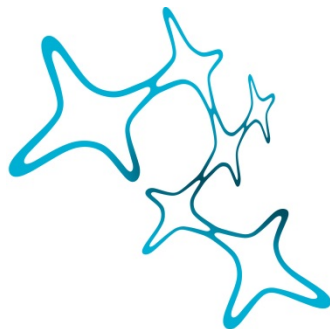


---

# OPTIMIZING TRANSCRIPTOMICS TO STUDY THE EVOLUTIONARY EFFECT OF *FOXP2*

---

Aleksandar Janjic



Graduate School of  
Systemic Neurosciences  
LMU Munich



Dissertation der Graduate School of Systemic Neurosciences  
der Ludwig-Maximilians-Universität München

October 2021

**Supervisor**

Professor Wolfgang Enard  
Anthropology and Human Genomics  
Department Biology II  
Faculty of Biology  
Ludwig-Maximilians University Munich

**First Reviewer:** Professor Wolfgang Enard

**Second Reviewer:** PD Dr. Carsten Wotjak

**Date of Submission:** October 15, 2021

**Date of Defense :** January 10, 2022

# TABLE OF CONTENTS

<b>Summary .....</b>	<b>v</b>
<b>Zusammenfassung .....</b>	<b>vii</b>
<b>Acknowledgements .....</b>	<b>xi</b>
<b>Abbreviations .....</b>	<b>xiii</b>
<b>Introduction .....</b>	<b>1</b>
Gene expression contributes to biological function .....	1
Quantifying gene expression to gain biological insight .....	5
Selecting an appropriate method.....	9
RNA sequencing: the past, present, and future .....	11
Developing and benchmarking RNA-seq protocols.....	17
Investigating Foxp2 and its role in the evolution of human speech using global transcriptomics.....	19
Study rationale.....	23
<b>Results .....</b>	<b>25</b>
Sensitive and powerful single-cell RNA sequencing using mc- SCRB-seq. ....	25
Benchmarking single-cell RNA-sequencing protocols for cell at- las projects. ....	35
Prime-seq, efficient and powerful bulk RNA-sequencing .....	49
Investigating loss-of-function and human evolution of FOXP2 using global transcriptomic analysis .....	97
<b>Discussion .....</b>	<b>149</b>
Single cell transcriptomics and the race to the top .....	149
Bulk RNA-seq in a single cell world .....	151
Characterizing the effect of Foxp2 on dopaminergic striatal neu- rons.....	153

<b>References .....</b>	<b>155</b>
<b>Figure List .....</b>	<b>164</b>
<b>Publication List .....</b>	<b>165</b>
<b>Publication Copy Rights .....</b>	<b>167</b>
Sensitive and powerful single-cell RNA sequencing using mc- SCRB-seq.....	167
Benchmarking single-cell RNA-sequencing protocols for cell at- las projects.....	168
Prime-seq, efficient and powerful bulk RNA-sequencing.....	169

# SUMMARY

The field of genomics was established with the sequencing of the human genome, a pivotal achievement that has allowed us to address various questions in biology from a unique perspective. One question in particular, that of the evolution of human speech, has gripped philosophers, evolutionary biologists, and now genomicists. However, little is known of the genetic basis that allowed humans to evolve the ability to speak. Of the few genes implicated in human speech, one of the most studied is *FOXP2*, which encodes for the transcription factor Forkhead box protein P2 (FOXP2). *FOXP2* is essential for proper speech development and two mutations in the human lineage are believed to have contributed to the evolution of human speech. To address the effect of *FOXP2* and investigate its evolutionary contribution to human speech, one can utilize the power of genomics, more specifically gene expression analysis via ribonucleic acid sequencing (RNA-seq).

To this end, I first contributed in developing mcSCRB-seq, a highly sensitive, powerful, and efficient single cell RNA-seq (scRNA-seq) protocol. Previously having emerged as a central method for studying cellular heterogeneity and identifying cellular processes, scRNA-seq was a powerful genomic tool but lacked the sensitivity and cost-efficiency of more established protocols. By systematically evaluating each step of the process, I helped find that the addition of polyethylene glycol increased sensitivity by enhancing the cDNA synthesis reaction. This, along with other optimizations resulted in developing a sensitive and flexible protocol that is cost-efficient and ideal in many research settings.

A primary motivation driving the extensive optimizations surrounding single cell transcriptomics has been the generation of cellular atlases, which aim to identify and characterize all of the cells in an organism. As such efforts are carried out in a variety of research groups using a number of different RNA-seq protocols, I contributed in an effort to benchmark and standardize scRNA-seq methods. This not only identified methods which may be ideal for the purpose of cell atlas creation, but also highlighted optimizations that could be integrated into existing protocols.

Using mcSCRB-seq as a foundation as well as the findings from the scRNA-seq benchmarking, I helped develop prime-seq, a sensitive, robust, and most importantly, affordable bulk RNA-seq protocol. Bulk RNA-seq was frequently overlooked during the efforts to optimize and establish single-cell techniques, even though the method is still extensively used in analyzing gene expression. Introducing early barcoding and reducing library generation costs kept prime-seq cost-efficient, but basing it off of single-cell methods ensured that it would be a sensitive and powerful technique. I helped verify this by benchmarking it against TruSeq generated data and then helped test the robustness by generating prime-seq libraries from over seventeen species. These optimizations resulted in a final protocol that is well suited for investigating gene expression in comprehensive and high-throughput studies.

Finally, I utilized prime-seq in order to develop a comprehensive gene expression atlas to study the function of *FOXP2* and its role in speech evolution. I used previously generated mouse models: a knockout model containing one non-functional *Foxp2* allele and a humanized model, which has a variant *Foxp2* allele with two human-specific mutations. To study the effect globally across the mouse, I helped harvest eighteen tissues which were previously identified to express *FOXP2*. By then comparing the mouse models to wild-type mice, I helped highlight the importance of *FOXP2* within lung development and the importance of the human variant allele in the brain.

Both mcSCRB-seq and prime-seq have already been used and published in numerous studies to address a variety of biological and biomedical questions. Additionally, my work on *FOXP2* not only provides a thorough expression atlas, but also provides a detailed and cost-efficient plan for undertaking a similar study on other genes of interest. Lastly, the studies on *FOXP2* done within this work, lay the foundation for future studies investigating the role of *FOXP2* in modulating learning behavior, and thereby affecting human speech.

# ZUSAMMENFASSUNG

Der Bereich der Genomik wurde mit der Sequenzierung des menschlichen Genoms begründet, eine entscheidende Errungenschaft, die es uns ermöglicht hat, verschiedene Fragen der Biologie aus einer einzigartigen Perspektive zu betrachten. Vor allem die Frage nach der Evolution der menschlichen Sprache hat Philosoph:innen, Evolutionsbiolog:innen und nun auch Genomiker:innen beschäftigt. Allerdings ist nur wenig über die genetische Grundlage bekannt, die es dem Menschen ermöglichte, die Fähigkeit zu sprechen zu entwickeln. Von den wenigen Genen, die mit der menschlichen Sprache in Verbindung gebracht werden, ist eines der am besten untersuchten *FOXP2*, das für den Transkriptionsfaktor Forkhead Box Protein P2 (*FOXP2*) kodiert. *FOXP2* ist für die korrekte Sprachentwicklung essentiell, und es wird angenommen, dass zwei Mutationen in der menschlichen Abstammungslinie zur Evolution der menschlichen Sprache beigetragen haben. Um die Wirkung von *FOXP2* zu erforschen und seinen evolutionären Beitrag zur menschlichen Sprache zu untersuchen, kann man sich die Möglichkeiten der Genomik zunutze machen, insbesondere die Genexpressionsanalyse mittels Ribonukleinsäuresequenzierung (RNA-seq).

Zu diesem Zweck habe ich zunächst an der Entwicklung von mcSCRIB-seq mitgewirkt, einem hochsensiblen, leistungsstarken und effizienten Einzelzell-RNA-seq-Protokoll (scRNA-seq). Nachdem sich scRNA-seq als zentrale Methode zur Untersuchung der zellulären Heterogenität und zur Identifizierung zellulärer Prozesse herauskristallisiert hatte, war es zwar ein leistungsfähiges genomweites Verfahren, aber es fehlte die Empfindlichkeit und Kosteneffizienz etablierter Protokolle. Durch die systematische Bewertung der einzelnen Prozessschritte konnte ich helfen zu zeigen, dass die Zugabe von Polyethylenglykol die Empfindlichkeit erhöht, indem die Effizienz der cDNA-Synthesereaktion verbessert wird. Zusammen mit anderen Optimierungen führte dies zur Entwicklung eines empfindlichen und flexiblen Protokolls, das kosteneffizient und ideal für viele Forschungsbereiche ist.

Eine Hauptmotivation für die umfangreichen Optimierungen im Bereich der Einzelzelltranskriptomik war die Erstellung von Zellatlanten, die darauf abzielen, alle Zellen eines Organismus zu identifizieren und zu charakterisieren. Da solche Bemühungen in einer Vielzahl von Forschungsgruppen unter Verwendung einer Reihe unterschiedlicher RNA-seq-Protokolle durchgeführt werden, habe ich dazu beigetragen, scRNA-seq-Methoden zu vergleichen und zu standardisieren. Dabei wurden nicht nur Methoden ermittelt, die für die Erstellung von Zellatlanten ideal sind, sondern auch Optimierungen aufgezeigt, die in bestehende Protokolle integriert werden könnten.

Auf der Grundlage von mcSCRB-seq und den Erkenntnissen aus dem scRNA-seq-Methodenvergleichsstudie habe ich an der Entwicklung von prime-seq mitgewirkt, einem sensitiven, robusten und vor allem erschwinglichen Bulk-RNA-seq-Protokoll. Bulk-RNA-seq wurde bei den Bemühungen um die Optimierung und Etablierung von Einzelzelltechniken häufig übersehen, obwohl die Methode nach wie vor in großem Umfang zur Analyse der Genexpression eingesetzt wird. Durch die Einführung des frühen Barcodings und die Senkung der Kosten für die Bibliothekserstellung blieb prime-seq kosteneffizient, während durch die Zugrundelegung von Einzelzellmethoden sichergestellt wurde, dass es sich um eine empfindliche und leistungsstarke Technik handeln würde. Ich habe dazu beigetragen, dies durch einen Vergleich mit TruSeq generierten Daten zu verifizieren und dann die Robustheit zu testen, indem ich Prime-Seq-Bibliotheken von mehr als siebzehn Arten generierte. Diese Optimierungen führten zu einem endgültigen Protokoll, das sich gut für die Untersuchung der Genexpression in umfassenden Studien mit hohem Durchsatz eignet.

Schließlich nutzte ich prime-seq, um einen umfassenden Genexpression-satlas zur Untersuchung der Funktion von *FOXP2* und seiner Rolle in der Sprachevolution zu entwickeln. Ich verwendete zuvor generierte Mausmodelle: ein Knockout-Modell mit einem nicht funktionalen *Foxp2*-Allel und ein humanisiertes Modell, das ein variantes *Foxp2*-Allel mit zwei humanspezifischen Mutationen aufweist. Um die Auswirkungen auf die gesamte Maus zu untersuchen, half ich bei der Entnahme von achtzehn Geweben, bei denen zuvor festgestellt worden war, dass sie *FOXP2* exprimieren. Durch den anschließenden Vergleich der Mausmodelle mit Wildtyp-Mäusen konnte ich die Bedeutung von *FOXP2* in der Lungenentwicklung und die Bedeutung der menschlichen Variante im Gehirn aufzeigen.

Sowohl mcSCRB-seq als auch prime-seq wurden bereits in zahlreichen Studien verwendet und veröffentlicht, um eine Vielzahl von biologischen und biomedizinischen Fragen zu beantworten. Darüber hinaus liefert meine Arbeit zu *FOXP2* nicht nur einen umfassenden Expressionsatlas, sondern auch eine detaillierte und kosteneffiziente Strategie für die Durchführung einer ähnlichen Studie zu anderen Genen von Interesse. Schließlich legen die im Rahmen dieser Arbeit durchgeführten Studien zu *FOXP2* den Grundstein für künftige Studien, die die Rolle von *FOXP2* bei der Modulation des Lernverhaltens, welches die Entwicklung der menschlichen Sprache beeinflusst, untersuchen.

-

To all those that have supported my career,  
those which have instilled in me a deep sense of appreciation for science,  
and the many friends I have made along the way.

-

# ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor and *Doktorvater*, Wolfgang Enard, without whom this work would not have been possible. Not only did you provide me with the necessary resources to carry out this project, but you advised me every step of the way. Additionally, the lab you carefully created is composed of the most compassionate individuals, all of which have consistently helped me during my research and made the experience truly wonderful. A special thank you to Karin Bauer, Ines Bliessener, Steffi Färberböck, and Ming Zhao for running a tight ship and always getting things done; Lukas Schmitt, Beate Vieth, Swati Parekh, Zane Kliesmete, and Philipp Janssen for helping with analysis, plotting, and all things R; and Jessy Radmer, Daniel Richter, and Paulina Spürk for all of your help in the lab.

I would also like to thank my current mentors, Ines Hellmann, Carsten Wotjak, and Özgün Gökce, as well as my former mentors, James Mandell, Jen Sokolowski, Paul Fisher, Luni Emdad, Bridget Quinn, Eric Mayton, Danny Meilinger, Stefan Müller, Christoph Ziegenhain, and Johannes Bag-noli. You all provided me with invaluable feedback and opportunities during each step of my scientific career and I am truly thankful. I know I would not be where I am today without your guidance.

Additionally, I am thankful for the opportunities that have been afforded to me by LMU and the Graduate School of Systemic Neuroscience. I am thankful for the sequencing services provided by Stefan Krebs and the staff at the GeneCenter, and I am especially thankful for the mice that made this project possible, as well as Sabrina Schenk and Irena Stähler, who along with the rest of the animal caretakers tended to the mice and made sure they lived happy lives.

Even with the aforementioned network of support, none of this would have been possible without my family and friends. Thank you to my parents, Vesna and Nikola Janjic; my parents in-law, Belinda and Barry Lacy;

my siblings Inga and Igor Janjic and Kelly and Chanon Smith; and my nieces, Kennedy and Madison Smith, for their support and enthusiasm. Also, thank you to Ana, Dejan, Marko, Filip, and Anastasia Jovanovic; Maria, Marcel, and Emily Krammel; and Jelena Milic for making Munich feel like home.

Also, a special thank you to Johanna Geuder and Lucas Wange who were not only unparalleled colleagues, but fierce, fierce friends. Thank you for your guidance when I failed, for the celebrations when I succeeded, and for the coffee addiction you single-handedly created. We started this adventure together, and soon we will be ready to take on any challenges.

Finally, I would like to thank my extraordinary husband, Michael Lacy. Not only are you an inspirational and brilliant scientist, but more importantly, you are the most loving and supportive partner anyone could ask for. Without hesitation you encouraged me to begin my research and then throughout were a mainstay for a successful and thrilling doctoral candidacy. I am so delighted and fortunate we could embark on this great odyssey together, and I look forward to where the rest of our journey will take us.

# ABBREVIATIONS

<b>ATAC-seq</b>	transposase-accessible chromatin using sequencing
<b>ChIP-seq</b>	chromatin immunoprecipitation sequencing
<b>CBG</b>	cortico-basal ganglia
<b>DNA</b>	deoxyribonucleic acid
<b>DNase I</b>	deoxyribonuclease I
<b>ERCC</b>	external RNA controls consortium
<b>FACS</b>	fluorescence-activated cell sorting
<b>FOXP2</b>	forkhead box protein P2
<b>MAPit</b>	methylation accessibility protocol for individual templates
<b>mcSCRB-seq</b>	molecular crowding single cell RNA barcoding and sequencing
<b>MNase-seq</b>	micrococcal nuclease sequencing
<b>NGS</b>	next generation sequencing
<b>NOMe-seq</b>	nucleosome occupancy and methylome sequencing
<b>PEG</b>	polyethylene glycol
<b>qPCR</b>	quantitative polymerase chain reaction

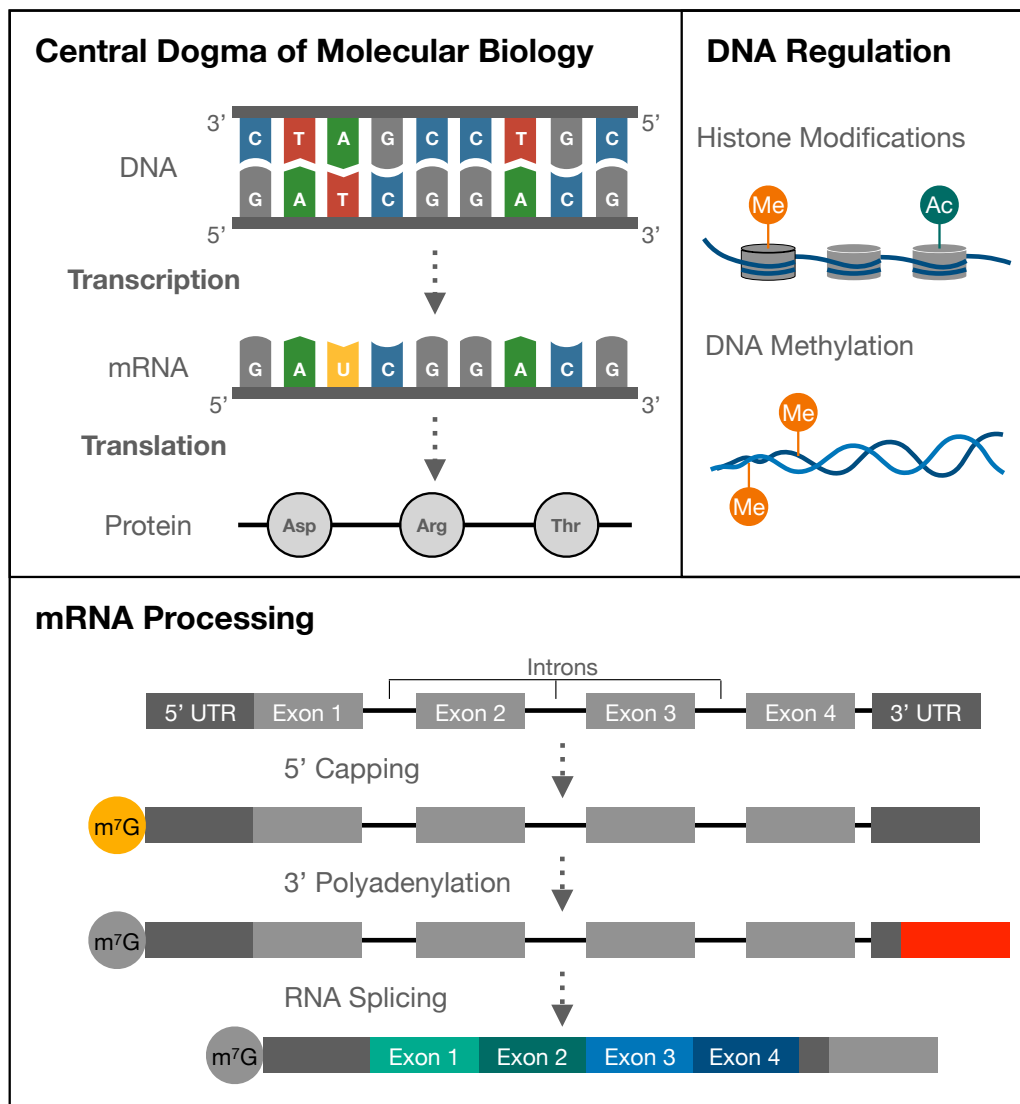
<b>RNA</b>	ribonucleic acid
<b>mRNA</b>	messenger ribonucleic acid
<b>rRNA</b>	ribosomal ribonucleic acid
<b>tRNA</b>	transfer ribonucleic acid
<b>RNA-seq</b>	ribonucleic acid sequencing
<b>scRNA-seq</b>	single cell ribonucleic acid sequencing
<b>SAGE</b>	serial analysis of gene expression
<b>SCRB-seq</b>	single cell RNA barcoding and sequencing
<b>SNP</b>	single nucleotide polymorphisms

# INTRODUCTION

## GENE EXPRESSION CONTRIBUTES TO BIOLOGICAL FUNCTION

**G**enes are a basic unit of heredity comprising a sequence of nucleotides that, when expressed, result in a gene product. In eukaryotes, a gene is a stretch of deoxyribonucleic acid (DNA) that is usually transcribed into messenger ribonucleic acid (mRNA) and then translated into protein; in some cases, however, RNA is the final gene product (e.g. transfer RNA and microRNA). In addition to the discrete steps of transcription (Roeder and Rutter, 1969) and translation (Crick, 1958), there are multiple points of regulation, such as chromatin accessibility (Kornberg, 1974), DNA silencing (Avery et al., 1944; Hotchkiss, 1948), and mRNA processing (Perry, 1976). This entire sequential flow of genetic information is summarized into the central dogma of molecular biology (Crick, 1958) (Figure 1). Thus, gene expression is an essential cornerstone in the study of biology, as it ultimately determines the phenotype, i.e. the observable trait, of the organism and thereby the function being investigated.

The realization of phenotype from genotype is carried out in a multi-step process which starts with the gene. In order for the gene to undergo expression, it must first be accessible. This is strictly regulated by the post-transcriptional modifications made to histones (Allfrey et al., 1964; Luger et al., 1997), which tightly wind and pack the DNA, as well as DNA methylation in vertebrates (Holliday and Pugh, 1975) (Figure 1). With the correct modifications, these histones will unwind regions of the DNA to make the gene accessible and serve as a template (Jenuwein and Allis, 2001), which is then read by RNA polymerase. As the RNA polymerase travels along the DNA, a complementary ribonucleotide is added to the RNA strand in a process known as transcription (Livingstone, 2010). This results in an RNA molecule complementary to the DNA molecule, with the exception of thymine which is replaced by uracil in RNA. In eukaryotes,



three various RNA polymerases exist, but RNA polymerase II is responsible for synthesis of protein-coding mRNA (Roeder and Rutter, 1969).

**Figure 1. A general overview of molecular biology.** The central dogma of molecular biology states that the flow of genetic information is unidirectional, where DNA is transcribed into mRNA, and then mRNA is translated into protein. This process is tightly regulated, among which is DNA regulation, including histone modifications and DNA methylation. These modifications either activate or repress genes by making them available for transcription. Once the mRNA is transcribed, it must be processed. The pre-mRNA is first capped on the 5' end and then polyadenylated on the 3' end. Finally, the introns are excised to form the final mRNA molecule.

