)
body mass	dS	-0.453 (0.98)	0.143 (0.68)
brain size	dS	-0.551 (0.99)	-0.332 (0.85)
GI	brain size	0.817 (1.00)	0.354 (0.85)
brain size	body mass	0.909 (1.00)	0.656 (0.95)
GI	body mass	0.681 (1.00)	-0.196 (0.76)

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

**Supplementary Table 12.** Model selection results between logistic regression models that predict the proportion of proliferating mouse NSCs in the presence of TRNP1 compared to a GFP control (LRT). n=donor mouse (batch). Proliferation is best predicted by the presence of TRNP1 (yes/no) together with the donor mouse to correct for the batch

Model	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
0: prolif(%) ~ 1	68	1491.35			
1: prolif(%) ~ TRNP1	67	1258.02	1	233.33	1.1E-52
2: prolif(%) ~ TRNP1+n	56	235.84	11	1022.18	3.2E-212

**Supplementary Table 13.** Model selection results between logistic regression models that predict the proportion of proliferating mouse NSCs in the presence of different TRNP1 orthologues (LRT). n=donor mouse (batch). Proliferation is best predicted by the respective TRNP1 orthologue together with the donor mouse to correct for the batch

Model	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
0: prolif(%) ~ 1	56	1067.98			
1: prolif(%) ~ orthologue	51	968.08	5	99.91	5.5E-20
2: prolif(%) ~ orthologue+n	40	121.49	11	846.58	1.9E-174

**Supplementary Table 14.** The induced NSC proliferation rates by the different TRNP1 orthologues (according to model 2 from Suppl. Table 13)

Species	Proliferation rate	Std. Error
macaque	0.49	0.028
galago	0.51	0.027
ferret	0.52	0.023
mouse	0.54	0.022
human	0.58	0.022
dolphin	0.65	0.025

**Supplementary Table 15.** Pairwise proliferation rate comparison between the TRNP1 orthologues of interest (Tukey test)

Comparison	Estimate	Std. Error	z value	Pr(> z )
human - mouse	0.15	0.069	2.188	0.0989
dolphin - human	0.30	0.089	3.378	0.0028
human - macaque	0.34	0.085	3.972	0.0003
human - galago	0.26	0.085	3.071	0.0082

**Supplementary Table 16.** PGLS model selection using LRT to test whether CRE activity of the 7 *TRNP1* regulatory regions is predictive for either brain size or gyrification (GI) across species. The reduced model contains intercept as the only predictor

Model	Value	Std. Error	df	L.Ratio	LRT p-value
log2(brain size) ~ log2(upstream1)	0.01	0.152	1	0.01	0.929
log2(brain size) ~ log2(upstream2)	-0.28	0.234	1	1.43	0.232
log2(brain size) ~ log2(upstream3)	-0.30	0.223	1	1.92	0.166
log2(brain size) ~ log2(exon1)	0.31	0.487	1	0.42	0.517
log2(brain size) ~ log2(intron)	0.35	0.399	1	0.78	0.377
log2(brain size) ~ log2(exon2)	0.28	0.286	1	0.97	0.324
log2(brain size) ~ log2(downstream)	0.18	0.474	1	0.16	0.693
log2(GI) ~ log2(upstream1)	0.01	0.073	1	0.01	0.929
log2(GI) ~ log2(upstream2)	-0.04	0.060	1	0.53	0.468
log2(GI) ~ log2(upstream3)	-0.06	0.048	1	1.51	0.219
log2(GI) ~ log2(exon1)	0.07	0.088	1	0.62	0.431
log2(GI) ~ log2(intron)	0.14	0.086	1	2.75	0.097
log2(GI) ~ log2(exon2)	0.10	0.086	1	1.33	0.250
log2(GI) ~ log2(downstream)	0.03	0.114	1	0.07	0.787

bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

**Supplementary Table 17.** PGLS model selection using LRT to test whether the association between GI and intron CRE activity on the Old World monkey and great ape branch is consistent across three independent cell lines. The reduced model contains intercept as the only predictor

Model	Value	Std.Error	df	L.Ratio	LRT p-value	Cell line
$\log_2(\text{GI}) \sim \log_2(\text{intron})$	0.20	0.059	1	8.66	0.003	human1
$\log_2(\text{GI}) \sim \log_2(\text{intron})$	0.14	0.071	1	3.81	0.051	human2
$\log_2(\text{GI}) \sim \log_2(\text{intron})$	0.10	0.059	1	3.31	0.069	macaque