Following synthesis of the mRNA, or more accurately pre-mRNA, the molecule is first processed and then exported (Köhler and Hurt, 2007). A chief component of mRNA processing is alternative splicing where the introns, or the non-coding region of the molecule, of pre-mRNA can be removed allowing multiple variant transcripts to be synthesized from one gene (Gilbert, 1978). Additionally, the pre-mRNA is capped on the 5' end and polyadenylated on the 3' end in order to stabilize and protect the molecule (Shatkin and Manley, 2000) (Figure 1). The mRNA molecule is then exported from the nucleus to the cytoplasm of the cell.

Once the mRNA has been processed and is present in the cytoplasm, ribosomes will bind the mRNA and direct transfer RNAs (tRNAs) to the mRNA. Unlike during transcription, where each nucleotide is read individually, during translation, the process of synthesizing proteins from mRNA, three nucleotides form on reading unit, or codon (Blanchet et al., 2018). Anticodon tRNAs will carry a specific amino acid and when they match a codon, the ribosome will bind the amino acid to the growing amino acid chain. The process of translating transcripts can be variable depending on the mRNA molecule and is dynamically modulated (Riba et al., 2019). The resulting proteins, each with their own function, are the final gene products and are responsible for the observed phenotype of the organism.



## QUANTIFYING GENE EXPRESSION TO GAIN BIOLOGICAL INSIGHT

**A**s stated within the central dogma of molecular biology, gene expression is a multistep process with numerous points for potential quantification, such as at the level of DNA regulation (e.g. DNA methylation and histone modification), transcription, and translation. Each, however, comes with its own advantages and disadvantages, making some methods particularly appropriate for evaluating gene expression between one or more conditions within an experiment.

### Quantifying active chromatin regions

DNA regulation, specifically, is the earliest point of expression and therefore can prove insightful, especially in predicting cell states or directions. Several methods for quantifying accessible regions of the genome have been developed and they primarily involve accessibility of the DNA to enzymatic cleavage reactions or to chemical changes by methyltransferase, followed by high-throughput sequencing (Minnoye et al., 2021).

One of the first methods to make use of enzymatic cleavage of accessible DNA was DNase-seq, which uses deoxyribonuclease I (DNaseI), an enzyme which digests DNA (Boyle et al., 2008; Song and Crawford, 2010). Since the accessibility of DNA is regulated via histones, cleavage most frequently occurs in open regions such as promoters, as well as in between nucleosomes. This technology eventually gave rise to newer more efficient protocols, such as Transposase-Accessible Chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013), which in place of DNaseI uses a genetically modified hyperactive transposase (Tn5) (Goryshin and Reznikoff, 1998) that simultaneously cleaves the DNA and then ligates adapters for downstream reactions. By combining these two steps in one reaction, the ATAC-seq protocol is more efficient; additionally it is more sensitive and requires a lower input material. On the other hand, micro-

coccal nuclease sequencing (MNase-seq) digests accessible regions and sequences the DNA that is protected by the nucleosomes, thereby providing the sequence of inaccessible DNA, opposite of the data produced by DNase-seq and ATAC-seq (Zaret, 2005). Lastly, chromatin immunoprecipitation sequencing (ChIP-seq) also uses chromatin accessibility as a basis for generating gene expression data; however, in this technology chromatin binding factors such as transcription factors or specifically modified histones are the target (Landt, 2012). Proteins of interest shield the DNA during digestion, and then using antibodies the protein of interest is precipitated and the DNA is then sequenced.

In addition to chromatin accessibility detected via enzymatic cleavage, another popular technology utilizes a methyltransferase to label open regions of DNA, which then undergo bisulfite conversion and sequencing (Carvin et al., 2003; Jessen et al., 2004). Initially this was only possible for localized analysis at targeted loci, until the advent of genome-wide methods such as the methylation accessibility protocol for individual templates (MAPit) (Darst et al., 2012; Pardo et al., 2015). More recent protocols have taken it one step further, such as nucleosome occupancy and methylome sequencing (NOMe-seq) (Kelly et al., 2012); the protocol allows one to not only identify regions of open chromatin but also endogenous methylation, which is a marker of gene repression. Therefore, one is able to get chromatin accessibility and methylation information from a single DNA strand.

## Quantifying messenger RNA

Messenger RNA (mRNA) is the intermediate step in gene expression. Precise and accurate quantification of mRNA was first developed in 1977 and carried out via northern blotting (Alwine et al., 1977). This technique involves first separating RNA molecules by gel electrophoresis, transferring them to a nylon membrane, and then hybridizing radioactively, or more recently, chemiluminescent probes to the RNA. The gel is then imaged, previously using autoradiography and more recently via chemiluminescence detection.

As northern blotting requires large quantities of mRNA, more sensitive protocols would eventually be needed. Initially developed for DNA quantification, quantitative polymerase chain reaction (qPCR) was adapted to then quantify gene expression (Wang et al., 1989; Chiang et al., 1996; Gibson et al., 1996). This highly sensitive method uses either inter-

calating dyes or hybridizing probes to emit a fluorescent signal during each PCR cycle, which in turn can be quantified to represent the number of detected mRNA copies in the sample. Although qPCR addresses the high RNA input limitation of northern blots, they are only efficiently performed on a few genes, and are therefore not appropriate for high throughput gene expression studies.

In order to accurately detect and quantify many genes simultaneously, i.e. on the order of thousands, methods such as microarrays (Schena et al., 1995), serial analysis of gene expression (SAGE) (Velculescu et al., 1995), and RNA sequencing (RNA-seq) (Bainbridge et al., 2006; Cheung et al., 2006; Emrich et al., 2007; Weber et al., 2007) were developed. Microarray technology uses a chip with surface-bound probes that can be specific for many genes (Schena et al., 1995). The RNA is reverse transcribed, labeled, and will then bind to complementary probes on the microarray surface. Binding is quantified by fluorescence intensity from the labeled sample, and thus is comparative, i.e. the signal of one gene in one condition is compared to the same gene in another condition. SAGE differs from microarrays in that it uses a tag-based approach, where cDNA is cleaved into small tags which are then ligated together, amplified, and sequenced (Velculescu et al., 1995). The resulting data contains the short sequence tags and the number of occurrences for each tag; the sequence can be matched to a database and the originating mRNA can be determined and the number of expressed genes thereby quantified. Finally, in RNA-seq, the RNA is first reverse transcribed and then the cDNA is sequenced using high-throughput next generation sequencing (NGS) (Wang et al., 2009). The resulting data is then usually mapped to a reference genome and the number of detected reads is totaled to provide information about gene expression. Unlike microarrays and SAGE, RNA-seq offers additional advantages as it does not require previous sequence knowledge and can therefore be used on non-model organisms as well as with *de novo* genes (Stark et al., 2019; Shendure, 2008).

## Quantifying proteins

Proteins, on the other hand, are the final gene product for most genes, and thus serve as the most accurate point of quantification for gene expression. Western blotting, the protein counterpart to northern blotting, is still one of the most popular techniques for protein quantification (Eisenstein, 2005; Renart et al., 1979; Towbin et al., 1979; Burnette, 1981). Denatured samples, usually consisting of cell or tissue lysate, are run on a

polyacrylamide gel and then transferred to a membrane (e.g. nitrocellulose or polyvinylidene difluoride). The membrane is incubated with an antibody specific to the protein of interest, and then either by using a fluorescent or chemiluminescent (e.g. horseradish peroxidase) label on the primary or secondary antibody, the signal is imaged and quantified.

As with other blotting approaches, western blotting suffers from low throughput. Thus for quantifying many proteins simultaneously or even the entire proteome, scientists turned to a century old technology, mass spectrometry (Thomson, 1897). Mass spectrometry measures the mass-to-charge ratio ( $m/z$ ) of ionized molecules by converting molecules to gas-phase ions, separating them by their  $m/z$  value, and recording the number of occurrences (Yates III, 2011). Advances in the field, especially the ability to analyze protein mixtures, led to shotgun-proteomics and allowed for more complex and high-throughput protein analysis (Eng et al., 1994; Link et al., 1999). However, as the resulting peptide mixture is highly complex, there is a higher chance of misidentification or a protein not being detected. This is especially problematic for lowly expressed proteins as the highly abundant proteins are preferentially sampled and may therefore prevent detection.

## SELECTING AN APPROPRIATE METHOD

**W**ith the plethora of options currently available to scientists, finding an ideal method for quantifying gene expression is determined by the question to be addressed, the conditions of the experiment, and the resources available to the experimenter. Each method has its own set of advantages and disadvantages, which must be considered during experimental planning.

Quantifying the activity of genes via chromatin accessibility, for example, is often too early of a point for accurate expression detection. Although methods such as ATAC-seq are used to accurately identify cell types and functions, their profiles do not strongly correlate to protein expression (Edfors et al., 2016; Starks et al., 2019). Rather, the methods themselves produce profiles which can help explain function.

Quantifying proteins, on the other hand, is the most accurate way to determine which genes are expressed. The available methods, however, are either not high-throughput or are too resource intensive to be regularly used in biological research (Aebersold and Mann, 2003). Although, with future developments these limitations may no longer impede wide-scale proteomics as a direct means of gene expression analysis.

Therefore, quantification on the transcriptional level, specifically using RNA-seq, has emerged to become the most popular gene expression quantification method (Stark et al., 2019). This powerful technique is not only highly sensitive but it is also highly efficient, allowing one to quantify practically all genes present in very large sample sizes. Unlike its predecessors, northern blotting and microarrays, RNA-seq does not require *a priori* sequence knowledge thereby making it a powerful tool not only in human biology but across the entire field (Kukurba and Montgomery, 2015). Furthermore, microarrays can exhibit cross-hybridization artifacts due to similar sequences, something which is not an issue when sequencing is used (Casneuf et al., 2007). A potential roadblock limiting the use of RNA-seq is the large-scale computational technologies necessary for downstream analysis. Fortunately, however, as the wet-lab technology

has developed, so too have the computational methods, further solidifying RNA-seq as the method of choice for gene expression analysis (Williams et al., 2017; Conesa et al., 2016).

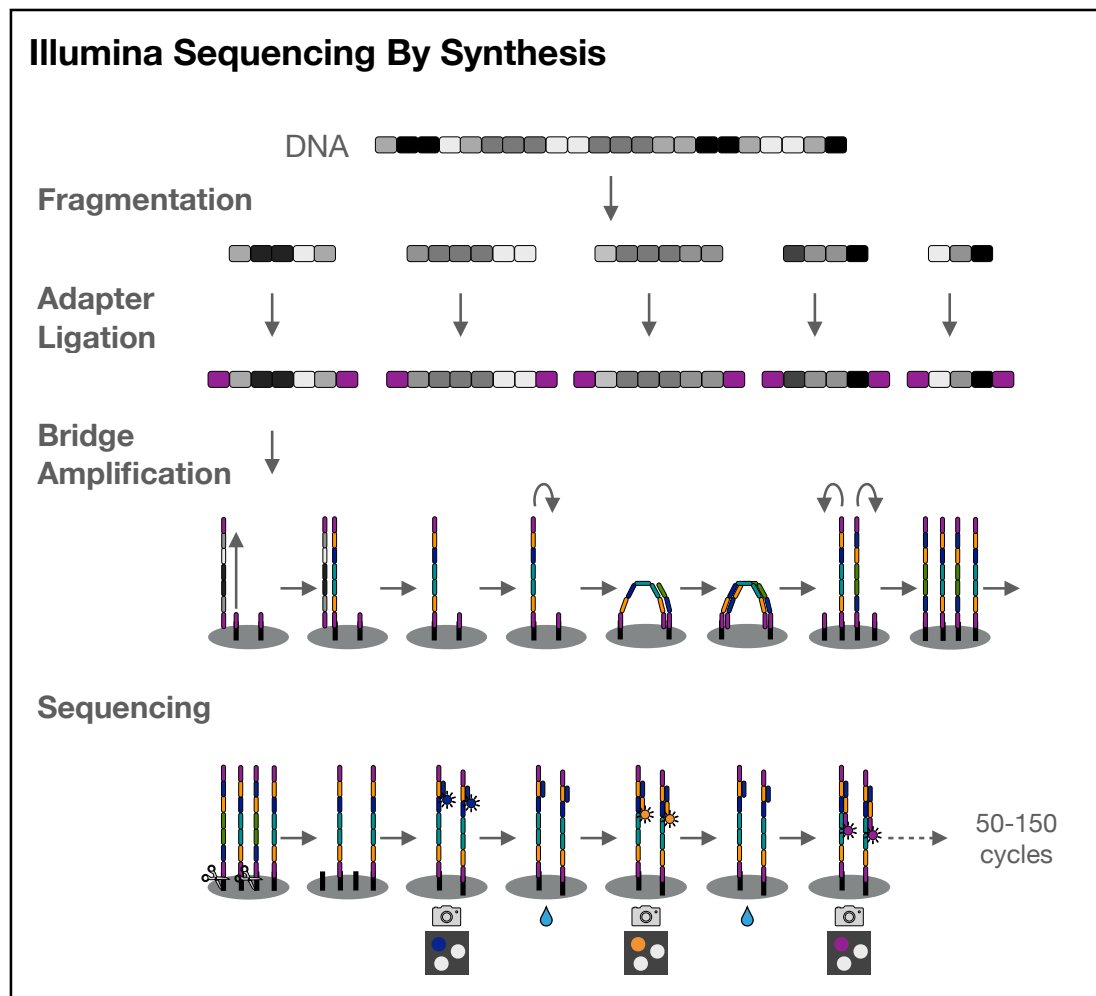
## RNA SEQUENCING: THE PAST, PRESENT, AND FUTURE

**D**eveloped over a decade ago, RNA-seq has now become a ubiquitous tool present across all molecular biology research; now an indispensable method for studying genomics, RNA-seq is primarily utilized in differential gene expression analysis but also useful in studying mRNA splicing, mRNA isoforms, regulation via non-coding RNA, RNA structure, and single nucleotide polymorphisms (SNPs) (Stark et al., 2019). Even as the method has firmly secured its position in the repertoire of molecular biologists, further development has continued and given rise to numerous advances, including single-cell transcriptomics (Tang et al., 2009; Stegle et al., 2015), spatial transcriptomics (Moor and Itzkovitz, 2017), and direct long-read RNA-seq (Garalde et al., 2018; Cartolano et al., 2016).

### Innovation begets innovation

The method we know as RNA-seq was only possible due to multiple advances in sequencing technology. Emerging in 2005, NGS brought higher throughput and lower costs than previous sequencing methods (Mardis, 2013), chiefly Sanger sequencing (Sanger et al., 1977). This allowed sequencing to be used not only as a direct means of analyzing and constructing genomes, but as a read-out in comparative studies with much larger sample sizes. This optimization was fueled by a period of fierce competition between NGS businesses, eventually leading to the dominance of Illumina's sequencing-by-synthesis approach (Shendure et al., 2017) (Figure 2).

Once NGS technologies became readily available, RNA-seq was able to be developed, first in the mid-2000s being used on various tissues including human prostate cancer cell lines (Bainbridge et al., 2006), *Medicago truncatula* (Cheung et al., 2006), *Zea mays* (Emrich et al., 2007), and *Arabidopsis thaliana* (Weber et al., 2007). These ground-breaking studies were able to more efficiently capture transcriptomic information,



**Figure 2. Illumina Sequencing by Synthesis.** Libraries are generated by first fragmenting the DNA and then ligating sequencing adapters. The library is then loaded onto the Illumina flow-cell, where it will bind to surface primers specific for the sequencing adapters. The libraries are amplified in small clusters in a process known as bridge amplification. A complementary strand is then synthesized using fluorescently labeled nucleotides. After the addition of each nucleotide, the flow-cell is imaged and this process repeated from 50 up to 150 cycles.

confirm previously known information generated by microarrays, SAGE, and pPCR, and even manage to discover new genes and new splice events. Bainbridge et al., for example, detected 10,117 genes with RNA-seq, which was more than previous transcriptomic studies on the same cell line using the Affymatrix profiling platform and Massively Parallel Signature sequencing (MPSS) were able to detect (2006). Additionally, Weber et al. found at least sixty likely protein coding sequences that were then not annotated genes (2007). Even in the early stages as the method

was being established, it was clear that RNA-seq would provide new and unparalleled power in gene expression analysis.

## Optimizations and standardizations

RNA-seq has evolved and adapted since its inception, but key components have largely remained the same (Figure 3). Cells or tissue are first lysed and RNA is extracted, most commonly through column-based kits or magnetic beads (Tavares et al., 2011). As mRNA, the protein coding portion of the transcriptome, makes up only 5% of total RNA (Warner, 1999), the mRNA molecules have to be enriched so that a majority of the sequencing resources are not sunk into unwanted data (i.e. ribosomal RNA). This is achieved either indirectly through ribosomal RNA (rRNA) depletion or directly through targeting mRNA via their poly(A) tail (Yi et al., 2011; O'Neil et al., 2013; Zhao et al., 2018). The RNA is then reverse transcribed into cDNA; in protocols utilizing rRNA depletion cDNA synthesis is usually done by priming with random hexamers, whereas in protocols where mRNA is enriched via the poly(A) tail this is usually done with oligo(dT) primers. The cDNA may then be amplified if necessary for the particular protocol being used, but it is ultimately used to generate sequencing libraries.

The type of libraries generated depends on the sequencing method to be used. In most cases this is Illumina sequencing due to their market dominance, and thus requires short insert sizes between 300-1,000 nt for optimal sequencing efficiency (Goodwin et al., 2016; Bentley et al., 2008) (Figure 2). To achieve this, either the mRNA or the cDNA must be fragmented. Fragmenting the RNA reduces 5':3' bias (Mortazavi et al., 2008) but only benefits one if they are using a full-length method. If one is using a 5' or 3' counting method, the cDNA can be fragmented into shorter fragments and then the necessary sequencing primers are ligated. This is commonly done with a fragmentase, but tagmentation through a Tn5 transposase is also popular as it performs the two-step fragmentation and adapter ligation in one step (Adey et al., 2010). Once the libraries are generated they are sequenced with the Illumina sequencing-by-synthesis approach and generate reads of 50-150 nt (Bentley et al., 2008).

Once the sequencing is completed the data is analyzed. Many computational tools and pipelines exist in order to analyze the RNA-seq data, and depending on the question at hand, some may be more advantageous

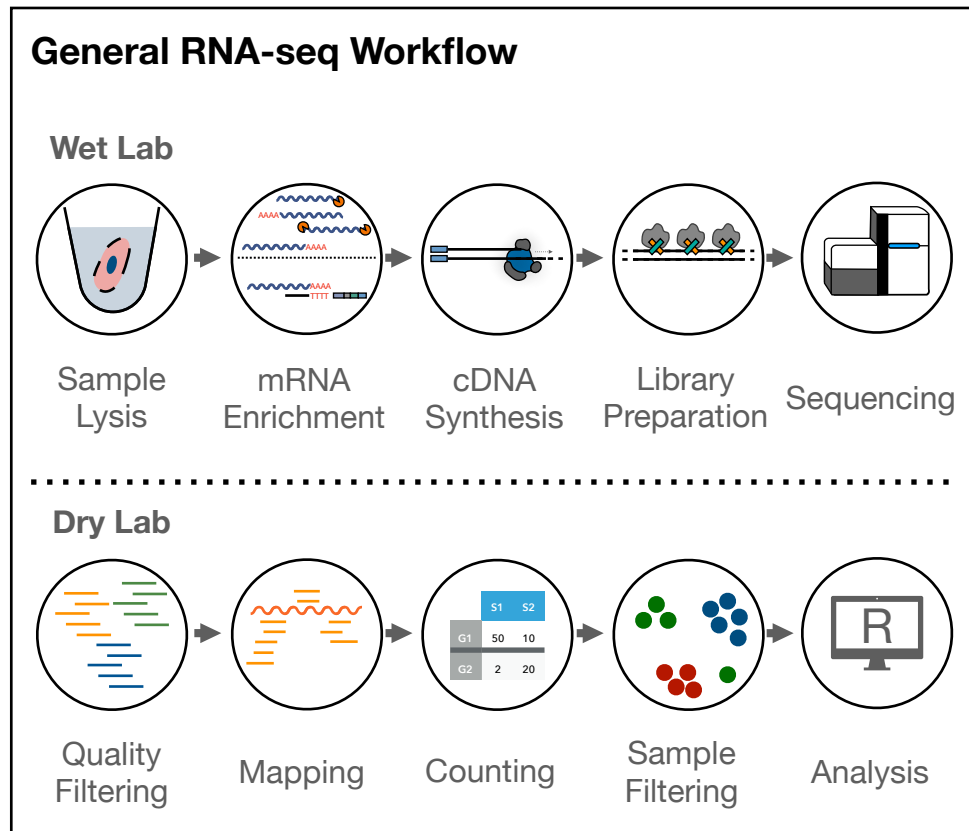
than others (Conesa et al., 2016; Sahraeian et al., 2017; Seyednasrollah et al., 2013). Generally, however, the data must first be pre-processed, a set of actions that include quality filtering based on the sequencing phred algorithm (Ewing and Green, 1998), mapping to an annotated reference genome, and counting the number of gene-assigned reads to generate a final expression or count matrix containing the number of detected reads for each gene in each sample. This count matrix is then filtered for lowly expressed genes and samples containing only a few reads, as well as normalized to address differences in sequencing depth and gene length. The normalization step is crucial and can greatly affect the final results, which is why it is essential that the correct normalization is performed on the data (Vieth et al., 2019). The final filtered and normalized count matrix may then be used for further analysis.

Downstream analysis of RNA-seq data is dependent on the biological question at hand. Often dimensionality reduction and clustering will be used to identify groupings within the samples (Kiselev et al., 2019). This may also allow one to further filter the data if there are any technical issues related to outliers. Differential expression analysis is also chiefly employed in finding expression differences between the tested conditions (Robinson et al., 2010; Love et al., 2014; Ritchie et al., 2015; Law et al., 2014). Once differentially expressed genes are detected functional annotation enrichment may be performed to interpret the observed differences between the conditions (Ashburner et al., 2000; Subramanian et al., 2005). These findings are then used to draw conclusions and answer the initial biological question.