**Supplementary Table 18.** Enriched Gene Ontology Terms (Fisher's  $p$ -value < 0.05) of the 22 TFs with binding site enrichment on the intron CRE sequences from the 10 catarrhine species. Background: all expressed TFs included in the motif binding enrichment analysis (392)

GO.ID	Term	Annotated	Significant	Expected	Fisher's P
GO:0010817	regulation of hormone levels	26	5	1.47	0.011
GO:0042127	regulation of cell population proliferation	117	12	6.60	0.012
GO:0008285	negative regulation of cell population proliferation	61	8	3.44	0.012
GO:0043523	regulation of neuron apoptotic process	20	4	1.13	0.020
GO:1901615	organic hydroxy compound metabolic process	21	4	1.18	0.024
GO:0051402	neuron apoptotic process	23	4	1.30	0.033
GO:1903706	regulation of hemopoiesis	47	6	2.65	0.037
GO:0006325	chromatin organization	35	5	1.97	0.037
GO:0008283	cell population proliferation	135	12	7.62	0.039
GO:0090596	sensory organ morphogenesis	36	5	2.03	0.042
GO:0006259	DNA metabolic process	25	4	1.41	0.044

**Supplementary Table 19.** PGLS model selection using LRT where GI was predicted using either standardized TRNP1 protein evolution rates ( $\omega$ ) combined with standardized intron CRE activity or the standardized  $\omega$  alone across the Old World monkey and great apes for which both measurements were available (n=9)

Model	Predictor	Value	Std.Error	df	logLik	L.Ratio	LRT p-value
$\log_2(\text{GI}) \sim \log_2(\omega)$	$\omega$	0.19	0.041	3	11.27		
$\log_2(\text{GI}) \sim \log_2(\omega) + \log_2(\text{intron})$	$\omega$	0.16	0.029				
$\log_2(\text{GI}) \sim \log_2(\omega) + \log_2(\text{intron})$	intron	0.05	0.015	4	15.66	8.78	0.003



bioRxiv preprint doi: <https://doi.org/10.1101/2021.02.05.429919>; this version posted February 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1. Warnke, P. Mitteilung neuer Gehirn-und Körpergewichtsbestimmungen bei Saugern. *Psychol. Neurol* **13**, 355–403 (1908).
2. Lewitus, E. *et al.* An adaptive threshold in mammalian neocortical evolution. *PLoS biology* **12**, e1002000 (2014).
3. Crile, G. & Quiring, D. P. A record of the body weight and certain organ and gland weights of 3690 animals (1940).
4. Bronson, R. T. Brain weight-body weight relationships in 12 species of nonhuman primates. *Am. J. Phys. Anthropol.* **56**, 77–81 (1981).
5. Rilling, J. K. & Insel, T. R. The primate neocortex in comparative perspective using magnetic resonance imaging. en. *J. Hum. Evol.* **37**, 191–223 (1999).
6. Hrdlička, A. Weight of the brain and of the internal organs in American monkeys. With data on brain weight in other apes. *Am. J. Phys. Anthropol.* **8**, 201–211 (1925).
7. Zilles, K. *et al.* Gyrification in the cerebral cortex of primates. *Brain Behav. Evol.* **34**, 143–150 (1989).
8. Boddy, A. M. *et al.* Comparative analysis of encephalization in mammals reveals relaxed constraints on anthropoid primate and cetacean brain scaling. *J. Evol. Biol.* **25**, 981–994 (2012).
9. Manger, P. R. An examination of cetacean brain structure with a novel hypothesis correlating thermogenesis to the evolution of a big brain. en. *Biol. Rev. Camb. Philos. Soc.* **81**, 293–338 (2006).
10. Kverková, K. *et al.* Sociality does not drive the evolution of large brains in eusocial African mole-rats. en. *Sci. Rep.* **8**, 9203 (2018).
11. Ventura-Antunes, L. *et al.* Different scaling of white matter volume, cortical connectivity, and gyrification across rodent and primate brains. en. *Front. Neuroanat.* **7**, 3 (2013).
12. Spitzka, E. A. Brain-weights of animals with special reference to the weight of the brain in the Macaque monkey. *J. Comp. Neurol.* **13**, 9–17 (1903).
13. Brodmann, K. Neuere Forschungsergebnisse der Großhirnrinden-anatomie mit besonderer Berücksichtigung anthropologischer Fragen. *Naturwissenschaften* **1**, 1120–1122 (1913).
14. Stephan, H. *et al.* New and revised data on volumes of brain structures in insectivores and primates. en. *Folia Primatol.* **35**, 1–29 (1981).
15. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13**, 555–556 (1997).
16. Lartillot, N. & Poujol, R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution* **28**, 729–744 (2010).