## Going beyond averages

RNA-seq has quickly established itself as an essential method within research, but as NGS costs have decreased the capabilities of RNA-seq could be pushed even further. Scientists were interested in questions such as understanding the heterogeneity present in many of their disease states (i.e. single cell RNA-seq), determining the organization of the cells in their tissues (i.e. spatial RNA-seq), and studying species beyond only model organisms (i.e. direct long-read RNA-seq).

The primary use of RNA-seq is still determining what genes are differentially expressed between two or more conditions. The generated expression profile, however, represents an average transcriptome of all the cells



**Figure 3. General RNA Sequencing Workflow.** The process begins in the wet lab, where the sample is first lysed and the mRNA enriched, either through polyA tail capture using oligo(dT) primers or by depleting rRNA and tRNA. The mRNA is then reverse transcribed and libraries are generated from the cDNA. The samples are then sequenced. Once the sequencing data is obtained the remaining analysis is carried out in the dry lab. First the data is filtered for low-quality reads and then the data is mapped using an annotated reference genome. The mapped reads are counted and compiled into a count matrix with the number detected genes for each sample. Outlier samples, either due to technical issues (e.g. low sequencing depth) or sampling issues (e.g. incorrect sampling) are removed, and additional data analysis is performed, such as differential gene expression analysis, functional annotation analysis, and network analysis.

analyzed; this of course can be very insightful in many experiments but is unable to address the diversity of the cells within the sample. Therefore to characterize the transcriptomes of every individual cell and better understand the effect of diverse populations on the condition being investigated, single cell RNA-seq was developed (Tang et al., 2009).

Although scRNA-seq methods build on the technology of bulk RNA-seq and in principle follow a similar workflow with the exception of the necessary cell dissociation step, two important points would have to be optimized to develop successful protocols. Firstly, in order to detect the few picograms of RNA present within one cell the methods would have to become far more sensitive (Picelli et al., 2013; Hashimshony et al., 2016; Bagnoli et al., 2018; Sasagawa et al., 2018; Hagemann-Jensen et al., 2020). Secondly, to capture rare cell subtypes and have sufficient biological replicates, the number of samples must be vastly increased compared to bulk RNA-seq experiments, thereby requiring methods to be more cost efficient (Hansen et al., 2011).

Comparisons between protocols have been performed and generally concluded that plate-based or microfluidic systems detect more genes (more sensitive) but profile fewer cells (less cost efficient) and droplet-based systems detect fewer genes (less sensitive) but profile many more cells (more cost efficient) (Ziegenhain et al., 2017). Thus, depending on the question at hand, one may choose one method over the other. For example, scRNA-seq is widely used in generating atlases detailing every cell present within an organism (Regev et al., 2017), and in such a case, where differences between cells are very large, one benefits more greatly from having more efficient methods over more sensitive methods.

RNA-seq technologies are continuing to be developed as evident by spatial RNA-seq protocols (Marx, 2021), direct RNA-seq systems (Amarasinghe et al., 2020), and new bulk and single-cell RNA-seq protocols aiming to address certain limitations. These methods will continue to shape the field of molecular biology, especially as they become increasingly more routine.

## DEVELOPING AND BENCHMARKING RNA-SEQ PROTOCOLS

**A**s with any technology there may be certain limitations and RNA-seq is certainly no exception. Even though RNA-seq is a catch-all term for various protocols, the main issue to consider is, and likely will continue to be, the balance between breadth (i.e. number of samples) and depth (i.e. the information from each sample) of the protocol (Stark et al., 2019). This balance will greatly determine the cost efficiency and ultimately the conclusions that can be drawn from the obtained data. With this in mind, I set out to address some of these limitations by contributing to the development of sensitive protocols that remain at their core cost-efficient and flexible.

Firstly, I tackled single cell RNA-seq by contributing to a systematic optimization of single cell RNA barcoding and sequencing (SCRB-seq) (Soumillon et al., 2014) which led to developing molecular crowding SCRB-seq (mcSCRB-seq). mcSCRB-seq exhibits high sensitivity, power, and accuracy, all while remaining a very flexible and low-cost option for many researchers. Chiefly, adding polyethylene glycol (PEG) during the reverse transcription reaction considerably enhanced cDNA synthesis and increased the number of detected genes. I then helped benchmark mcSCRB-seq against other scRNA-seq protocols via publicly available sequencing data containing widely used External RNA Controls Consortium (ERCC) spike-in molecules (Pine et al., 2016; Hardwick et al., 2017), which showed that mcSCRB-seq was among the most sensitive protocols investigated. Finally, I assisted in exemplifying the method's strengths by using human peripheral blood mononuclear cells (PBMCs).

Secondly, I set out to benchmark a variant protocol of mcSCRB-seq, called gmcSCRB-seq, by contributing to a thirteen-method comparison study for the Human Cell Atlas. This multi-lab consortium was interested in determining how well a protocol performs for the purpose of constructing cell atlases of tissues and whole organisms. Within this study, each participating group generated sequencing data from the same het-

erogeneous sample, which included numerous cell-types. I helped find that all protocols performed similarly in terms of clusterability, mappability, and mixability, however, there were noticeable differences between the protocols with regards to gene detection and marker expression. Therefore, these two components drove overall performance and led to Quartz-seq2 (Sasagawa et al., 2018) having the overall best performance among all methods with respect to constructing cell atlases as part of a greater consortium.

Lastly, I turned my focus back to bulk RNA-seq as it is a widely used method and will continue to be widely used. Much of the optimization that has occurred in the field of RNA-seq has focused on improving single-cell methods. Therefore, by using SCRB-seq (Soumillon et al., 2014) and mcSCRB-seq (Bagnoli et al., 2018) as a basis, I contributed in developing prime-seq, a bulk protocol that is sensitive, robust, and cost efficient, thereby making it ideal for research. I then helped in benchmarking prime-seq against TruSeq data produced during the MAQC-III study (SEQC/MAQC-III Consortium, 2014), and exemplified its strengths in two proof-of-principle experiments.

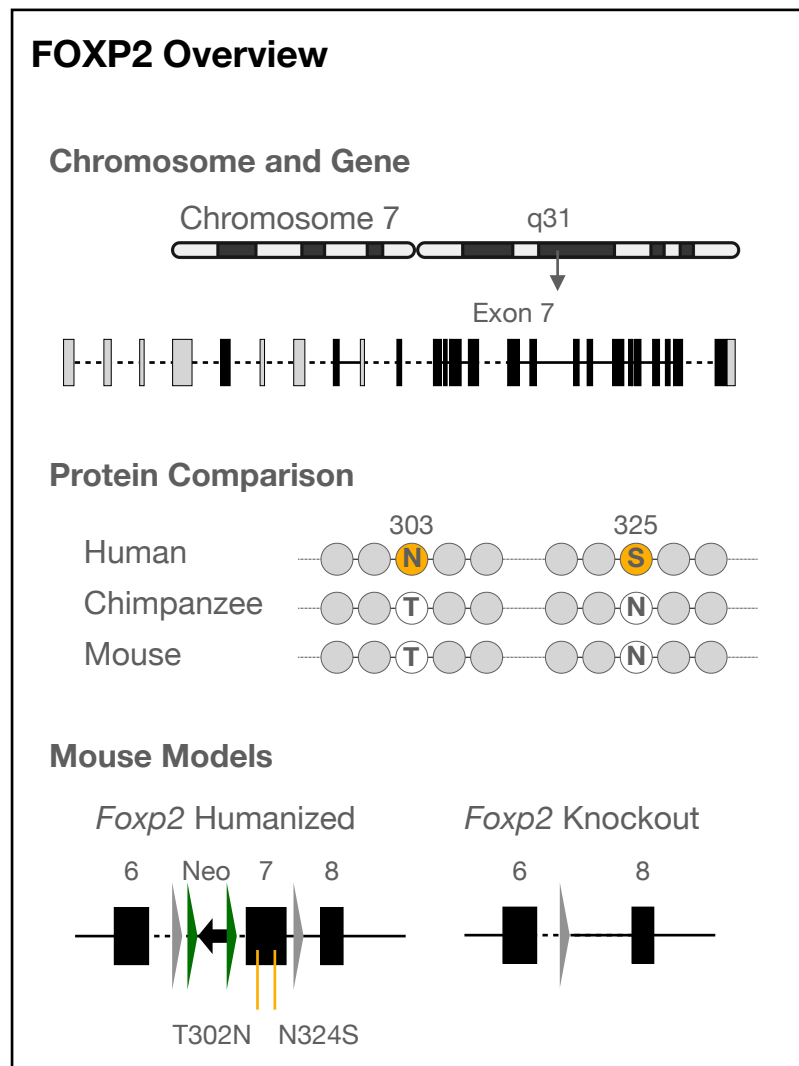
The development of RNA-seq has undeniably altered the field of genomics. For example, as of 2021, over thirty-eight thousand published studies in PubMed used RNA-seq. What is more impressive is that the field is continuing to evolve, with optimizations in experimental methodology and developments in computational analysis. This work, in particular, contributes directly to the progress of this technology. mcSCRB-seq has been cited fifty-four times (CrossRef) and the step-by-step protocol (Bagnoli et al., 2018) has been accessed over forty-seven thousand times to date; additionally some researchers have used optimizations from mcSCRB-seq in order to improve their own protocols, such as using PEG during reverse transcription (Hagemann-Jensen et al., 2020). prime-seq, on the other hand, has been used to successfully generate RNA-seq libraries from 17 organisms across 132 experiments, which ultimately resulted in 19 publications, 3 pre-prints, and 6 conference posters. The success of prime-seq is even more noteworthy when one takes into account that the method has thus far only been published as a pre-print and step-by-step protocol (Janjic et al., 2020), highlighting its robustness, ease of use, and appeal across various fields of research.

## INVESTIGATING *FOXP2* AND ITS ROLE IN THE EVOLUTION OF HUMAN SPEECH USING GLOBAL TRANSCRIPTOMICS

**G**ene expression determines an observable phenotype, and using RNA-seq to quantify expression allows for efficient and powerful studies to be performed linking the two. One phenotype, in particular, that has long interested humans is our ability to speak. Understanding the genetic reason behind why humans have evolved complex language and our closest relatives have not is a question of extreme curiosity but substantially less insight. Forkhead box protein P2 (*FOXP2*) is a transcription factor encoded by the *FOXP2* gene that is essential in immune function, development, and possibly most interesting, necessary for proper speech (Kim et al. 2019; Lam et al. 2013).

The role of *FOXP2* in human speech was first observed in a loss-of-function state (i.e. mutant allele) in a three-generation pedigree (Fisher et al., 1998; Lai et al., 2001). Those heterozygous for a non-functional copy of *FOXP2* exhibited developmental verbal dyspraxia, with all aspects of speech affected (Hurst et al., 1990; Vargha-Khadem et al., 1995). As this was the first genetic link found to human speech, further studies investigated if there was an evolutionary explanation as to why *FOXP2* may be necessary for proper speech development. Interestingly, human *FOXP2* differs from the chimpanzee allele, as well as mouse, at two amino acid substitutions (exon 7, position 303 and 325) (Enard et al., 2002).

In order to properly elucidate the function of both *FOXP2* and the human-specific mutations, in vivo experiments are essential. However, such experiments would be impossible in humans and exceptionally difficult in primates. Fortunately, evolutionary approaches have shown us that in such cases mouse models can be appropriate, specifically as mouse *Foxp2* is functionally the same as the ancestral version. Therefore, a mouse model was previously developed to investigate the effect of a loss-of-function due to a non-functional allele (*Foxp2<sup>wt/ko</sup>*) and the effect of a gain-of-function due to the human-specific amino acid substitutions (*Foxp2<sup>hum/hum</sup>*) (Enard et al., 2009) (Figure 4).



**Figure 4. *FOXP2* Overview.** *FOXP2* is found on chromosome 7q31.1, with Exon 7 being of particular interest due to the location of two human specific mutations resulting in a change from asparagine to threonine on position 303 and serine to asparagine on position 325. The developed mouse models to study *FOXP2* function and evolution possess the human specific substitutions in the humanized model and a removal of exon 7 in the knockout model.

Overall the mice are healthy and do not exhibit any obvious differences, however, less pronounced phenotypic changes were observed including: changes to ultrasonic vocalizations, decreased exploratory behavior, decreased dopamine concentrations, increased dendrite length in medium spiny neurons of the striatum, and increased neural plasticity (Enard et al., 2009). Based on these findings, as well as the fact that the opposite findings were observed in the knockout mice, *FOXP2* has been implicated to be involved in the cortico-basal ganglia (CGB) circuit. Confirmatory studies showed that learning and striatal neuroplasticity were specifically affected, which may be explained by the transition from declarative to procedural learning (Schreiweis et al., 2014). Thus, faster behavioral automatization could explain the link between *FOXP2* and the evolution of human speech.

Although both the *Foxp2* knockout and humanized mouse model have been extensively studied and their phenotypes observed in great detail (Enard et al., 2009), global transcriptomic analysis across various tissues is still lacking. In order to address this gap, I contributed in carrying out a comprehensive study to characterize the effect of *FOXP2* in most major tissues of the mouse models. To create this expression atlas I helped sampled eighteen tissues in eight mice per genotype (wild-type, knockout, and humanized), for a total of twenty-four mice. I then used prime-seq on all 421 samples simultaneously and generated eighteen libraries which were then analyzed using quality filtering, power analysis, differential gene expression analysis, functional annotation analysis, regulatory network inference, and transcription factor binding motif identification.

Within the expression atlas, I observed the strongest signal in the lungs of knockout mice compared to wild-type mice. This was also the only tissue present where *Foxp2* was a differentially expressed gene. I also observed strong differences in expression within the brain of our humanized mice compared to the wild-type mice, where numerous differentially expressed genes were related to neurite outgrowth and control. Although the scale of this study is already larger than any previous *FOXP2* transcriptional analysis, I helped confirm that the effect signal of *FOXP2* is rather small, as was hypothesized in Enard et al. (2009) and Schreiweis et al. (2014).



## STUDY RATIONALE

In their pivotal review, *RNA sequencing: the teenage years*, Stark, Grzelak, and Hadfield state “any predictions of how RNA-seq might develop over the next decade are likely to be too conservative” (2019, p. 652). Additionally it will be difficult to predict if standard practices feature primarily one protocol, as was the story with NGS, or a plethora of methods each with their own strengths. However, what is clear is that this work, through the development and benchmarking of mcSCRB-seq, as well as the development and use of prime-seq will have contributed substantially to the innovation that typifies the field of genomics. Additionally, the application of prime-seq to investigate the effect of *FOXP2* within this work not only serves as exemplary data to detail each component of the RNA-seq workflow, but also serves as a blueprint for future studies looking to utilize the power of RNA-seq to its fullest potential.



# RESULTS

## SENSITIVE AND POWERFUL SINGLE-CELL RNA SEQUENCING USING MCSCRB-SEQ.

### 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 SCRБ-seq (mcSCRБ-seq), which we show to be one of the most sensitive, efficient, and flexible scRNA-seq methods to date.

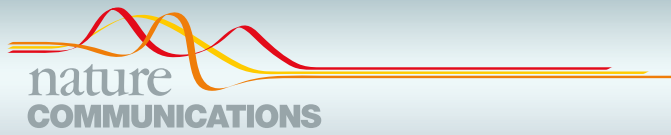
### Declaration of Contribution

CZ and WE conceived the study. JWB, CZ, **AJ** and LEW performed experiments and prepared sequencing libraries. JG and JWB cultured mouse ES and human iPS cells. Sequencing data were processed by SP and CZ. JWB, CZ, **AJ** and BV analyzed the data. JWB., CZ, **AJ**, IH and WE wrote the manuscript.

### Availability

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










## 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 SCRБ-seq (mcSCRБ-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))

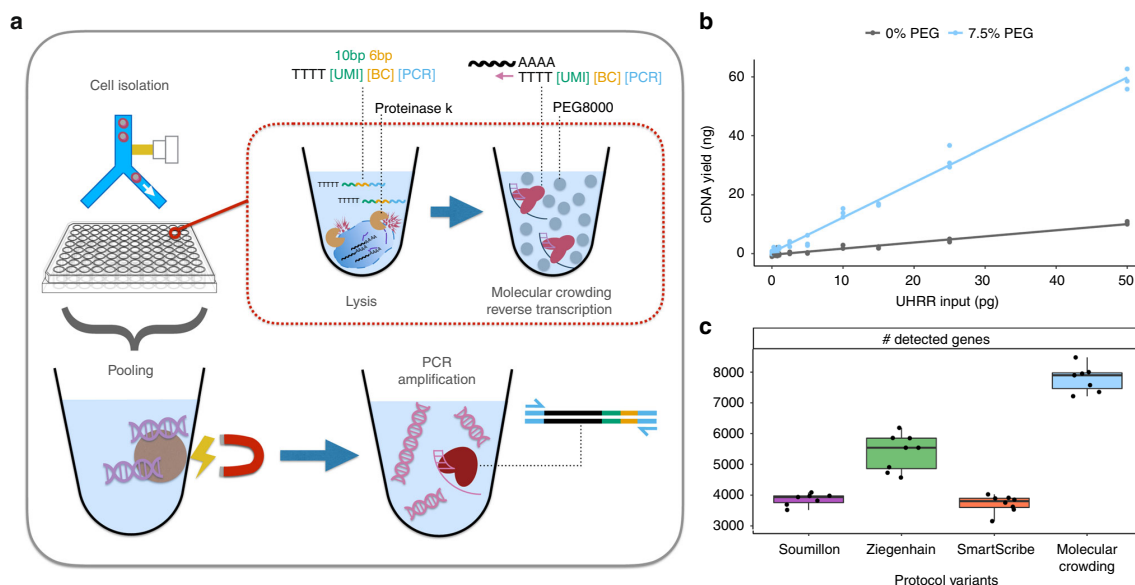
Whole transcriptome single-cell RNA sequencing (scRNA-seq) is a transformative tool with wide applicability to biological and biomedical questions<sup>1,2</sup>. Recently, many scRNA-seq protocols have been developed to overcome the challenge of isolating, reverse transcribing, and amplifying the small 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 SCRБ-seq and generate molecular crowding SCRБ-seq (mcSCRБ-seq), one of the most powerful and cost-efficient plate-based methods to date (Fig. 1a).

## Results

**Systematic optimization of SCRБ-seq.** We started to test improvements to SCRБ-seq by optimizing the cDNA yield and quality generated from universal human reference RNA (UHRR)<sup>9</sup> in a standardized SCRБ-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 SCRБ-seq generate cDNA by oligo-dT priming, template switching, and PCR amplification, we were surprised that these additives decreased cDNA yield for SCRБ-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 SCRБ-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** mcSCRБ-seq workflow and the effect of molecular crowding. **a** Overview of the mcSCRБ-seq protocol workflow. Single cells are isolated via FACS in multiwell plates containing lysis buffer, barcoded oligo-dT primers, and Proteinase K. Reverse transcription and template switching are carried out in the presence of 7.5% PEG 8000 to induce molecular crowding conditions. After pooling the barcoded cDNA with magnetic SPRI beads, PCR amplification using Terra polymerase is performed. **b** cDNA yield dependent on the absence (gray) or presence (blue) of 7.5% PEG 8000 during reverse transcription and template switching. Shown are three independent reactions for each input concentration of total standardized RNA (UHRR) and the resulting linear model fit. **c** Number of genes detected (>=1 exonic read) per replicate in RNA-seq libraries, generated from 10 pg of UHRR using four protocol variants (see Supplementary Table 1) at a sequencing depth of one million raw reads. Each dot represents a replicate (n=8) and each box represents the median and first and third quartiles per method with the whiskers indicating the most extreme data point, which is no more than 1.5 times the length of the box away from the box

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

To test whether these increases in cDNA yield indeed correspond to increases in sensitivity, we generated and sequenced 32 RNA-seq libraries from 10 pg of total RNA (UHRR) using eight replicates for each of the following four SCRb-seq protocol variants (Supplementary Tables 1, 2): the original SCRb-seq protocol<sup>8</sup> (“Soumilion”; 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 “Soumilion” 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 SCRb-seq protocol, in particular due to the addition of PEG 8000, increases the sensitivity compared to previous protocol variants at reduced costs.

#### Terra retains more complexity during cDNA amplification.

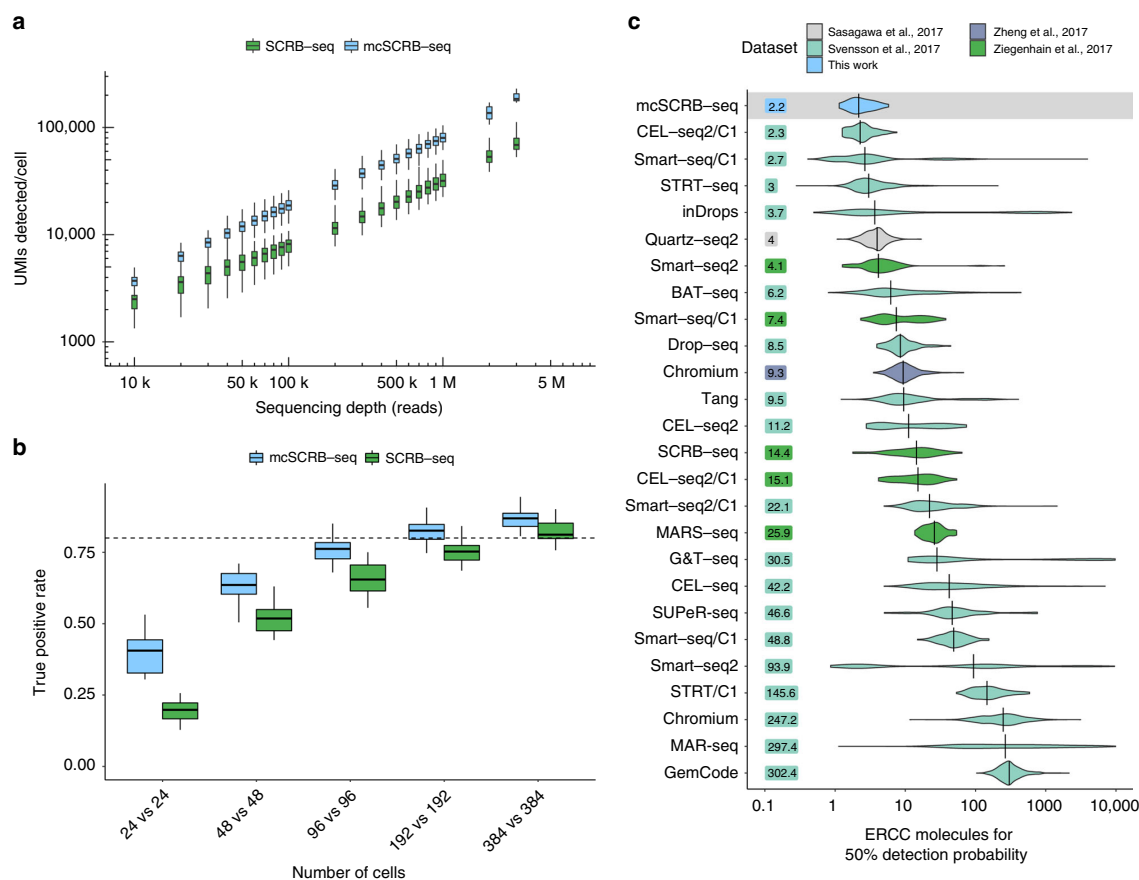
Next, we aimed to increase the efficiency of this protocol by optimizing the cDNA amplification step. Depending on the number of cycles, reaction conditions, and polymerases, substantial noise and bias is introduced when the small amounts of cDNA molecules are amplified by PCR<sup>15,16</sup>. While 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 SCRb-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/μl) from mouse ESCs to prepare libraries with Nextera XT (~0.8 ng needed). Depending on the investigated cells, which may have a lower or higher RNA content than ESCs, the cycle number might need to be adapted to generate enough cDNA while avoiding overcycling.

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

#### mcSCRb-seq increases sensitivity 2.5-fold more than SCRb-seq.

To directly compare the entire mcSCRb-seq protocol to the previously benchmarked SCRb-seq protocol used in Ziegenhain et al.<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, mcSCRb-seq clearly outperforms SCRb-seq and detects 2.5 times as many UMIs per cell at depths above 200,000 reads (Fig. 2a and Supplementary Fig. 8a). At two million reads, mcSCRb-seq detected a median of 102,282 UMIs per cell and a median of 34,760 ERCC molecules, representing 48.9% of all spiked in ERCC molecules (Supplementary Fig. 8b). Assuming that the efficiency of detecting ERCC molecules is representative of the efficiency to detect endogenous mRNAs, the median content per mESC is 227,467 molecules (Supplementary Fig. 8c and 8d), which is very similar to previous estimates using mESCs and STRT-seq, a 5' tagged UMI-based scRNA-seq protocol<sup>20</sup>. As expected, the higher number of UMIs in mcSCRb-seq also results in a higher number of detected genes. For instance, at 500,000 reads, mcSCRb-seq detected 50,969 UMIs that corresponded to



**Fig. 2** Comparison of 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-MLV 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.

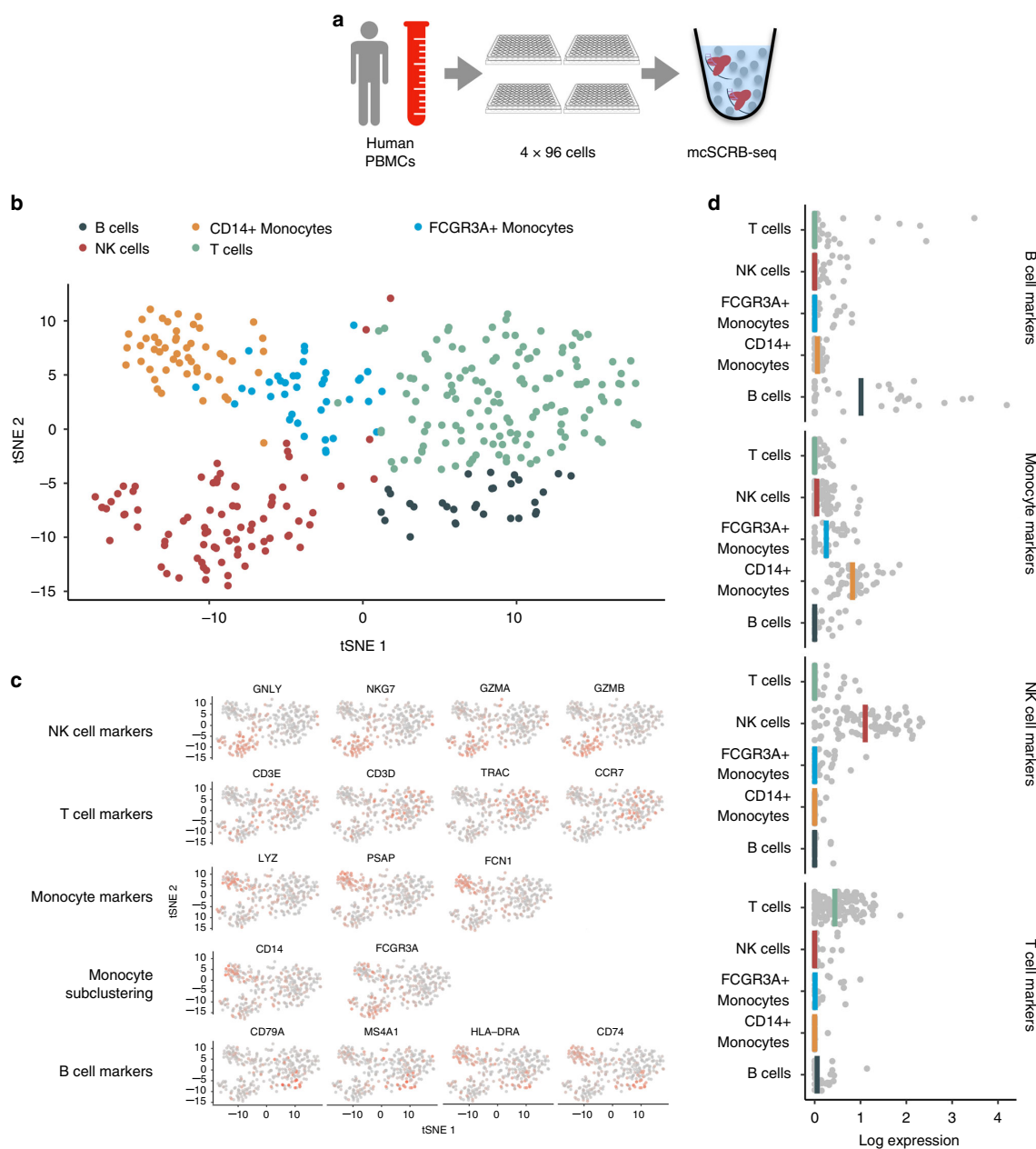
**Comparison of template-switching oligos (TSO).** Unblocked (IDT) and blocked (Eurogentec) template-switching oligonucleotides were compared to determine yield when reverse transcribing 10 pg UHRR and primer-dimer formation without UHRR input. Reaction conditions for RT and PCR were as described above.

**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 (Biotac), 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>8</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.eummc.org](http://www.eummc.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).



**Fig. 3** mSCRB-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 mSCRB-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>.

**SCRB-seq cDNA synthesis.** Cells were dissociated using trypsin and resuspended in 100  $\mu$ l of RNeasy Protect 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>6,8</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.

**mcSCRB-seq cDNA synthesis.** A full step-by-step protocol for mcSCRB-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 (P5NEXTPT5, IDT). Libraries were pooled and size-selected using 2% E-Gel Agarose EX Gels (Life Technologies), cut out in the range of 300–800 bp, and extracted using the MinElute Kit (Qiagen) according to manufacturer's recommendations.

**Sequencing.** Libraries were paired-end sequenced on high output flow cells of an Illumina HiSeq 1500 instrument. Sixteen bases were sequenced with the first read to obtain cellular and molecular barcodes and 50 bases were sequenced in the second read into the cDNA fragment. When several libraries were multiplexed on sequencing lanes, an additional 8 base i7 barcode read was done.

**Primary data processing.** All raw fastq data were processed using zUMIs together with STAR to efficiently generate expression profiles for barcoded UMI data<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 GRCm38.75). Downsampling to fixed numbers of raw sequencing reads per cell were performed using the “-d” option in zUMIs.

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

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

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

**Complex tissue analysis.** PBMCs were obtained from a healthy male donor with written informed consent in accordance with the ethical standards of the responsible committee on human experimentation (216–08, Ethikkommission LMU München) and with the current (2013) version of the Declaration of Helsinki. Cells were sorted into 96-well plates containing 5  $\mu$ 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 ddH<sub>2</sub>O. After the addition of 1  $\mu$ l 2  $\mu$ M barcoded oligo-dT primer (E3V6NEXT, IDT), cDNA was synthesized according to the mcSCRB-seq protocol (see above). Pooling was performed by adding SPRI bead buffer. Sequencing and primary data analysis was performed as described above using the human reference genome (hg38). We retained only high-quality cells with at least 50,000 reads and a mapping rate above 75%. Furthermore, we discarded potential doublets that contained more than 40,000 UMIs and 5000 genes. Next, we used Seurat<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>217</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 SCRB-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.

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).
- SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).
- 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. *protocols.io* <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 iPS cells. Sequencing data were processed by S.P. and C.Z. J.W.B., C.Z., A.J. and B.V. analyzed the data. J.W.B., C.Z., A.J., I.H. and W.E. wrote the manuscript.

## Additional information

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

**Competing interests:** The authors declare no competing interests.

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

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



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

© The Author(s) 2018

# BENCHMARKING SINGLE-CELL RNA-SEQUENCING PROTOCOLS FOR CELL ATLAS PROJECTS.

## 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 multi-center 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.

## Declaration of Contribution

HH designed the study. EM and AL performed all data analyses. CM, AAV and EB prepared the reference sample. CZ, DJM, SP and OS supported the data analysis. MG and IG provided technical and sequencing support. S, DG, JKL, SCB, CS, AO, RCJ, KK, CB, YT, YS, KT, TH, CB, CF, SS, TT, CC, XA, LTN, AR, JZL, **AJ**, LEW, JWB, WE, RS and IN provided sequencing-ready single-cell libraries or sequencing raw data. HH, EM and AL wrote the manuscript with contributions from the co-authors. All authors read and approved the final manuscript.

## Availability

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



# 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>12</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)

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 capacity 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 µm; HEK293 cells: ~15 µm) 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

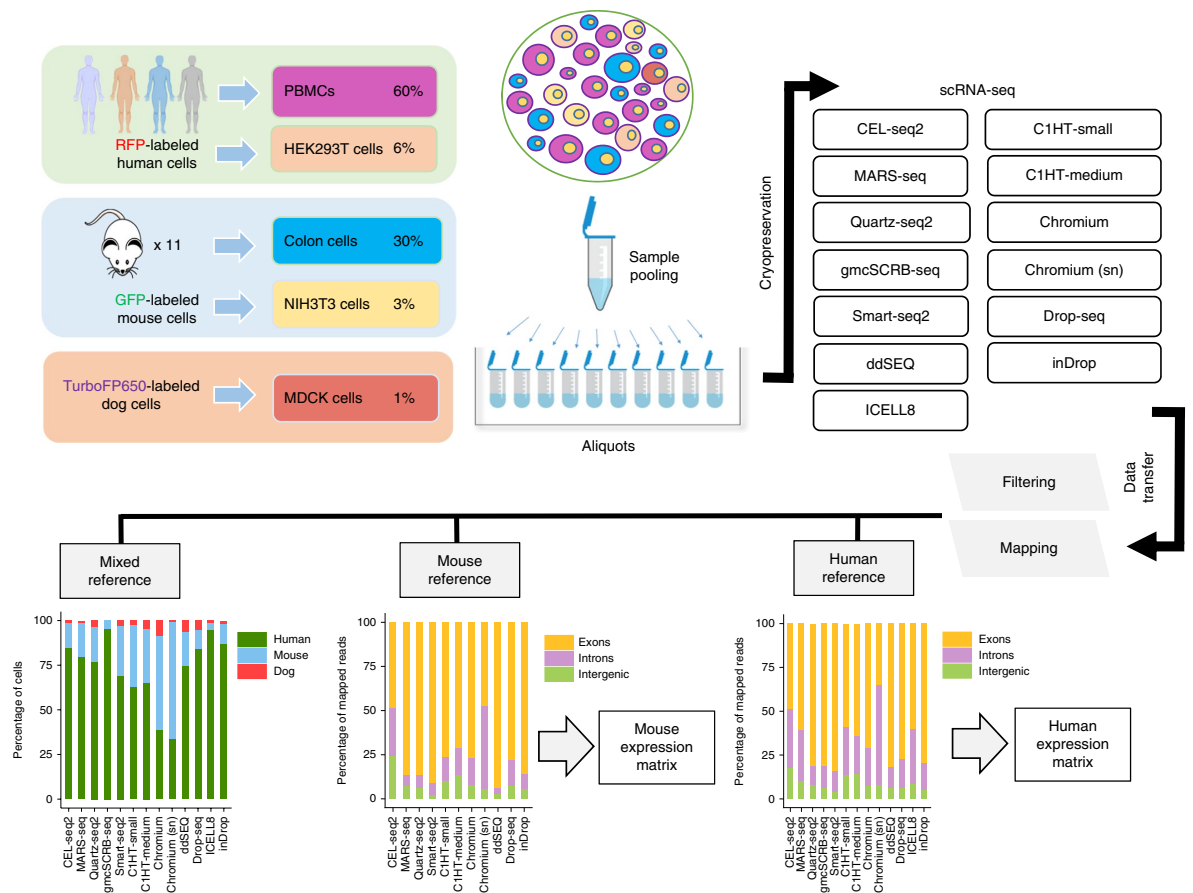
introduced during sample preparation. In addition to the intra-tissue complexity, the fluorescence-labeled, spiked-in cell lines allowed us to monitor cell-type composition during sample processing, and to identify batch effects and biases introduced during cell capture and library preparation.

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



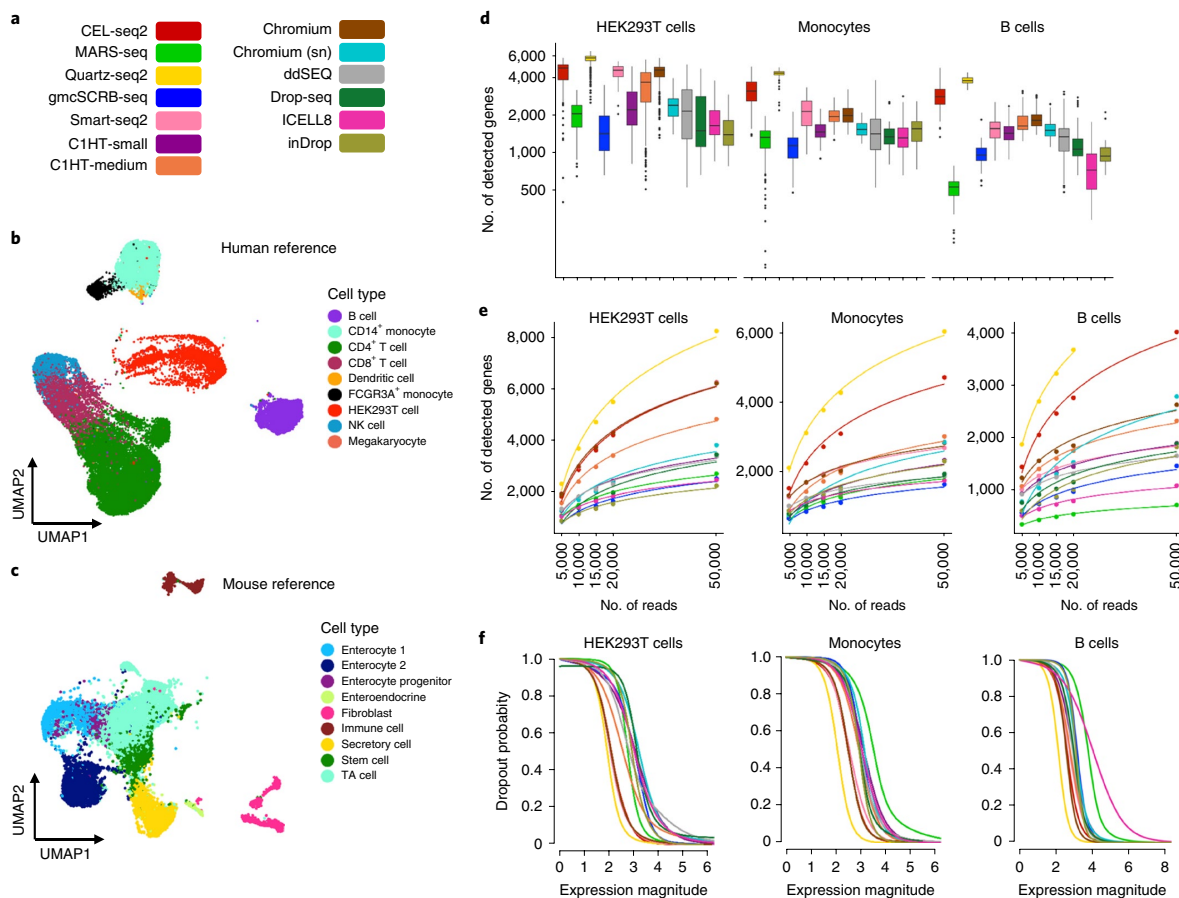
**Fig. 1 | Overview of the experimental design and data processing.** The reference sample consists of human PBMCs (60%), and HEK293T (6%), mouse colon (30%), NIH3T3 (3%) and dog MDCK cells (1%). The sample was prepared in one single batch, cryopreserved and sequenced by 13 different sc/sRNA-seq methods. 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 gmcSCRB-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 v2; 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/sRNA-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



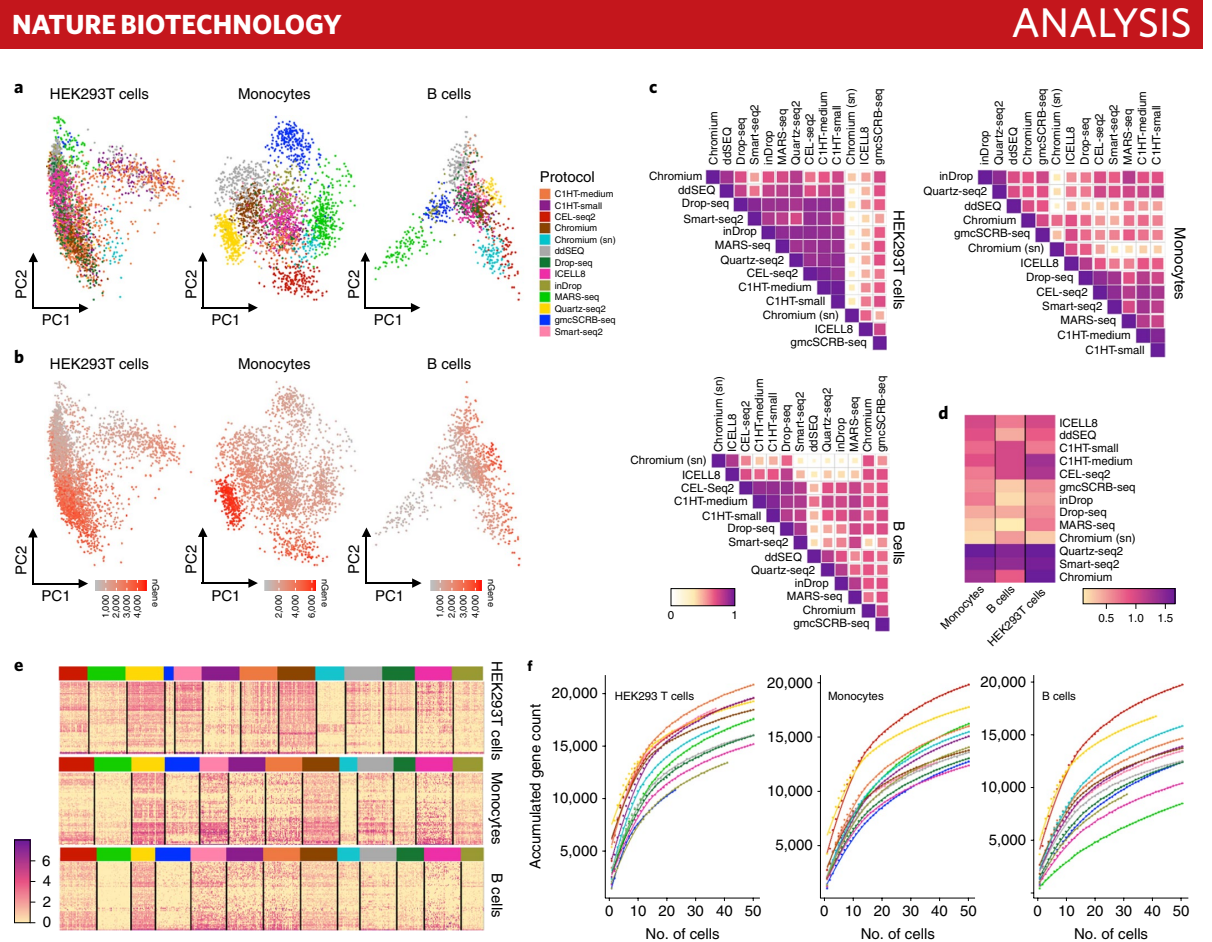
**Fig. 2 | Comparison of 13 sc/snRNA-seq methods.** **a**, Color legend of sc/snRNA-seq protocols. **b**, UMAP of 30,807 cells from the human reference sample (Chromium) colored by cell-type annotation. **c**, UMAP of 19,749 cells from the mouse reference (Chromium) colored by cell-type annotation. **d**, 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 v.2, 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 v.2 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



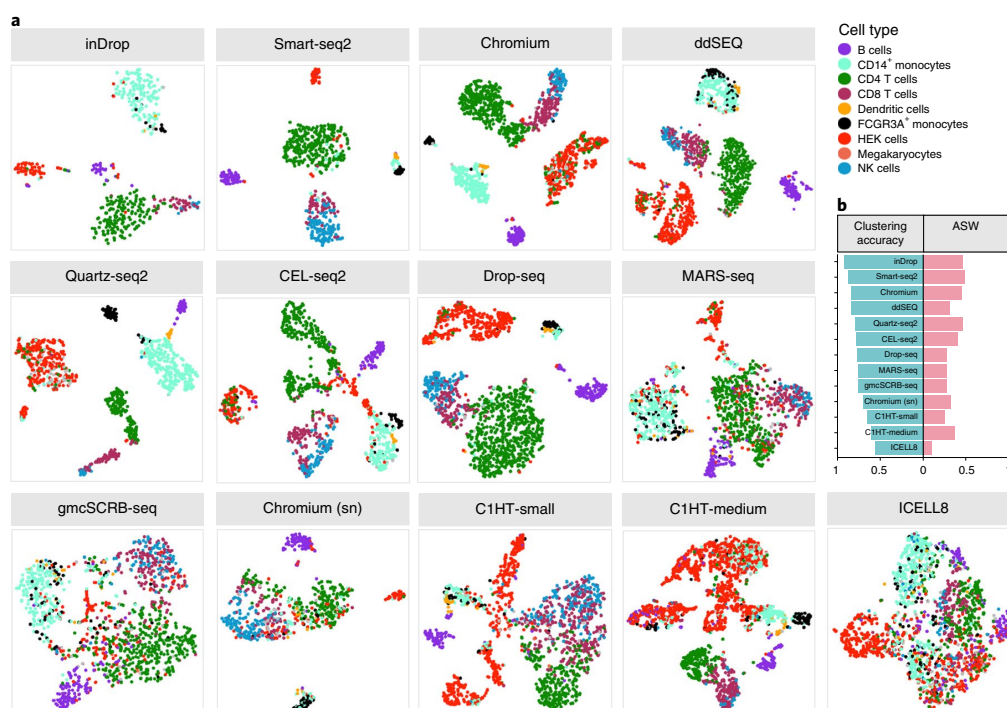
**Fig. 3 | Similarity measures of sc/snRNA-seq methods.** **a, b**, Principal component analysis on down-sampled data (20,000) using highly variable genes between protocols, separated into HEK293T cells, monocytes and B cells, and color coded by protocol (**a**) and number of detected genes per cell (**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(expression) values of cell-type-specific reference markers for down-sampled (20,000) HEK293T cells, monocytes and B cells. **e**, Log(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.