# Acknowledgements

This is the most difficult but also one of the most important parts of this thesis. I am thankful to so many people and I will not be able to mention all of you in sufficient detail without this getting too long for anyone to read. First I want to thank everyone, that I was fortunate enough to meet in the Enard/Hellmann WG (in this case I mean Wohngemeinschaft rather than working group) over the last six! years. That includes everyone from A to Z (Andi to Zeynep). I like to believe that it took me as long as it did to complete my PhD because I was enjoying my time here so much that I didn't see the point in ending it earlier.

I want to thank Wolfi for his unique inspiring way of leading a group, thinking and talking about science and creating an environment where I was never afraid to babble out any random idea I had (and I had many, often random). For showing me what kind of scientist I would like to be, transparent, kind, clear thinking and with a scientific idea, a vision that reaches beyond the next paper or project. Similarly I am thankful to Ines who always gave me great advice (even if I sometimes didn't want to hear it :)) and provoked scientific discussions that helped me to see the bigger picture. I want to thank Mari for being one of the best scientists I know and also one of the coolest most impressive people.

I want to thank all of the students I "supervised" over the years (in chronological order Matthias, Jessy, Nik, Amy, Fiona and Isabella) and who probably taught me more than I taught them. Some about scientific rigor others about artistic and scientific creativity and exceptional motivation and many about organization, planning and all of the other things you do while I procrastinate.

I'm thankful to the previous generations of PhD students who taught me everything to make it in this lab, particularly my first supervisor Beate, my RNA lab mentors and buddies

Christoph and Johannes and Daniel who regularly grounds me (in a good way) when I think I know stuff and then realize that he knows so much more while at the same time claiming he knows nothing.

I'm thankful to the current and future generations of PhD students, that carry on what we have started. Zane and Philipp particularly for always being worthy opponents (usually more than that) in all the disciplines of scientific combat (eating, drinking, football, headis, darts, beer pong). Zane I hope my secret masterplan to work as an assistant in your research group one day works out. I want to thank the Twitter team (Fiona) for hours of what might have looked like procrastination but really was important scientific communication and outreach #Parafim #TwitterMonday(?). I want to thank Anita the most humble data analysis superstar for making sure that this group continues to be one of the greatest places to work. I want to thank Jessie for saying that some things (mostly sequencing related) will not be possible anymore when I'm gone. It's of course not true and you guys are absolutely capable of handling it, but it still is nice to hear that and feel indispensable :).

While PhD students, research course students, master students, bachelor students came and went, Karin and Ines were the constant, the backbone of this group. Without your often underestimated work we would not get anything done. Thank you Karin for great stories and friendship. For being who you are, always fun to talk to, never boring, always unexpected.

Finally, I want to thank Aco and Raissa. I don't even know where to start, maybe at the beginning. First, it was just Johanna and me, we got along well but were after all a dilemma. Then after half a year Aleks joined as a PhD student and we became known as Trilemma. We shared many experiences some legal others more in a legal grey zone. We celebrated the life as PhD students relentlessly, sometimes we maybe even overdid it a bit, and we comforted each other when things did not go as planned. Sometimes we also did science together and it was awesome. You we're the best collaborators ever and are the best friends one can wish for.

Now comes the difficult task of making a smooth transition from the lab family to my family and who would be better for this transition than Paulina? Part of this lab and part of my family by now! Danke das du mir nicht Übel nimmst, wenn ich mal wieder so tief in Gedanken oder irgendein Medium versunken bin das ich unansprechbar bin, besonders



in diesen letzten Wochen des Thesis Schreibens. Danke das du mit mir nach Barcelona gehst, ich freu mich auf diesen nächsten Schritt mit dir. Egal wie es mit der Wissenschaft in Barcelona läuft, es wird eine grandiose Erfahrung, weil wir es zusammen machen. Danke das du manchmal über meine Witze lachst und immer da bist auch wenn ich dass/das nicht beherrsche und viel zu Laut nieße.

Besonders erwähnen will ich natürlich auch meine Familie und Freunde aber auch hier kann ich nicht allen angemessen danken, weil es nur noch sechs Stunden zur Abgabe sind ;). Ich danke in einem Rundumschlag allen meinen Freunden dafür, dass sie meine Freunde sind, den alten aus der Schulzeit, den neuen aus dem Master und allen dazwischen. Gleichermäßen danke ich natürlich auch all meinen Geschwistern, für alles.