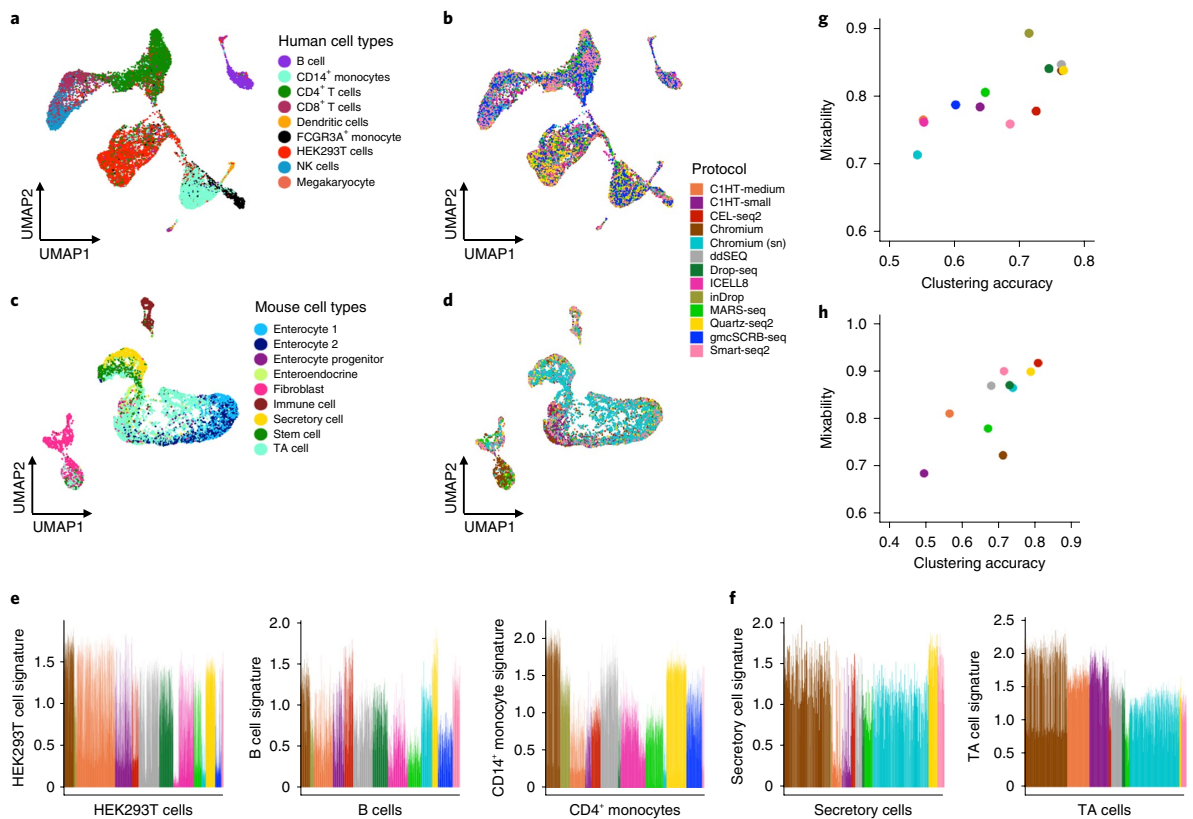
have higher power for cell-type identification. As marker genes are particularly important for data interpretation (for example, annotation), low marker detection levels could severely limit the interpretation of poorly explored tissues, or when trying to identify subtle differences across subpopulations. 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>24</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 integrability 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).

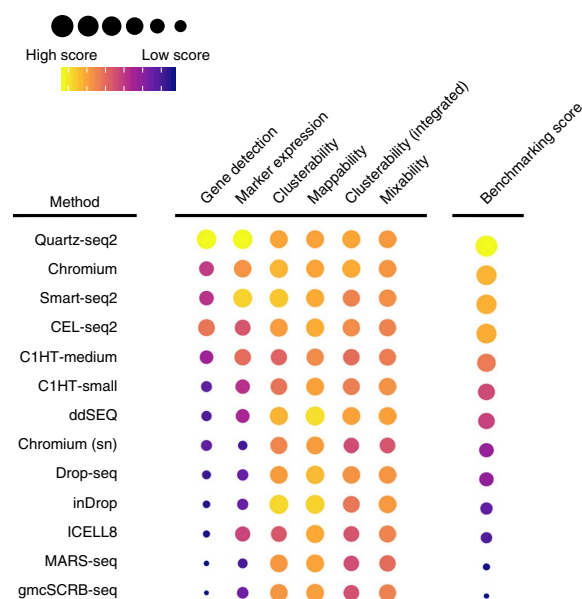
Cell atlas datasets will serve as a reference for annotating cell types and states in future experiments. Therefore, we assessed cells' ability to be projected on to our reference sample (Fig. 2b,c). We used the population signature model defined by matchScore2 and evaluated the protocols based on their 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, integrability 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 integrable 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. 37), 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 gmcSCRB-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).
- Karaiskos, 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**, eaaq1723 (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).
- Alioti, 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. Cvanden 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 mCSCRB-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

## Methods

**Ethical statement.** The present study was approved by the Parc de Salut MAR Research Ethics Committee (reference no. 2017/7585/1) 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% (v/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  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , 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  $M(N, \pi = (\pi_1, \dots, \pi_n))$ , 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 gmSCR-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

SCDE's `scde.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>47</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/elimereu/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 Manton 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).

# PRIME-SEQ, EFFICIENT AND POWERFUL BULK RNA-SEQUENCING

## Abstract

With the advent of Next Generation Sequencing, RNA-sequencing (RNA-seq) has become the major method for quantitative gene expression analysis. Reducing library costs by early barcoding has propelled single-cell RNA-seq, but has not yet caught on for bulk RNA-seq. Here, we optimized and validated a bulk RNA-seq method we call prime-seq. We show that with respect to library complexity, measurement accuracy, and statistical power it performs equivalent to TruSeq, a standard bulk RNA-seq method, but is four-fold more cost-efficient due to almost 50-fold cheaper library costs. We also validate a direct RNA isolation step that further improves cost and time-efficiency, show that intronic reads are derived from RNA, validate that prime-seq performs optimal with only 1,000 cells as input, and calculate that prime-seq is the most cost-efficient bulk RNA-seq method currently available. We discuss why many labs would profit from a cost-efficient early barcoding RNA-seq protocol and argue that prime-seq is well suited for setting up such a protocol as it is well validated, well documented, and requires no specialized equipment.

## Declaration of Contribution

**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 bar-coded 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.

## Availability

<https://doi.org/10.1101/2021.09.27.459575>



bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

## 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>, Wolfgang Enard<sup>1+</sup>

\* contributed equally

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

<sup>2</sup> Graduate School of Systemic Neurosciences, Department of Biology II, 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, Germany

<sup>5</sup> Department of Pediatrics, Dr. von Hauner Children's Hospital, LMU, Munich, Germany

<sup>6</sup> Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden

<sup>+</sup> Corresponding author, Lead contact:

Wolfgang Enard

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)

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

## Abstract

With the advent of Next Generation Sequencing, RNA-sequencing (RNA-seq) has become the major method for quantitative gene expression analysis. Reducing library costs by early barcoding has propelled single-cell RNA-seq, but has not yet caught on for bulk RNA-seq. Here, we optimized and validated a bulk RNA-seq method we call prime-seq. We show that with respect to library complexity, measurement accuracy, and statistical power it performs equivalent to TruSeq, a standard bulk RNA-seq method, but is four-fold more cost-efficient due to almost 50-fold cheaper library costs. We also validate a direct RNA isolation step that further improves cost and time-efficiency, show that intronic reads are derived from RNA, validate that prime-seq performs optimal with only 1,000 cells as input, and calculate that prime-seq is the most cost-efficient bulk RNA-seq method currently available. We discuss why many labs would profit from a cost-efficient early barcoding RNA-seq protocol and argue that prime-seq is well suited for setting up such a protocol as it is well validated, well documented, and requires no specialized equipment.

## 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 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 (Figure 1, Figure 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 four-fold 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 (Table S1, Figure S2). Additionally it has low setup costs as it does not require specialized equipment and is well validated and documented. Hence, it will be a very useful protocol for

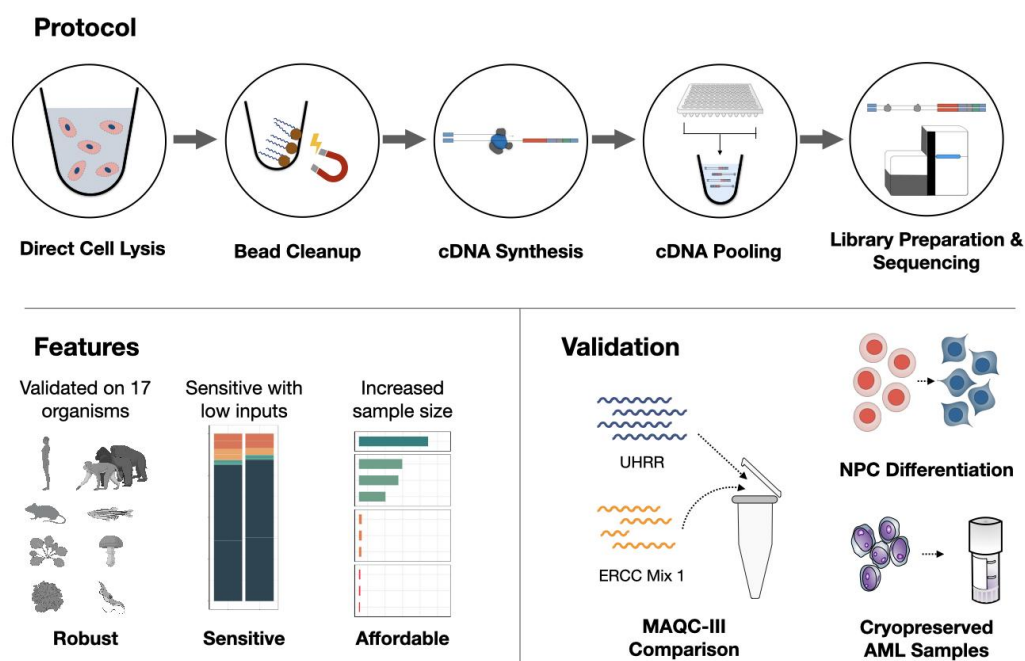
many labs or core facilities that quantify gene expression levels on a regular basis and have no cost-efficient protocol available yet.

## Results

### Development of the prime-seq protocol

The prime-seq protocol is based on the scRNA-seq method SCRB-seq [44] and our optimized derivative mcSCRB-seq [11]. It uses the principles of poly(A) priming, template switching, early barcoding, and UMIs to generate 3' tagged RNA-seq libraries (Figure 1 and Figure 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://dx.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 >4,800 RNA-seq libraries in 97 projects from vertebrates (mainly mouse and human), plants, and fungi (Table S1 and Figure 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 (Table S1). Notably, a large fraction (21%) of all UMIs map to introns with considerable variation among samples (Figure 2A).

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.



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

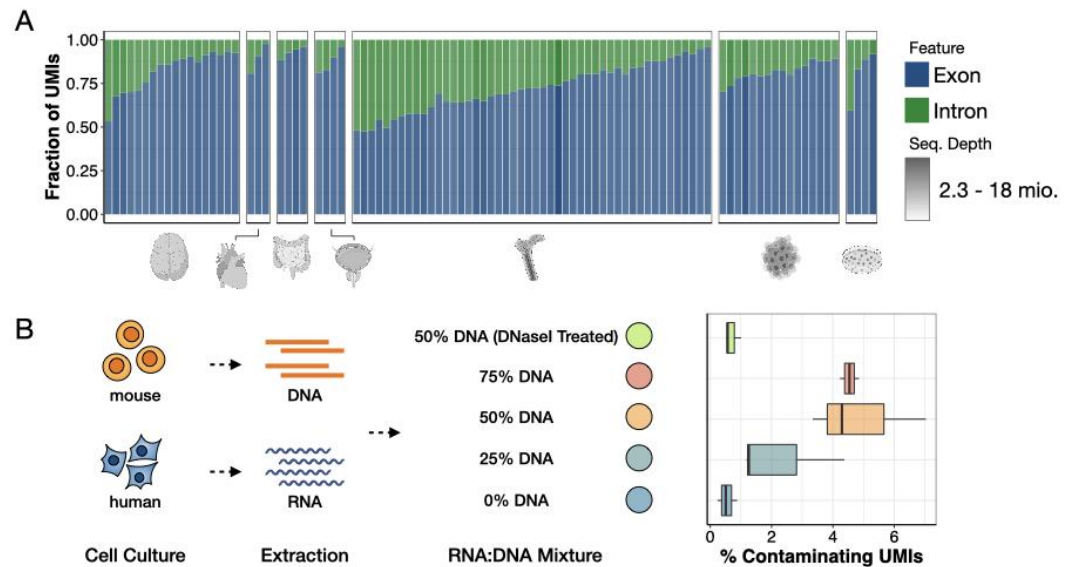
About 8,000 genes are detected only by exonic reads, ~ 8,000 by exonic and intronic reads, and ~ 4,000 by intronic reads only (Figure 2B, Table S1). Intronic reads correlate well with exonic reads of the same gene in scRNA-seq [45] and bulk RNA-seq data sets [46] and intronic reads

are also used to infer expression dynamics in scRNA-seq data [47]. Hence, intronic reads can in principle be informative for quantifying gene expression. However, 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 1,000 HEK cells, this resulted in ~5 ng of “cDNA” generated from gDNA in addition to the 12-32 ng of cDNA generated from RNA. (Figure 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 (Figure 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 (Figure 2B). In the sample that did not contain any mouse DNA, ~0.5% of all exonic and intronic UMIs mapped to the mouse genome, which represents the background level due to mismapping. Of the human mapped reads in this sample, ~70% mapped to exons or introns and 10% to intergenic regions. (Figure S3B). Importantly, the DNase I treated sample had the same distribution of mapped UMIs (0.7% mapped to mouse), 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.

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.



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

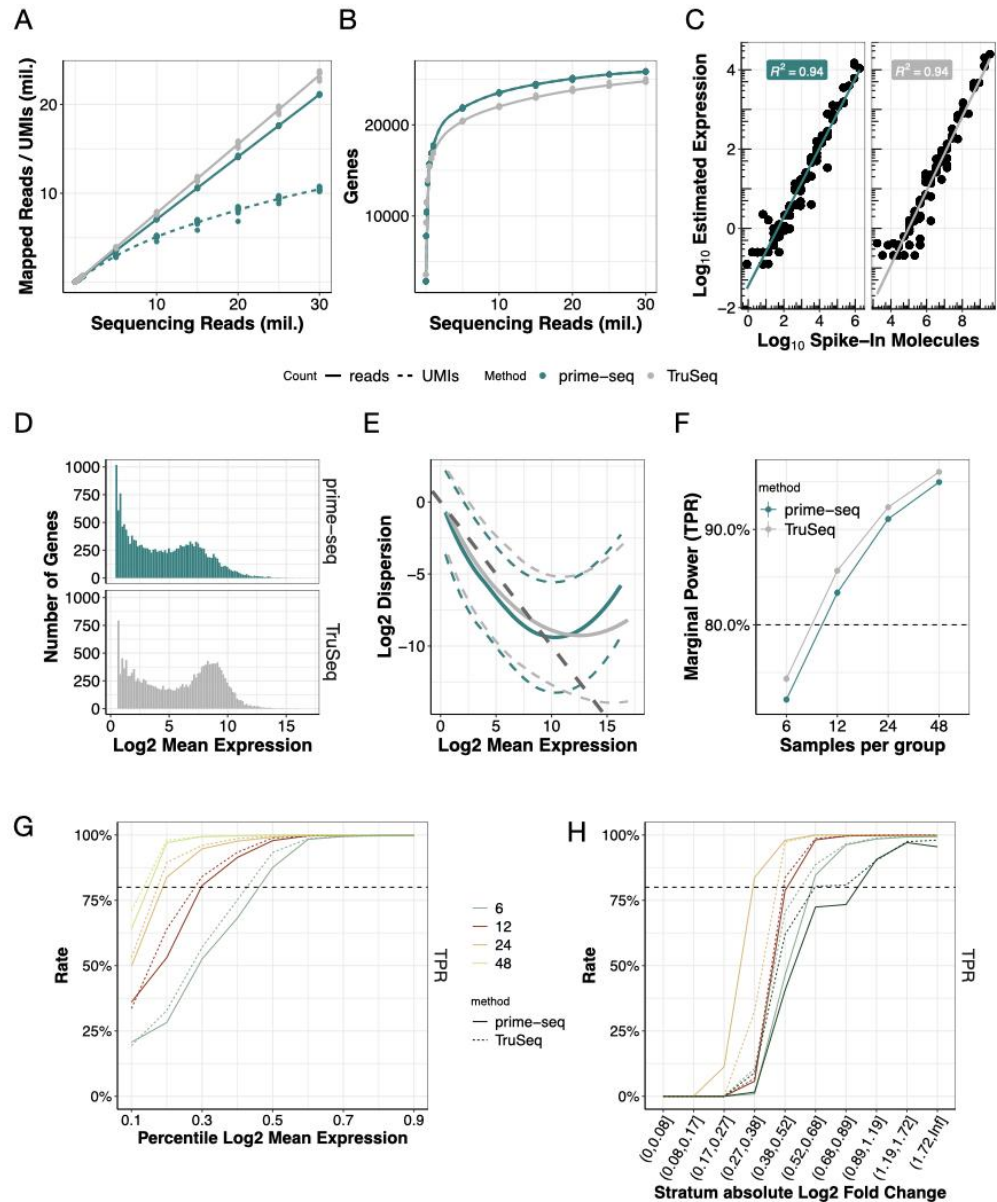
As expected, with increasing amounts of mouse DNA the proportion of mouse mapped UMIs increased (Figure 2B), but even with 75% of the sample being mouse DNA, only 4.5% of the UMIs map to the mouse genome, suggesting that also for gDNA containing samples the impact of genomic reads on expression levels is likely small. Notably, with increasing amounts of gDNA, the fraction of unmapped reads also increased (Figure 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. 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 (Figure S4A) in the same ratio but at a 1,000-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).

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.



**Figure 3. prime-seq has similar sensitivity and power compared to TruSeq (MAQC-III data).** (A) Mapped reads, UMIs (dashed line, only prime-seq), and (B) detected genes at varying sequencing depths between TruSeq data from the MAQC-III Study and matched prime-seq data, shows prime-seq and TruSeq are similarly sensitive (filtering parameters:

## Results

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

*detected UMI  $\geq 1$ , detected gene present in at least 25% of samples and is protein coding).*  
(C) Accuracy, measured by spike-in molecules, is similarly high in both methods ( $R^2=0.94$ ).  
(D) The distribution of genes across mean expression is similar for both methods, as well as the (E) dispersion, which follows a poisson distribution (dark grey dashed line) for lower expressed genes and then increases as technical variation increases for highly expressed genes. The local polynomial regression fit between mean and dispersion estimates per method is shown in solid lines with 95% variability band per gene shown in dashed lines. (F) Power analysis at a sequencing depth of 10 million reads shows almost identical power between prime-seq and TruSeq, and a similar increase at varying sample size for (G) mean expression and (H) absolute log2 fold change. Data filtering parameters: detected UMI  $\geq 1$ , detected gene present in at least 25% of samples.