Mama, Papa, Saskia, Franz und alle weiteren Elternteile die ich potentiell vergessen habe ;) Nur wenn man solche Eltern hat und gleich vier davon kann man ohne größere Sorgen 11 Jahre lang studieren. Danke das ihr immer alles unterstützt habt, bedingungslos. Zu guter Letzt, Danke Opa das du unermüdlich gewartet hast und mir immer geglaubt hast, dass ich bald fertig bin. Es ist soweit ich bin fertig!



# Curriculum Vitae

# LUCAS ESTEBAN WANGE

ORCID: 0000-0002-3275-9156 · Twitter: @le\_wange

---

## UNIVERSITY EDUCATION

**FEB. 2017 – CURRENT**

**DOCTORAL CANDIDATE**, LUDWIG MAXIMILIANS UNIVERSITY,  
PROF. WOLFGANG ENARD

Establishing methods to measure gene expression evolution using single cell and bulk RNA sequencing in primates.

From data generation to data analysis.

**OCT. 2014 - OCT. 2016**

**MASTER OF SCIENCE (BIOLOGY)**, LUDWIG MAXIMILIANS UNIVERSITY

Final grade: 1.23

Thesis title: Identification of active Enhancers across the primate lineage using a massively parallel reporter assay (MPRA)

**OCT. 2011 - JUL. 2014**

**BACHELOR OF SCIENCE (BIOLOGY)**, LUDWIG MAXIMILIANS UNIVERSITY

Final grade: 2.22

Thesis title: Visualizing changes in cAMP-levels in the sleeping sickness parasite *Trypanosoma brucei* via FRET

## PROFESSIONAL EXPERIENCE

**OCT. 2015 – DEZ 2016**

**STUDENT ASSISTANT**, MYRIAD GMBH

Participation in companies research projects including data analysis.

Internship in company head quarter Salt Lake City, UT

**OCT. 2012 – MAY 2016**

**STUDENT ASSISTANT**, CENTER FOR HUMAN GENETICS AND LABORATORY  
DIAGNOSTICS

Routine diagnostic immunotyping using NGS.

## EXPERTISE

### WETLAB

- Single cell and bulk RNA-sequencing
- scRNA-seq CRISPR screens (CROP-seq)
- single cell and bulk ATAC-seq
- single cell genotyping
- Massively parallel reporter assays (MPRA)
- NOMe-seq

### DRYLAB

- Computational analysis of RNA-seq data
- Preprocessing of raw sequencing data
- Data analysis and visualization in R
- Comparative genomic analyses

## LANGUAGES

German     ● ● ● ● ●  
 English    ● ● ● ● ●  
 Spanish    ● ● ● ● ●

R            ● ● ● ● ●  
 Bash        ● ● ● ● ●  
 Python      ● ● ● ● ●

## TEACHING

- Supervision of Master Thesis projects (2017- current)
- Co-supervisor Seminar Cancer Evolution (2017 -2018)
- Co-supervisor Seminar Current Topics in Statistical Genomics (2018 – current)
- Co-supervisor Seminar RNA-seq analysis (2022)

## MEMBERSHIPS

- Single Cell Omics Germany (SCOG)
- European Society of Human Genetics (2017)
- Scientists For Future, Munich

## PRESENTATIONS AND CONFERENCES

**EUROPEAN SOCIETY OF HUMAN GENETICS (ESHG) 2017, BARCELONA ES**

**Poster:** Experience of BRCA clinical testing for 3,200 patients: An international perspective

**INTERNATIONAL SYMPOSIUM ACUTE LEUKEMIAS XVI 2017, MUNICH DE**

**CANCER EVOLUTION, MUNICH DE**

**Poster:** Single-cell RNA sequencing of ALL patient samples before and after treatment

**SINGLE CELL BIOLOGY 2019, BEAVERCREAK US**

**Poster:** Single cell genotyping and transcriptome analysis of AML patients with subclonal FLT3 mutations

**INTERNATIONAL SYMPOSIUM ACUTE LEUKEMIAS XVII 2019, MUNICH DE**

**Poster:** Analyzing transcriptional profiles of childhood ALL at single cell resolution

**SELECTED PUBLICATIONS****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>

**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 mcSCR-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>

**Benchmarking single-cell RNA-sequencing protocols for cell atlas projects.**

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D. J., Álvarez-Varela, A., Batlle, E., Sagar, Grün, D., Lau, J. K., Boutet, S. C., Sanada, C., Ooi, A., Jones, R. C., Kaihara, K., Brampton, C., Talaga, Y., Sasagawa, Y., Tanaka, K.,...**Wange, L. E.** ... Heyn, H. (2020). *Nature Biotechnology*, 38(6), 747–755. <https://doi.org/10.1038/s41587-020-0469-4>

**\*Shared first author**