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%; Figure 3A, Figure S5A). But despite the slightly lower number of reads that can be used, prime-seq does detect at least as many genes as TruSeq (Figure 3B). Both methods also show a similar distribution of gene expression levels (Figure 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 (Figure 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 (Figure 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 (Figure 3E). Prime-seq shows a slightly lower variance for medium expression levels where most genes are expressed and a higher one for a handful of genes with very high expression levels (Figure 3E). 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 Figure S5B). 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 (Figure 3F), the expression levels (Figure 3G), and the fold changes (Figure 3H) 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 (Figure S5C) and the (marginal) TPR across all expression levels and fold changes is 80% for both methods at ~12 replicates per condition (Figure 3F). The power increases for both methods in a similar manner with increasing expression levels (Figure 3G) and increasing fold

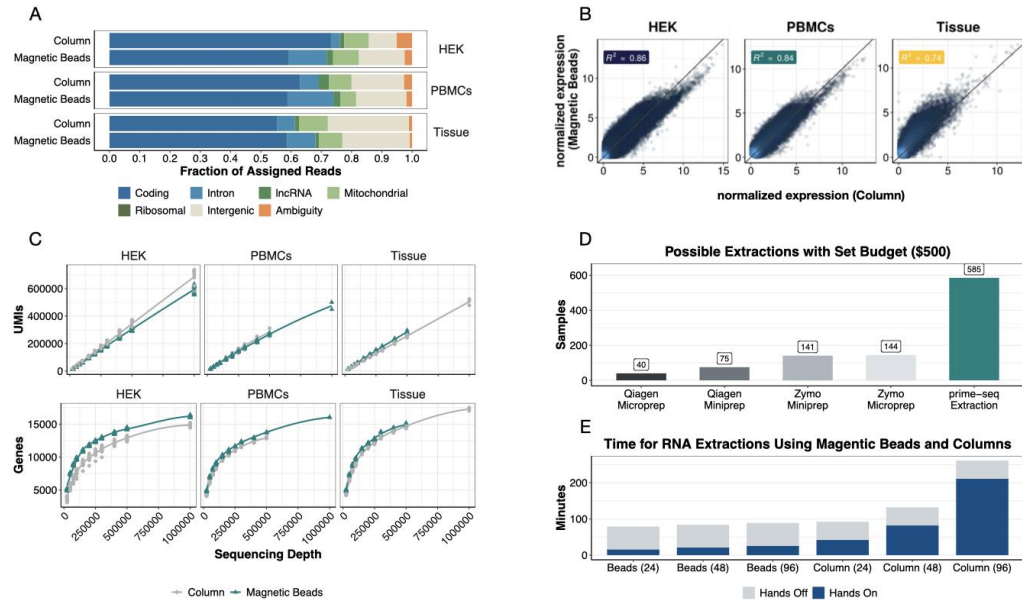
changes (Figure 3H). This is also the case when using only exonic reads for the power analysis (Figure S5C and S5F-G). In summary, prime-seq and TruSeq perform very similarly in estimating gene expression levels with respect to library complexity, accuracy, and statistical power.

### **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 magnetic beads. 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 (Figure 4A and S6) and considerable differences in expression levels (Figure 4B and S7).

To further explore these differences, we tested the influence of the Proteinase K digestion and its associated heat incubation (50°C for 15 minutes and 75°C for 10 minutes), 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 (Figure S6A and B).

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.



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

Hence, bead-based extraction does create a different expression profile than column based extraction, especially due to the often necessary Proteinase K incubation step. This confirms the general influence of RNA extraction protocols on gene expression profiles [51]. Importantly, the complexity of the two types of libraries is similar, with a slightly higher number of genes detected

in the bead-based isolation (Figure 4C, Figure S6C and S6D), potentially due to a preference for longer transcripts with lower GC contents (Figure S7C).

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 (Figure 4D, Table S2). In addition, RNA isolation using magnetic beads is twice as fast and without robotics more amenable to high throughput experiments (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 1,000 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 1,000 (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 (Figure S8A). 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; Figure S8B), indicating that expression profiles from 1,000 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 1,000 cells does not compromise the complexity of prime-seq libraries and hence that prime-seq is a very sensitive RNA-seq protocol.

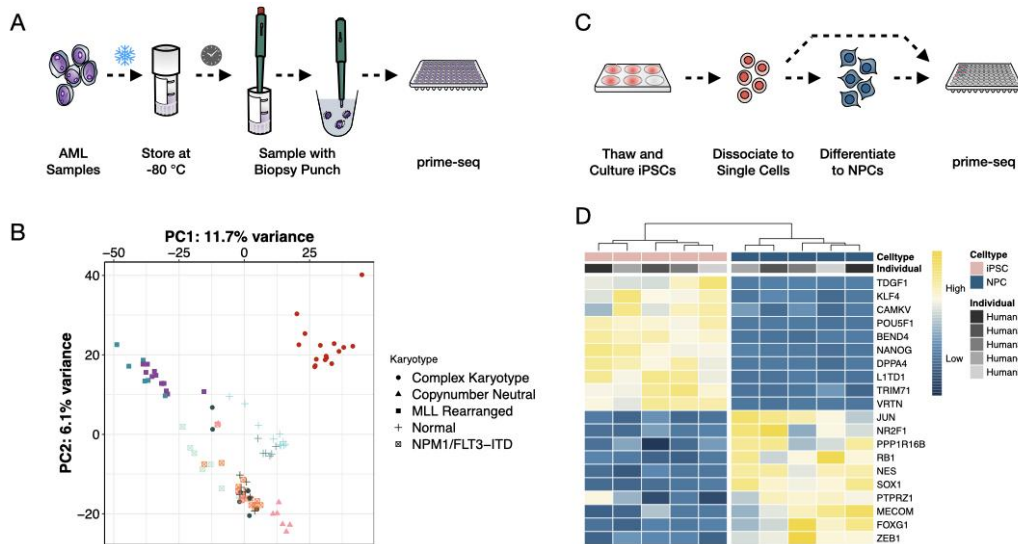
### **Two exemplary applications of prime-seq**

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 [23,52]. These consisted of different donors and AML subtypes and were stored in freezing medium at -80°C for up to 3.5 years (Figure 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 (Figure 5B) as expected [53].

To further exemplify the performance of prime-seq, we investigated its ability to detect known differences in a well established differentiation system [54]. We differentiated five human induced pluripotent stem cell (iPSCs) lines [36] to neural progenitor cells (NPCs) and generated expression profiles using prime-seq (Figure 5C). In a hierarchical clustering of well known marker genes [55], 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 (Figure 5D).

## Results

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.



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

### 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 [9,23–35,42,43,56,57]. 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 [16,22,58–61]. This

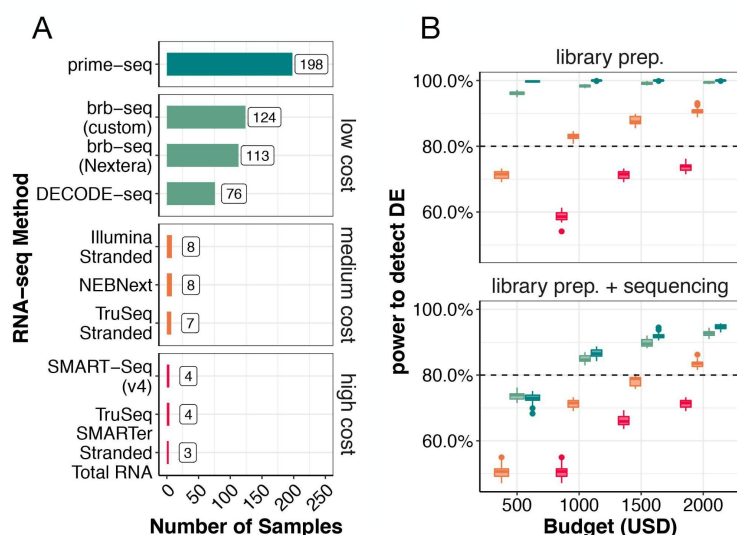
includes BRB-seq that uses poly(A) priming and DNA-Pol I for second strand synthesis and also performs similarly to TruSeq [22]. Decode-seq also uses poly(A) priming and template switching like prime-seq, but adds sample-specific barcodes and UMIs at the 5' end [16]. In a direct comparison, Decode-seq 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 [62] 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, SmartSeq, 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 (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 (Figure 6A). Note that these costs include for all methods \$1.39 per sample for two Bioanalyzer (Agilent) Chips (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) (Figure 6B, Figure S9A). For example, to reach an 80% TPR at a desired FDR of 5%, one needs to spend \$715 including sequencing costs for

## Results

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

prime-seq, \$795 when using Decode-seq, \$1,625 when using Illumina Stranded, and \$3,485 when using TruSeq (Figure S9B).



**Figure 6. Prime-seq is very cost-efficient.**

(A) With a set budget of \$500, prime-seq allows one to process 198 samples, which is 1.6 times more samples than the next cost-efficient method. (B) The compared methods were grouped into low, middle, and high cost methods and

the TruSeq MAQCII data was used as a basis for power analysis for all methods but prime-seq. The increase in sample size due to cost efficiency directly impacts the power to detect differentially expressed genes, as evident by the increased performance of prime-seq and other low cost methods (BRB-seq and Decode-seq), even when sequencing costs are included in the comparison (sequencing depth of 10 mio. reads at a cost of \$3.40 per 1 mio. reads).

Cost-efficiency with respect to time can also matter and we calculated hands-on and hands-off time for the different methods (Table S5). Hands-on times vary from 30-35 minutes for the non-commercial, early barcoding methods to 52-191 minutes 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 NEBNext that individually process RNA samples and generate full-length cDNA profiles (“full-length protocols”) from protocols like prime-seq, Decode-seq, or BRB-seq that use early barcoding and generate 5’ or 3’ tagged cDNA libraries (“tag protocols”).

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

Initially, early barcoding 3’ tagged protocols generated slightly less complex libraries (i.e. detected fewer genes for the same number of reads), especially due to a considerable fraction of unmapped reads [22,63]. 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 (Figure 6) and as argued before [16,22,63], 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 [64] 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 [65], 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.

Finally, while early barcoding and pooling enable the cost-efficiency of tag protocols, this 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://dx.doi.org/10.17504/protocols.io.s9veh66>). Early barcoding and pooling also entails the danger of barcode swapping, i.e. the formation of chimeric molecules during PCR, resulting in a contamination of a cell's expression profile with transcripts from another cell. This is especially an issue for scRNA-seq [66] as the number of PCR cycles and on the polymerase likely play a role [67]. To verify that this is not an issue in prime-seq, we pooled human and mouse samples at each possible point in the protocol; we detected low rates of cross-contamination when samples were pooled as RNA (0.59%), cDNA (0.76%), or libraries (0.83%).

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 1,000 cells (~10-20ng total RNA) as input. However, the decisive advantage of tag protocols is their drastically higher cost-efficiency (Figure 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 10M reads or more (Figure 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 [68]. 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 6,000 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 ~\$2,000 for 96 primers, which will be sufficient for processing more than 240 thousand 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://dx.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

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

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://dx.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 1,000-10,000 cells, were treated with 20 µg of Proteinase K (Thermo Fisher, #AM2546) and 1µL 25 mM EDTA (Thermo Fisher, EN0525) at 50°C for 15 minutes with a heat inactivation step at 75°C for 10 minutes. The samples were then cleaned using cleanup beads, a custom made mixture containing SpeedBeads (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 minutes with a heat inactivation step at 65°C for 5 minutes.

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), 1x Maxima H- Buffer (Thermo Fisher), 1 mM each dNTPs (Thermo Fisher), 1 µM template-switching oligo (IDT), 1 µM barcoded oligo(dT) primers (IDT). The reaction was incubated at 42°C for 90 minutes.

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 minutes followed by a heat inactivation

step at 80 °C for 10 minutes. 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 1x 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 minutes, denaturation at 98 °C for 15 seconds, annealing at 65 °C for 30 seconds, elongation at 68 °C for 4 minutes, and a final elongation at 72 °C for 10 minutes. 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 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 seconds; 10-12 cycles of 98°C for 10 seconds, 65°C for 1 minute 15 seconds, 65°C for 5 minutes; and 65°C for 4 minutes.

## Results

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

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 (Figure 4) as well as the AML experiment (Figure 5B). These libraries were prepared as previously described [11].

Briefly, three replicates of 0.8 ng of DNA were tagged 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 minutes, initial denaturation at 95 °C for 30 seconds, denaturation at 95 °C for 10 seconds, annealing at 62 °C for 30 seconds, elongation at 72 °C for 1 minute, and a final elongation at 72 °C for 5 minutes. 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 bp - 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 [69] 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 [70] 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 [71]) 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,[72]) and the reads counted using RSubread (version 1.32.4,[73]).

## Results

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

### Sensitivity and Differential Gene Expression Analysis of RNA-seq Data

The count matrix generated by zUMIs was loaded into RStudio (version 1.3.1093 [74]) using R (version 4.0.3 [75]), bioMart (version 2.46.0 [76]), dplyr (version 1.0.2 [77]), and tidyr (version 1.1.2 [78]) were used for data processing and calculating descriptive statistics (i.e. detected genes, reads, and UMIs). DESeq2 (version 1.30.0 [79]) was used for differential gene expression analysis. ggplot2 (version 3.3.3 [80]), cowplot (version 1.1.1 [81]), ggbeeswarm (0.6.0 [82]), ggsignif (version 0.6.0 [83]), ggsci (version 2.9 [84]), ggrepel (version 0.9.0 [85]), EnhancedVolcano (1.8.0 [86]), ggpointdensity (version 0.1.0 [87]) and pheatmap (version 1.0.12 [88]) 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, RNAseq 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', RNAseq = '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 (Figure S5B).

### 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

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

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 (1,000 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 laboured breathing. The mice were then removed from the container, and a cervical dislocation was performed. The mice were briefly washed with 80% EtOH, the head decapitated, and the brain removed. The brain was transferred to a dish with ice-cold PBS and placed in a 1 mm slicing matrix.

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

## Results

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

### 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, manufacturer instructions were followed and both the Proteinase K and DNase digestion steps were performed as outlined in the protocol. For the magnetic bead isolation, the prime-seq protocol was used as outlined in the “Magnetic Beads” condition. For “No Incubation” condition the Proteinase K digestion was skipped entirely. For the “Incubation” condition, the Proteinase K digestion was performed but with no enzyme; that is the heat cycling of 50°C for 15 minutes and 75°C for 10 minutes was carried out but no enzyme was added to the lysate.

### gDNA Priming Experiment

For a graphical overview of the gDNA Priming experiment, see Figure 2B. Frozen vials of mouse embryonic stem cells (mESC), that 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 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 Figure S7A. 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.

#### 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 100x NEAA (Thermo Fisher), 100mM Sodium Pyruvate (Thermo Fisher), 50mM 2-Mercaptoethanol (Thermo Fisher) supplemented with

## Results

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

500nM A-83-01 (Sigma Aldrich), 100nM LDN 193189 (Sigma Aldrich) and 30 $\mu$ M Y27632 (biozol). A half-medium change was performed on day 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, counted and 2x10<sup>4</sup> 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 [52]. AML-PDX cells were cryopreserved as 10 Mio cells in 1mL of freezing medium (90% FBS, 10% DMSO) and stored at -80°C for biobanking purposes. To avoid thawing these samples and thus harming or even destroying them, the frozen cell stocks were first transferred to dry ice under a cell culture hood. Next a sterile 1 mm biopsy punch was used to punch the frozen cells in the vial and transfer the extracted cells to one well of a 96 well plate containing 100  $\mu$ 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 1x10<sup>5</sup> 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.

## 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 [57].

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

## Results

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

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

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the ArrayExpress repository under the following accession numbers E-MTAB-10133, 10138-10142, 10175. The code required to generate the figures can be found at <https://github.com/Hellmann-Lab/prime-seq>.

Competing interests

The authors declare that they have no competing interests.

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

## Funding

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.

## 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.

## Acknowledgements

We would like to thank Karin Bauer and Ming Zhao for lab support, Ines Bliesener and Maik 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 Figure 1, Figure 3 and Figure S2 were created with BioRender.com

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

## References

1. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet.* 2019;20:631–56.
2. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell.* 2017;65:631–43.e4.
3. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun.* 2019;10:4667.
4. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc.* 2018;13:599–604.
5. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol.* 2020;38:747–55.
6. Ziegenhain C, Vieth B, Parekh S, Hellmann I, Enard W. Quantitative single-cell transcriptomics. *Brief Funct Genomics.* 2018;17:220–32.
7. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods.* nature.com; 2011;9:72–4.
8. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* 2012;2:666–73.
9. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. *Sci Rep.* 2016;6:25533.
10. Hagemann-Jensen M, Ziegenhain C, Chen P, Ramsköld D, Hendriks G-J, Larsson AJM, et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat Biotechnol.* 2020;38:708–14.
11. Bagnoli JW, Ziegenhain C, Janjic A, Wange LE, Vieth B, Parekh S, et al. Sensitive and powerful single-cell RNA sequencing using 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, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell.* 2015;161:1202–14.
14. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell.* 2015;161:1187–201.
15. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

of cell type deconvolution pipelines for transcriptomics data. *Nat Commun.* 2020;11:5650.

16. Li Y, Yang H, Zhang H, Liu Y, Shang H, Zhao H, et al. Decode-seq: a practical approach to improve differential gene expression analysis. *Genome Biol.* 2020;21:66.

17. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics.* 2014;30:301–4.

18. Lazic SE, Clarke-Williams CJ, Munafò MR. What exactly is “N” in cell culture and animal experiments? *PLoS Biol.* 2018;16:e2005282.

19. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell.* 2017;171:1437–52.e17.

20. Uzbass F, Opperer F, Sönmezer C, Shaposhnikov D, Sass S, Krendl C, et al. BART-Seq: cost-effective massively parallelized targeted sequencing for genomics, transcriptomics, and single-cell analysis. *Genome Biol.* 2019;20:155.

21. Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Zachery Cogan J, et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat Biotechnol.* Nature Publishing Group; 2020;38:954–61.

22. Alpern D, Gardeux V, Russeil J, Mangeat B, Meireles-Filho ACA, Breyse R, et al. BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol.* 2019;20:71.

23. Ebinger S, Özdemir EZ, Ziegenhain C, Tiedt S, Castro Alves C, Grunert M, et al. Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia. *Cancer Cell.* 2016;30:849–62.

24. Schreck C, Istvánffy R, Ziegenhain C, Sippenauer T, Ruf F, Henkel L, et al. Niche WNT5A regulates the actin cytoskeleton during regeneration of hematopoietic stem cells. *J Exp Med.* 2017;214:165–81.

25. Gegenfurtner FA, Zisis T, Al Danaf N, 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 [Internet].* 2018;131. Available from: <http://dx.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.

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

29. Wang S, Gegenfurtner FA, Crevenna AH, Ziegenhain C, Kliesmete Z, Enard W, et al. Chivosazole A Modulates Protein-Protein Interactions of Actin. *J Nat Prod.* 2019;82:1961–70.
30. Ebinger S, Zeller C, Carlet M, Senft D, Bagnoli JW, Liu W-H, et al. Plasticity in growth behavior of patients' acute myeloid leukemia stem cells growing in mice. *Haematologica.* 2020;105:2855–60.
31. Garz A-K, Wolf S, Grath S, Gaidzik V, Habringer S, Vick B, et al. Azacitidine combined with the selective FLT3 kinase inhibitor crenolanib disrupts stromal protection and inhibits expansion of residual leukemia-initiating cells in FLT3-ITD AML with concurrent epigenetic mutations. *Oncotarget.* 2017;8:108738–59.
32. Mulholland CB, Nishiyama A, Ryan J, Nakamura R, Yiğit M, Glück IM, et al. Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals. *Nat Commun.* 2020;11:5972.
33. Redondo Monte E, Wilding A, Leubolt G, Kerbs P, Bagnoli JW, Hartmann L, et al. ZBTB7A prevents RUNX1-RUNX1T1-dependent clonal expansion of human hematopoietic stem and progenitor cells. *Oncogene.* 2020;39:3195–205.
34. Shami A, Atzler D, Bosmans LA, Winkels H, Meiler S, Lacy M, et al. Glucocorticoid-induced tumour necrosis factor receptor family-related protein (GITR) drives atherosclerosis in mice and is associated with an unstable plaque phenotype and cerebrovascular events in humans. *Eur Heart J.* 2020;41:2938–48.
35. LaClair KD, Zhou Q, Michaelsen M, Wefers B, Brill MS, Janjic A, et al. Congenic expression of poly-GA but not poly-PR in mice triggers selective neuron loss and interferon responses found in C9orf72 ALS. *Acta Neuropathol.* 2020;140:121–42.
36. Geuder J, Ohnuki M, Wange LE, Janjic A, Bagnoli JW, Müller S, et al. A non-invasive method to generate induced pluripotent stem cells from primate urine [Internet]. Cold Spring Harbor Laboratory. 2020 [cited 2021 Jan 21]. p. 2020.08.12.247619. 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. BioMed Central;* 2021;22:1–19.
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 [Internet]. bioRxiv. 2020 [cited 2021 Sep 3]. p. 2020.05.02.073577. Available from:

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

<https://www.biorxiv.org/content/10.1101/2020.05.02.073577v1>

41. Kempf JM, Weser S, Bartoschek MD, Metzeler KH, Vick B, Herold T, et al. Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance in AML. *Sci Rep.* 2021;11:5838.
42. Pekayvaz K, Leunig A, Kaiser R, Brambs S, Joppich M, Janjic A, et al. Protective immune trajectories in early viral containment of non-pneumonic SARS-CoV-2 infection [Internet]. Cold Spring Harbor Laboratory. 2021 [cited 2021 Feb 19]. p. 2021.02.03.429351. Available from: <https://www.biorxiv.org/content/10.1101/2021.02.03.429351v1>
43. Kliesmete Z, Wange LE, Vieth B, Esgleas M, Radmer J, Hülsmann M, et al. TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals [Internet]. Cold Spring Harbor Laboratory. 2021 [cited 2021 Feb 19]. p. 2021.02.05.429919. 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 [Internet]. Cold Spring Harbor Laboratory. 2014 [cited 2021 Jan 21]. p. 003236. 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* [Internet]. 2018;7. Available from: <http://dx.doi.org/10.1093/gigascience/giy059>
46. 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* [Internet]. Oxford Academic; 2020 [cited 2021 Jan 21];2. Available from: <https://academic.oup.com/nargab/article-pdf/2/3/lqaa073/34054975/lqaa073.pdf>
47. 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.
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. 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

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

subgroups and in vivo bioluminescence imaging. PLoS One. 2015;10:e0120925.

53. 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.

54. 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.

55. Liu Y, Yu C, Daley TP, Wang F, Cao WS, Bhate S, et al. CRISPR Activation Screens Systematically Identify Factors that Drive Neuronal Fate and Reprogramming. Cell Stem Cell. 2018;23:758–71.e8.

56. 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. American Society of Hematology; 2016;128:602–602.

57. 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.

58. 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.

59. 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.

60. 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.

61. 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.

62. 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.

63. 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.

64. 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.

65. Westermann AJ, Vogel J. Cross-species RNA-seq for deciphering host-microbe interactions. Nat Rev Genet. 2021;22:361–78.

66. Dixit A. Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments [Internet]. bioRxiv. 2021 [cited 2021 Aug 26]. p. 093237. Available from:

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

<https://www.biorxiv.org/content/10.1101/093237v2>

67. 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.

68. 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* [Internet]. 2021;10. Available from: <http://dx.doi.org/10.7554/eLife.66274>

69. Buschmann T, Bystrykh LV. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics.* 2013;14:272.

70. 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.

71. Andrews S. FastQC: A quality control analysis tool for high throughput sequencing data [Internet]. Github; [cited 2021 Sep 14]. Available from: <https://github.com/s-andrews/FastQC>

72. 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.

73. 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.

74. Team R. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA, 2020. 2020.

75. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: <https://www.r-project.org/>

76. Steffen Durinck, Wolfgang Huber. biomaRt [Internet]. Bioconductor; 2017. Available from: <https://bioconductor.org/packages/biomaRt>

77. Wickham H, Francois R, Henry L, Müller K. dplyr: A grammar of data manipulation [Internet]. 2021. Available from: <https://github.com/tidyverse/dplyr>

78. Wickham H, Henry L. Tidy: Tidy messy data. R package version. 2020;1:397.

79. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550.

80. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer New York; 2010.

81. Wilke CO. cowplot: streamlined plot theme and plot annotations for “ggplot2.” 2019.

82. Clarke E, Sherrill-Mix S. ggbeeswarm: Categorical Scatter (Violin Point) Plots [Internet]. 2017. Available from: <https://CRAN.R-project.org/package=ggbeeswarm>

83. Constantin A-E, Patil I. ggsignif: R Package for Displaying Significance Brackets for “ggplot2” [Internet]. PsyArxiv. 2021. Available from: <https://psyarxiv.com/7awm6>

## Results

bioRxiv preprint doi: <https://doi.org/10.1101/2021.09.27.459575>; this version posted September 28, 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-NC-ND 4.0 International license.

84. Xiao N. ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for “ggplot2” [Internet]. 2018. Available from: <https://CRAN.R-project.org/package=ggsci>
85. Slowikowski K. ggrepel: Automatically position non-overlapping text labels with “ggplot2.” 2018.
86. Blighe K, Rana S, Lewis M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version. 2019;
87. Kremer LPM. ggpointdensity: A Cross Between a 2D Density Plot and a Scatter Plot [Internet]. 2019. Available from: <https://CRAN.R-project.org/package=ggpointdensity>
88. Kolde R. Pheatmap: pretty heatmaps [Internet]. 2012. Available from: <https://cran.r-project.org/web/packages/pheatmap/index.html>

# INVESTIGATING LOSS-OF-FUNCTION AND HUMAN EVOLUTION OF *FOXP2* USING GLOBAL TRANSCRIPTOMIC ANALYSIS

## Abstract

*FOXP2* was the first gene implicated in human language and still remains one of the few speech-associated genes. Although this has led to over two decades of investigation, the exact mechanisms are still unknown. *FOXP2* research has primarily focused on the lungs, due to its role in lung development, and the brain as it is involved in cortico-basal ganglia circuitry. However, a comprehensive multi-organ transcriptomic analysis is still lacking. To this end, we utilize a cost-efficient bulk RNA-sequencing method to process 421 samples consisting of 18 tissues. In this study we aim to elucidate the functional role of *FOXP2* and the evolutionary aspect of the human-specific mutations possibly contributing to the evolution of human speech. Therefore, we compare wild-type mice to heterozygous mice with one non-functional *Foxp2* allele (knockout) and homozygous mice with variant *Foxp2* alleles containing two human-specific mutations (humanized). We perform power analysis to determine the effect size of *Foxp2* and observe that it is too small to readily interpret differences between the genotypes in most tissues. Even with such a small effect size, we detect the strongest difference in the lungs of knockout mice, which supports previous findings. Additionally, within the humanized-wild-type comparison for the brain we observe differences related to neuron projection localization and regulation.

## Declaration of Contribution

**AJ** and WE conceived the study. **AJ**, ML, IO, and SP prepared the tissue samples. **AJ** and ML processed the tissue samples. **AJ**, LEW, and JG conducted the RNA-seq experiments. **AJ** performed differential gene expression analysis, functional annotation analysis, network analysis, and motif identification. LEW performed power analysis. ZK provided assistance at all steps of analysis. BV and IH provided computational and statistical support. WE, IH, and DA provided laboratory equipment. **AJ**, IO, and WE wrote the manuscript. All authors read and approved the manuscript.

## Availability

Unpublished manuscript. Currently only available within this work.



## Investigating loss-of-function and human evolution of FOXP2 using global transcriptomic analysis

Aleksandar Janjic<sup>1,2</sup>, Lucas E. Wange<sup>1</sup>, Johanna Geuder<sup>1</sup>, Zane Kliesmete<sup>1</sup>, Michael Lacy<sup>3,4</sup>, Sara Pagella<sup>2,5</sup>, Isabella Ogusuku<sup>1</sup>, Beate Vieth<sup>1</sup>, Dorothee Atzler<sup>3,4</sup>, Ines Hellmann<sup>1</sup>, Wolfgang Enard<sup>1+</sup>

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

<sup>2</sup> Graduate School of Systemic Neurosciences, Department of Biology II, Ludwig-Maximilians University, Martinsried, Germany

<sup>3</sup> Institute for Cardiovascular Prevention, Ludwig-Maximilians University, Munich, Germany

<sup>4</sup> German Center for Cardiovascular Research (DZHK), Partner Site Munich Heart Alliance, Munich, Germany.

<sup>5</sup> Division of Neurobiology, Department of Biology II, Ludwig-Maximilians University, Martinsried, Germany

<sup>+</sup> Corresponding author, Lead contact:

Wolfgang Enard

Anthropology and Human Genomics, Department of Biology II

Ludwig-Maximilians University

Großhaderner Str. 2, 82152 Martinsried, Germany

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

E-Mail: enard@bio.lmu.de

## Abstract

*FOXP2* was the first gene implicated in human language and still remains one of the few speech-associated genes. Although this has led to over two decades of investigation, the exact mechanisms are still unknown. *FOXP2* research has primarily focused on the lungs, due to its role in lung development, and the brain as it is involved in cortico-basal ganglia circuitry. However, a comprehensive multi-organ transcriptomic analysis is still lacking. To this end, we utilize a cost-efficient bulk RNA-sequencing method to process 421 samples consisting of 18 tissues. In this study we aim to elucidate the functional role of *FOXP2* and the evolutionary aspect of the human-specific mutations possibly contributing to the evolution of human speech. Therefore, we compare wild-type mice to heterozygous mice with one non-functional *Foxp2* allele (knockout) and homozygous mice with variant *Foxp2* alleles containing two human-specific mutations (humanized). We perform power analysis to determine the effect size of *Foxp2* and observe that it is too small to readily interpret differences between the genotypes in most tissues. Even with such a small effect size, we detect the strongest difference in the lungs of knockout mice, which supports previous findings. Additionally, within the humanized-wild-type comparison for the brain we observe differences related to neuron projection localization and regulation.

## Introduction

Human language is one, if not the most, defining characteristics of our species. Finding a genetic explanation behind this unique evolutionary ability has been a long-term goal not only of the scientific community but a question that has been posited by people over the ages. Forkhead box protein P2 (FOXP2) is 715 amino acid transcription factor encoded by the *FOXP2* gene and expressed in various tissues, including brain, lung, thyroid, bladder, and muscle (Campbell et al. 2009; Ferland et al. 2003; Lai et al. 2003; Takahashi et al. 2003; Lai et al. 2001; Shu et al. 2005, 2007). FOXP2 is a member of the FOX superfamily of transcription factors, and more specifically the FOXP subfamily, which is characterized by various functions but is especially important in immune function, differentiation, and development (Kim et al. 2019; Lam et al. 2013).

Although a functional role of *FOXP2* has been implicated in lung (Shu et al. 2007) and brain development (Spiteri et al. 2007; Ferland et al. 2003), the gene is most interestingly associated with human speech. Specifically, missense mutations (R553H and R328X) in one allele of *FOXP2* results in developmental verbal dyspraxia (Lai et al. 2001; MacDermot et al. 2005; Vargha-Khadem et al. 2005), whereas two non-functional copies are fatal (French et al. 2007; Fujita et al. 2008; Groszer et al. 2008; Shu et al. 2005). *FOXP2* is hypothesized to affect speech by modulating cortico-basal ganglia circuits (Enard et al. 2009). Thus it potentially affects learning by accelerating transitions from declarative to procedural behavior, and thereby allowing humans to develop proper speech (Schreiweis et al. 2014).

As with other genetic mutations, an evolutionary approach can provide invaluable insight into the mechanistic explanation responsible for the phenotypic state observed. Therefore, *FOXP2* has naturally been extensively studied in model organisms, such as mice, as well as non-human primates. The chimpanzee, gorilla, and rhesus FOXP2 proteins were found to be identical to one another, and, in addition to the mouse *Foxp2* protein, differ from humans by two amino acids (exon 7, position 303 and 325) (Fig. 1A - C) (Enard et al. 2002). Even though FOXP2 is a highly conserved protein (Enard et al. 2002; Teramitsu et al. 2004; Webb and Zhang 2005), the two amino acid differences in the human lineage, which occurred after separation of human and chimpanzee, are of special interest as they have been linked to speech and learning (Enard et al. 2009; Zhang, Webb, and Podlaha 2002). Rigorously studying these amino acid

substitutions by genetic manipulation in a controlled experiment, however, would be impossible in humans and be prohibitively difficult in chimpanzees or other primates. Fortunately, since mouse *Foxp2* can be treated as the ancestral version of the human gene, genetic manipulation of *FOXP2* can be studied using a mouse model (Enard et al. 2009).

To further elucidate the role of *FOXP2*, both the effect of the entire gene as in the case of patients with developmental language impairment as well as the human-specific amino acid substitutions potentially responsible for human speech, we used previously generated *Foxp2*<sup>wt/ko</sup> (knockout) and *Foxp2*<sup>hum/hum</sup> (humanized) mouse lines and compared them against wild-type mice, from the same C57BL/6J background (Enard et al. 2009). The knockout mice were bred heterozygously as a complete knockout is lethal. These mice were extensively studied as outlined in Enard et al. (2009); although they were healthy and displayed no large physical differences, the humanized mice exhibited differences to the wild-type mice in behavior (e.g. ultrasonic vocalizations and decreased exploratory behavior) and brain phenotype (e.g. decreased dopamine concentrations and increased dendrite length in medium spiny neurons), and the knockout mice showed the opposite effects.

Despite the *Foxp2* knockout and humanized mice being subjected to extensive physical and behavioral studies, comprehensive gene expression analysis across many tissues has yet to be performed, and could be of notable insight. RNA-sequencing (RNA-seq) is the current standard for quantifying the entire transcriptome, however, such wide-scale global analysis is only recently a practical possibility due to the advent of exceptionally low-cost bulk RNA-seq protocols (Y. Li et al. 2020; Alpern et al. 2019; Janjic et al. 2021); thus allowing us to go beyond just cortico-basal ganglia circuitry, as was the focus of previous *FOXP2* studies (Enard et al. 2009; Reimers-Kipping et al. 2011; Murugan et al. 2013).

Therefore, here we sampled eighteen tissues from three different *Foxp2* genotypes (wild-type, knockout, and humanized), for a total of 421 samples (Fig. 1D and E). We performed RNA-seq on all samples simultaneously and conducted quality checks, including power analysis to determine the effect size of *Foxp2* in our knockout and humanized mice, differential gene expression analysis, functional annotation enrichment, gene regulatory network inference, and transcription factor binding motif identification (Fig. 1F). We detected the strongest signal between wild-type and knockout in the brain, bladder, liver, and lung while between wild-type and humanized it was in the brain, bladder, liver, and kidney. Additionally, a *Foxp2* binding motif was detected among the differentially expressed

genes in the lungs of knockout compared to wild-type mice. Although the scale of this study is already larger than any *Foxp2* transcriptomic analysis to date, we determine that the signal of *Foxp2* is, as previously hypothesized (Enard et al. 2009; Schreiweis et al. 2014), very small. Ultimately, this study serves as a blueprint for other genes of interest, allowing the scientific community to adapt the study to gain greater insight into relevant, yet poorly understood genes.

## Results

### Experimental design and sample quality analysis

With the advent of highly cost-efficient bulk RNA-seq protocols, we were able to look beyond the tissues primarily studied in a *FOXP2* context (i.e. lung and brain) and perform a comprehensive transcriptional analysis of *Foxp2* wild-type, knockout, and humanized mice (Fig. 1C). Therefore, we harvested tissue from 24 mice ( $n = 8$  per genotype), including males and females all of a similar age (3-4 months) (Supp. Table 1). We included the pons, medulla, motor cortex, ventral striatum, dorsal striatum, amygdala, thalamus, midbrain, and cerebellum, as well as the heart, lung, spleen, small intestines, colon, liver, kidney, bladder, and testis in our analysis (Fig. 1E).

Of the 421 samples, 410 samples (97%) passed our filters for sequencing quality and depth (Supp. Fig. 1A) with a median of 2,684,329 UMIs and a median of 20,020 detected genes per sample (Supp. Fig. 1B). We observed that samples cluster by tissue, supporting that harvesting and processing of different tissues was performed accurately (Supp. Fig. 2A and 2B). To test the accuracy of our harvesting and determine if there was contamination of neighboring regions within the brain, we performed bulk deconvolution based on single cell RNA-seq data (Tabula Muris Consortium et al. 2018), and found that all samples within each region were homogenous (chi-square test of homogeneity = 1) (Supp. Fig. 3).

Power analysis was performed as a final quality check across all tissue types for each genotype. At a sample size of 8, as was present in our design, we reach 80% conditional power for most tissues with a mean expression percentile of 50% (i.e. the top half most highly expressed genes) (Supp. Fig. 4A), and with an approximate fold change  $> 1.8$  between the genotypes (Fig. 2A). However, what we observe in our data is that most

genes have a fold change between 1 and 1.2 (Fig. 2A and Supp. Fig. 4B), which means we are underpowered for genes with low fold changes. In the lung and brain, tissues which have been previously implicated with *FOXP2*, we observe more genes with larger fold changes than the average across all tissues (Fig. 2A). Additionally, with the exception of spleen, the tissues show fairly uniform marginal power at a sample size of 8, with most reaching 65-75% power (Fig. 2B). Doubling the sample size from 8 to 16, however, would increase marginal power above 80% for all tissues except spleen (Supp. Fig. 4C).

*Foxp2* deletion most strongly impacts gene expression in lungs, potentially targeting genes regulating Notch signaling

We first investigated the effect of *FOXP2* by using our *Foxp2*<sup>wt/ko</sup> (knock-out) mice. In this case, we are able to study the effect of one non-functional copy of *Foxp2* on the transcriptome of 18 mouse tissues. We performed pairwise comparisons of the different genotypes for each individual tissue, as well as larger groupings of tissue based on hierarchical clustering, taking into account genotype, sex, batch, or tissue type based on our reduced model testing (Supp. Fig. 2B, Supp. Table 2). In many tissues or tissue groupings we saw no or few ( $\leq 2$ ) differentially expressed genes (DEG) (Fig. 3A and Supp. Table 2). However, we found significant differential expression (false discovery rate (FDR) adj. p-value < 0.05) between knockout and wild-type mice in the lungs (DEG = 303), liver (DEG = 12), bladder (DEG = 8), and brain (DEG = 12) (Fig. 3A, Supp. Table 2 - 6).

Not only did we observe the strongest difference in the lungs, but it was also the only tissue where *Foxp2* was detected as a differentially expressed gene ( $\log_2$  fold change of -0.50 wild-type compared to knockout, FDR adj. p-value = 0.03). Additionally, we observe *Foxp2* expression in the lung samples of knockout mice to be higher than that of wild-type mice (p-value = 0.004, Wilcoxon test), a pattern that is not found in any of the other tissues (Supp. Fig. 5).

Within the brain, another region that is heavily implicated with *Foxp2* function, we find that individual brain regions exhibit very few ( $\leq 2$ ) or no significant DEGs. However, when we group the samples into similar tissue types based on hierarchical clustering of the samples (e.g. entire brain, amygdala-motor cortex, ventral-dorsal striatum, pons-medulla, thalamus-midbrain, colon-small intestines) we detect more DEGs (Fig. 3A). Across the entire brain (n = 72 samples per genotype), we detect 12 DEGs (FDR adj. p-value < 0.05) when comparing knockout and wild-type mice, factoring tissue into our model (Supp. Table 2 and 4).

To determine if there were any multi-tissue effects caused by *Foxp2* deletion we looked at which DEGs were detected multiple times across all tissues. We only found three DEGs: *Rpl18a* and *Rpl13*, both involved in translation, as well as *Tmem98*, a gene involved in oligodendrocyte myelination and differentiation (Huang et al. 2018).

To better understand the differences we detect between the *Foxp2* knockout and wild-type mice, we performed functional annotation enrichment, including gene ontology (GO) and reactome analysis. Terms relating to translation and cardiac development are enriched for the DEG in the bladder, metabolic processes in the liver, and cell migration, metabolism, and learning in the lungs (Supp. Fig. 6 and Fig. 3B). Across the entire brain, we find GO terms relating to protein regulation, regulation of cell projections, and regulation of myelination/oligodendrocytes when comparing wild type and knockout samples (Supp. Fig. 2A). Reactome analysis, however, only detected pathways corresponding to ribosomal proteins, translation, or cell stress, likely due to the limited number of pathways and genes annotated compared to GO.

Differential gene expression analysis, and subsequent functional annotation enrichment, considers all genes independently, assumes they are regulated independently, and often requires a large signal when in-group sample numbers are low. We therefore performed weighted correlation network analysis (WGCNA) on the grouped brain tissues, as in this case we would have enough samples to potentially identify clusters or highly correlated genes, which can be used to identify relevant networks. These clusters (or modules) can also be related to an external trait using eigen-gene network methodology, such as in our case the *Foxp2* genotype. Although network analysis did identify 21 modules, none of these modules were significantly correlated to the *Foxp2* genotype (Supp. Fig. 8).

Finally, as *Foxp2* is a transcription factor, we wanted to identify if the binding motif is over-represented in any of our tissues where differential expression was detected. We identified *Foxp2* in the lung DEG list, which consisted of 303 genes. Of these DEGs, 18 were enriched for the *Foxp2* binding motif (scertf\_badis.HCM1) (Fig. 4) (Badis et al. 2008), including *Zeb1*, *Mecom*, *Maml3*, *Atx1*, and *Foxp2* itself. *Zeb1* is involved in neuronal differentiation and acts with *Ctbp1* (Siles et al. 2013), a protein known to interact with *Foxp2* (S. Li, Weidenfeld, and Morrissey 2004). *Mecom* is a transcription factor present in developing mouse lungs (Hawkins et al. 2017). Additionally, *Maml3* and *Atxn1* are known to regulate Notch signaling (Tong et al. 2011; Oyama et al. 2011), which has been shown to be es-

essential for proper lung development (Guseh et al. 2009; Tsao et al. 2009; Rock et al. 2011).

Human-specific amino acid substitutions in *Foxp2* effect genes primarily involved in neuron outgrowth, signaling, and maturation

Unlike our knockout mice, the *Foxp2*<sup>hum/hum</sup> mice have two functional copies of *Foxp2* but differ from wild-type mice in that they possess human-specific amino acid substitutions. Here, we wanted to study the effect of these amino acid substitutions, which could have been important during human evolution. As with the knockout mice, we investigated the difference between humanized and wild-type mice in each tissue individually, as well as the grouped tissues (Supp. Fig. 3B, Supp. Table 2 for model information). We also saw no or few ( $\leq 2$ ) DEGs in most tissues as well as most groupings, except for bladder (DEG = 31), liver (DEG = 18), kidney (DEG = 7), and brain (DEG = 25) (FDR adj. p-value < 0.05) (Fig. 3A and Supp. Table 2, 7-10). Within the brain DEGs, *Tmem106b* (Stagi et al. 2014), *Hspb1* (Ackerley et al. 2006), *Xlr3b* (Cubelos et al. 2010) have been shown to regulate neurites and *Manf* has been shown to selectively promote the survival of dopaminergic neurons of the ventral midbrain and modulate GABAergic transmission to the dopaminergic neurons of the substantia nigra (Petrova et al. 2003). Across all tissues, only *Hspa1a* and *Hspa8* were significantly differentially expressed, both of which are implicated in a variety of functions, including protein folding, stress response, and interaction with *Stub1*, which is involved with *Foxp3* degradation (Mayer 2013; Stricher et al. 2013; Ono 2020).

We then performed gene ontology and reactome analysis between humanized and wild-type mice to draw functional conclusions among the differentially expressed genes we detect. Specifically, we found terms corresponding to protein regulation and transportation in the bladder, glycosylation and RNA regulation in the liver, and cell communication and signaling, as well as neuron maturation and axon guidance in the kidney (Supp. Fig. 6). Within the combined brain sample analysis, we find GO terms relating to protein regulation and localization and regulation of neuron projections (Fig. 2C).

Reactome analysis, network analysis, and transcription factor binding motif identification between humanized and wild-type mice was inconclusive. As with the knockout mice, reactome analysis detected pathways corresponding to ribosomal proteins, translation, or cell stress, with the exception of bladder which showed enrichment in a kainate receptor pathway. WGCNA showed no significant modules explained by the geno-

type effect (Supp. Fig. 7). And motif identification found no enriched motif for DEGs detected in humanized-wild-type comparisons.

## Discussion

In this study we aimed to identify both the transcriptional effect of *FOXP2* by using the *Foxp2*<sup>wt/ko</sup> (knockout) mouse model, as well as the evolutionary effect of two human-specific *Foxp2* amino acid substitutions by using the *Foxp2*<sup>hum/hum</sup> (humanized) model. We harvested 18 tissues from 24 mice, which resulted in 421 samples, making it the largest *FOXP2* transcriptional study to date. We then used prime-seq to prepare RNA-seq libraries and various computational packages to carry out down-stream analysis (see methods).

Differential gene expression analysis confirms previous findings, while providing novel insights

As with most bulk RNA-sequencing studies, a main objective was to determine differential expression globally (i.e. across many major organ systems) in our knockout and humanized mice when compared to wild-type mice. These atlas-style studies are currently a major component of single-cell genomics. However, with the advent of affordable, high-throughput bulk RNA-seq protocols, we are able to map expression across not only the entire mouse, but across numerous mice, something that would be prohibitively expensive with a single-cell resolution. This has the added benefit that we can investigate tissues previously not extensively studied without substantially increasing overall costs.

The most apparent difference we observed was in the lung samples between knockout and wild-type mice, which supports previous findings as *FOXP2* is extensively implicated in lung development. We only differentially detected *Foxp2* in the lungs of knockout mice compared to wild-type mice. More interestingly, however, the expression pattern of *Foxp2* in the lungs is counter-intuitive to what one would expect as we detected higher *Foxp2* expression in the knockout mice than in the humanized and wild-type mice. This could potentially be explained because our knockout mouse model is heterozygous and has one functional allele present, which may act as a self-regulator and thereby result in increased *Foxp2* expression. However, as we only observe this pattern in the lungs, there must also be a lung-specific regulation resulting in this observation. To

better understand this observation, one would have to investigate expression and resulting changes either at earlier time points (i.e. prior to death caused by homozygous deletion of *Foxp2*) or by using an inducible knockout.

The exact mechanisms of *FOXP2* in lung development have yet to be determined, but what is known is that decreased *Foxp2* in mouse studies results in increased *Pax2*, *Pax8*, *Pax9*, *Hoxa9-13*, and *Pdpn* expression and decreased *Nkx2-1*, *Sox2*, *Sox9*, and *Scgb1a1* expression (S. Li et al. 2016; Shu et al. 2007; Yang et al. 2010). None of these genes were found to be differentially expressed in our study, possibly as the aforementioned research was performed using embryonic or early postnatal mice (E10.5 to P14) and our work used developed adult mice. However, at less conservative p-values *Sox5* (FDR adj. p-value = 0.06), an interactor of *Sox6/9* (Lefebvre, Li, and de Crombrughe 1998), and *Sox7* (FDR adj. p-value = 0.07), a regulator of Pax genes and Wnt signaling (Takash et al. 2001), were detected. Additionally, a number of genes related to Notch signaling were identified, including *Notch2*, *Maml3*, *Atxn1*, and *Psenen* (FDR adj. p-value = 0.03, 0.02, 0.04, and 0.05 respectively). This is particularly striking as Notch signaling has been shown to be necessary for proper lung development (Tsao et al. 2016; Kong et al. 2004). The effect of *FOXP2* on Notch signaling has been confirmed in a neuronal (Sin, Li, and Crawford 2015), bone/cartilage (Xu et al. 2018), and cell culture (SH-SY5Y and HEK293T) (Vernes et al. 2007) context. However, in *Foxp1/4* mutants, Notch signaling in the lungs was unaffected (S. Li et al. 2012); thus it may be that the regulation of Notch signaling during lung development is specific to *Foxp2*, rather than *Foxp1/4*.

While we observed the strongest effect in the lungs, we also found considerable differences in the brain when comparing the knockout and humanized mice to wild-type mice. The individual brain regions exhibited few to no DEGs; however, when all brain regions were taken together and the tissue effect included in the analysis design model (Supp. Table 2), considerably more DEGs were detected in both the knockout-wild-type and humanized-wild-type comparison.

Knockout models can have wide-spread effects on the transcriptome with many genes being dysregulated; nevertheless, we only observe 12 significant DEGs among our knockout-wild-type brain sample comparison. Our knockout model (*Foxp2*<sup>wt/ko</sup>) is heterozygous, which could possibly explain the low number of detected DEGs. Additionally, *Foxp2* homo- and hetero-dimerization is required for proper function, and as *Foxp1/3/4* are not effected sufficient functional protein could likely be present for prop-

er brain function and development. Despite the low number of DEGs, we detected genes relating to myelination (*Tmem98*) as well as cell projections (*Hspb1*) (Supp. Fig. 5A), which supports previous studies (Oswald et al. 2017).

When comparing brain samples between the humanized and wild-type mice, we observe more than twice as many DEGs as in the knockout comparison, possibly due to both alleles being altered resulting in a stronger effect (*Foxp2*<sup>hum/hum</sup>). Previous studies have speculated that *Foxp2*, and more specifically the human-specific amino acid substitutions, is involved in dopaminergic signaling of medium spiny neurons in the striatum, and thereby affects learning and speech (Enard et al. 2009). One potential mechanism is by neurite and projection regulation (Vernes et al. 2011), which is supported by the DEGs we detect (*Tmem106b*, *Hspb1*, *Xlr3b*, *Hspa5*, *Adamts1*). Additionally, *Manf* was among the detected DEGs and promotes the survival of dopaminergic neurons and modulates their transmission to the substantia nigra, which receives the afferent connections from medium spiny neurons in the striatum (Petrova et al. 2003). Interestingly, genes related to neuron maturation, axonal guidance, and cell communication (*Ntn4*, *Lrrk2*, *Lgr4*, *Rhog*, *Nfib*, *Cxcl12*, and *Ank3*) are also detected in the kidney samples in the humanized-wild-type comparison, although at a lower p-value cutoff (FDR adj.  $p < 0.1$ ) (Supp. Fig. 6B). No studies have specifically investigated the role of *Foxp2* in the kidneys and *Foxp2* is not highly expressed in the tissue (Uhlén et al. 2015). However, a possible explanation for these specific genes is that they have diverse functions including neuron maturation as well as kidney development and growth, and some have even been shown to be regulated by *Foxp2* (Moralli et al. 2015; Oswald et al. 2017; Hickey, Berto, and Konopka 2019).

Power analysis should be utilized to understand the signal effect size, and shape follow up studies

Genomic studies have for a long time been characterized by a low sample number, even when the studied genotypic effect exerts a weak phenotypic effect. This has been the case for *Foxp2*, which has primarily been studied by qPCR and microarray based transcriptomics investigating a handful of genes using few biological and technical replicates. A primary issue with such analysis is that the power to detect differences is then significantly reduced. With our own study, although the main objective may have been to define the DEGs of various tissues, likely the result that will be the most relevant and beneficial to future studies and scientists is the power analysis performed on our dataset, which provides substantial insight into the *Foxp2* effect in various tissues.

Although a biological sample size of 8 is already rather high compared to many genomic studies, we find that it is insufficient to elucidate the effect of *Foxp2* in most tissues. Specifically, we further highlight this to be the case when investigating the brain. Initially, analysis was performed on all individual brain regions separately ( $n = 8$ ), and provided almost no DEGs. However, when brain regions are combined ( $n = 72$ ) and the various tissues are modeled into the analysis, we detect DEGs for both knockout-wild-type and humanized-wild-type comparisons. This is likely attributed to the increased sample number and therefore higher power to detect differential expression. Thus a main finding of our study is that we observe the previously hypothesized small effect of *Foxp2*, especially in the brain. Additionally, as both male and female mice were used, this could potentially affect the power. Both sexes were used as there was no reason to suspect *Foxp2* effects to be sex specific. However, this added variability, even though integrated into the model, may still further reduce the power to detect differential expression.

The experimental design serves as a blueprint for investigating additional genes of interests

Our particular study was motivated by our interests in understanding the function and the evolutionary effect of *Foxp2*. Although rather specific the study itself serves as an ideal blueprint for many other investigators, as the aim can be altered but the experiments can be carried out as they are. This enables other genes to be investigated in a similar manner and may provide invaluable insight into the mechanisms governing the effect, as well as their location. Thus, although not a methodological study in principle, we wanted to discuss both the limitations and advantages of such an experimental design so that they may be more easily utilized in future studies.

A clear limitation, as discussed above, is the power required to detect differential expression, which ultimately is the main question to be addressed. In our context, we primarily discuss this in terms of sample size, as this is one of the variables where the experimenter has the most control (i.e. more samples results in more power). However, the other aspect of power is the effect size itself, which for some gene-tissue combinations may be very strong, while for others the signal may be very weak. The signal can also be manipulated based on when the tissue is sampled. In the context of *Foxp2*, for example, this would mean sampling at different time points of development, sampling during a specific activity such as behavioural learning (Schreiweis et al. 2014) or cocaine administration (Medvedeva 2015), or using an inducible construct to activate the knock-

out or humanized *Foxp2* effect at a specific point in time. These changes to experimental design, however, are very gene specific and thus a study such as the one carried out within this work can be especially beneficial in providing an initial foundation.

In addition to studying the global genotypic effect across the entire organism, the study design benefits from other advantages. Firstly, by using the entire mouse we can study the total genotypic effect along its entire lifespan, something that would be impossible with in vitro experiments or difficult with model organisms with longer gestation periods. Thus even if the signal is active at a specific point in time, if it is strong enough it will have altered the tissue in a way that this can be detected even when sampling a fully developed mouse, as is the case in our knockout-wild-type comparison of lung tissue. Secondly, a very detailed step-by-step protocol for carrying out such an experiment has been published, and the code to carry out the necessary analysis is all publicly available. This allows researchers with various levels of experience, either in molecular or computational biology, to carry out such a study. And finally, the study itself is affordable as it utilizes prime-seq, a protocol specifically developed with affordability in mind. For example, it costs approximately \$9,500 to process, make libraries, and sequence (10 million reads per sample) all 421 samples. A dataset with similar resolution using qPCR, for example, would cost roughly 491 times as much as our study (\$0.56 per reaction) assuming that there would even be enough RNA to perform the number of required reactions. Thus, realistically it is impossible to carry out a global, transcriptomic study without using an affordable RNA-seq protocol.

## Declarations

### Availability of data and materials

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

### Authors' contributions

AJ and WE conceived the study. AJ, ML, IO, and SP prepared the tissue samples. AJ and ML processed the tissue samples. AJ, LEW, and JG con-

## Results

ducted the RNA-seq experiments. AJ performed differential gene expression analysis, functional annotation analysis, network analysis, and motif identification. LEW performed power analysis. ZK provided assistance at all steps of analysis. BV and IH provided computational and statistical support. WE, IH, and DA provided laboratory equipment. AJ, IO, and WE wrote the manuscript. All authors read and approved the manuscript.

## Acknowledgements

We would like to thank Sabrina Schenk, Irena Stähler, Petra Haussner, and the entire LMU animal facility for their diligence and dedication in tending to the mice. We would like to thank LAFUGA, specifically Dr. Stefan Krebs, for some of the sequencing services provided. We would also like to thank Dr. Ozgun Gokce and Dr. Carsten Wotjak for project feedback. Finally, we would like to thank Karin Bauer, Ming Zhao, and Ines Bliessener for lab maintenance and assistance.

## Competing interests

The authors declare that they have no competing interests.

## Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the LMU Excellence Initiative and DFG EN 1093/2-1.

## Methods

### Mouse maintenance and tissue harvesting

Male and female *Foxp2*<sup>wt/wt</sup>, *Foxp2*<sup>wt/ko</sup>, and *Foxp2*<sup>hum/hum</sup> mice were housed and bred according to institutional guidelines at the Biocenter Animal Facility (Ludwig-Maximilian University of Munich, Germany). The tissue samples were collected in accordance with institutional ethical standards and the German Animal Welfare Act Paragraph 4 (organ removal for scientific reasons). For harvesting, the animals were divided into batches balanced for sex and genotype. The tissues were collected from each group separately. Mice were individually anesthetized by administration of isoflurane in an enclosed chamber. After confirming the anesthesia by checking for a negative pinch response, animals were weighed and de-

capitated. The brain was immediately removed and placed in a cooled and oxygenated artificial cerebrospinal fluid (partially frozen slush). Tissue was collected from the motor cortex, midbrain, cerebellum, dorsal striatum, ventral striatum, amygdala, thalamus, pons, and medulla. Simultaneously, the body was moved to an individual dissection tray, where samples were collected from the lung, heart, liver, bladder, small intestines, colon, spleen, kidney, and testicles (for male individuals). Tail ends were also collected from all animals and the genotypes were verified by PCR (Enard et al. 2009). Each tissue was placed in individual 2 mL tubes and stored at -80°C until use.

### Sample preparation

Once all samples were harvested, 500  $\mu$ L of lysis buffer (Qiagen RLT+, 1%  $\beta$ -mercaptoethanol) was added to the 2 mL tube and the samples were homogenized using a TissueLyser LT (Qiagen) (brain tissue: 30 Hz 2 x 1 min, body tissue: 50 Hz 2 x 1 min). Samples were visually inspected to verify homogenization and then stored at -80°C. Using the weight and an approximate RNA content of various tissues (Walker et al. 2016; “RNA Yields from Tissues and Cells” n.d.), the samples were diluted to 25 ng RNA / $\mu$ L. The exact RNA content was verified by qPCR. SYBRGreen Nucleic Acid Gel Stain (1:5000 dilution, Thermo Fisher) was added to the pre-amplification step as outlined below and quantified using a QuantStudio5 Real-Time PCR Instrument (Applied Biosystems). Approximate weight of tissue in 100  $\mu$ L of lysis buffer for proper normalization of samples was as follows: 3 mg for all brain tissues, 2 mg for heart and lung, 1 mg for spleen, small intestines, colon, liver, kidney, bladder, and testes. However, cycle verification with qPCR is ideal and should be performed in cases when sample input is as variable as a multi-organ study.

### RNA-sequencing

A step-by-step protocol of the RNA-sequencing is available on protocols.io (Janjic et al. 2018). Additionally, the method was carried out as outlined in Janjic, Wange, et al. (2021). The prime-seq protocol is briefly outlined below, including any changes specific to this work.

The lysate (50  $\mu$ L) was treated with proteinase K (Thermo Fisher) and 25 mM EDTA (Thermo Fisher) and the nucleic acids were then concentrated using a custom made bead mixture (SpeedBeads, Sigma-Aldrich). Remaining DNA was digested using DNaseI (Thermo Fisher) and the samples were once again cleaned and concentrated using the custom bead mixture. The RNA was reverse transcribed using Maxima H- enzyme (Thermo

Fisher), with a custom template-switching oligo (IDT) and custom barcoded oligo(dT) primers (IDT). Following cDNA synthesis, the samples were pooled, cleaned, and concentrated with the custom beads and residual primers were digested using Exonuclease I (Thermo Fisher). The samples were cleaned with the custom beads and eluted in ddH<sub>2</sub>O. Second strand synthesis and pre-amplification were performed in a 50  $\mu$ L reaction, consisting of 1x KAPA HiFi Ready Mix (Roche) and SingV7 primer (IDT). The final DNA product was cleaned using cleanup beads and eluted with ddH<sub>2</sub>O. The quantity and quality was assessed using a Quant-iT PicoGreen dsDNA assay kit (Thermo Fisher) and a 2100 Bioanalyzer (Agilent), respectively. The cDNA was stored at -20 °C until needed for library construction.

Libraries were then constructed using the NEBNext Ultra II FS Library Preparation Kit (NEB) with the prime-seq specific protocol rather than the manufacturer instructions. Fragmentation was performed on 2.5  $\mu$ L of cDNA (generally 2 - 20 ng). Ligation was performed with a custom prime-seq adapter (IDT). The samples were then double-size selected using SPRI-select Beads (Beckman Coulter) (0.5,0.7 cutoffs). The samples were then amplified using Q5 Master Mix (NEB), i7 Index primer (Sigma-Aldrich), and i5 Index primer (IDT). Double-size selection was performed once more as before and the quantity and quality of the libraries were assessed as with the cDNA. Paired-end (150 bp) sequencing was performed on an Illumina NovaSeq (NovoGene).

### Data Processing

The fastq files were trimmed to remove any bases from the poly(A) tail using Cutadapt (v1.12, Martin 2011). The quality was then assessed using fastqc (v0.11.8, Andrews n.d.) and then the zUMIs pipeline (v2.9.4d, Parekh et al. 2018) was used to process the data and generate a count matrix. Reads with a Phred quality score threshold of 20 for 3 BC bases and 4 UMI bases were filtered, mapped to the mouse genome (GRCm38) with the Gencode annotation (vM25) using STAR (v2.7.3a, Dobin et al. 2013), and then counted using RSubread (v1.32.4, Liao, Smyth, and Shi 2019). The count matrix was loaded into RStudio (v1.4.1717, Team 2020) using R (v4.1.0, R Core Team 2016) for further analysis.

### Sample Quality Analysis

Data processing, visualization, and descriptive statistics were done using bioMart (v2.48.2, Steffen Durinck, Wolfgang Huber 2017), dplyr (v1.0.7, Wickham et al. 2021), and tidyr (v1.1.3, Wickham and Henry 2020), ggplot2

(v3.3.5, Wickham 2010), cowplot (v1.1.1, Wilke 2019), ggrepel (version 0.9.1, Slowikowski 2018), and pheatmap (v1.0.12, Kolde 2012). Sample characterization was performed using SingleR (v1.6.1, Aran et al. 2019). Deconvolution of bulk samples was performed using SCDC (v0.0.0.9000, Dong et al. 2020) with the Tabula Muris dataset (Tabula Muris Consortium et al. 2018) as a reference. The data was filtered based on the samples as well as the genes. Sample based filtering consisted of removing samples which did not cluster with samples of the same tissue type, resulting in 11 out of 421 samples being excluded. Gene based filtering consisted of removing lowly expressed genes; specifically, genes which were present in only fewer than 25% of the samples from the same tissue type (less than 6 out of 24 samples) were excluded from further analysis.

### Power Analysis

Power analysis was performed using powsimR (v1.2.3, Vieth et al. 2017). RNA-seq data per method was simulated based on the mean variance relationship per tissue and genotype. 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', RNAseq = 'bulk', Protocol = 'Read/UMI', Distribution = 'NB', ngenes = 20000, nsims = 20, p.DE = 0.1. Log<sub>2</sub> fold changes were sampled from a gamma distribution (shape = 1, scale = 0.5) that was used for power analysis of similar data in the past (Janjic et al. 2021).

### Differential Gene Expression Analysis

Differential gene expression analysis was performed using DESeq2 (v3.0.0, Love, Huber, and Anders 2014) and limma (v3.48.1, Ritchie et al. 2015). Batch and sex information for the samples was included in the model. In cases where multiple tissue types were analyzed simultaneously (i.e. entire brain), then tissue type was also integrated into the model. In comparisons where the sample number was low (i.e. n=8, individual tissue) DESeq2 was used, as it performs more robustly with a low sample number. In comparisons where the sample number was high (i.e. combined tissues) limma:voom was used, as one is able to account for repeated measures (e.g. sampling the ventral and dorsal striatum from one mouse). Functional annotation analysis was performed using topGO (v2.44.0, Alexa and Rahnenfuhrer 2021) with the following parameters: algorithm = "elim", statistic = "Fisher", and reactome (v1.36.0, Yu and He 2016).

### Regulatory Network Analysis and Transcription Factor Motif Identification

Weighted correlation network analysis (WGCNA, v1.70-3, Langfelder and Horvath 2008) was used for finding clusters (modules) of highly correlated genes. Modules were then related to the *Foxp2* genotype using eigen-gene network methodology. RcisTarget (v1.12.0, Aibar et al. 2017) was used for identifying *Foxp2* motifs from gene lists (i.e. DEGs from individual tissue comparisons).

## References

Ackerley, Steven, Paul A. James, Arran Kalli, Sarah French, Kay E. Davies, and Kevin Talbot. 2006. "A Mutation in the Small Heat-Shock Protein HSPB1 Leading to Distal Hereditary Motor Neuronopathy Disrupts Neurofilament Assembly and the Axonal Transport of Specific Cellular Cargoes." *Human Molecular Genetics* 15 (2): 347–54.

Aibar, Sara, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, et al. 2017. "SCENIC: Single-Cell Regulatory Network Inference and Clustering." *Nature Methods* 14 (11): 1083–86.

Alexa, A., and J. Rahnenfuhrer. 2021. topGO: Enrichment Analysis for Gene Ontology.

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 Biology* 20 (1): 71.

Andrews, Simon. n.d. FastQC: A Quality Control Analysis Tool for High Throughput Sequencing Data. Github. Accessed September 14, 2021. <https://github.com/s-andrews/FastQC>.

Aran, Dvir, Agnieszka P. Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, et al. 2019. "Reference-Based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage." *Nature Immunology* 20 (2): 163–72.

Badis, Gwenaël, Esther T. Chan, Harm van Bakel, Lourdes Pena-Castillo, Desiree Tillo, Kyle Tsui, Clayton D. Carlson, et al. 2008. "A Library of Yeast Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting Nucleosome Exclusion at Promoters." *Molecular Cell* 32 (6): 878–87.

Campbell, Polly, Roger L. Reep, Margaret L. Stoll, Alexander G. Ophir, and Steven M. Phelps. 2009. "Conservation and Diversity of *Foxp2* Expression in Muroid Rodents: Functional Implications." *The Journal of Comparative Neurology* 512 (1): 84–100.

Cubelos, Beatriz, Alvaro Sebastián-Serrano, Leonardo Beccari, Maria Elisa Calcagnotto, Elsa Cisneros, Seonhee Kim, Ana Dopazo, et al. 2010. "Cux1 and Cux2 Regulate Dendritic Branching, Spine Morphology, and Synapses of the Upper Layer Neurons of the Cortex." *Neuron* 66 (4): 523–35.

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.

Dong, Meichen, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M. Perou, Fei Zou, and Yuchao Jiang. 2020. "SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References." *Briefings in Bioinformatics* 22 (1): 416–27.

Enard, Wolfgang, Sabine Gehre, Kurt Hammerschmidt, Sabine M. Hölter, Torsten Blass, Mehmet Somel, Martina K. Brückner, et al. 2009. "A Humanized Version of Foxp2 Affects Cortico-Basal Ganglia Circuits in Mice." *Cell* 137 (5): 961–71.

Enard, Wolfgang, Molly Przeworski, Simon E. Fisher, Cecilia S. L. Lai, Victor Wiebe, Takashi Kitano, Anthony P. Monaco, and Svante Pääbo. 2002. "Molecular Evolution of FOXP2, a Gene Involved in Speech and Language." *Nature* 418 (6900): 869–72.

Ferland, Russell J., Timothy J. Cherry, Patricia O. Preware, Edward E. Morrissey, and Christopher A. Walsh. 2003. "Characterization of Foxp2 and Foxp1 mRNA and Protein in the Developing and Mature Brain." *The Journal of Comparative Neurology* 460 (2): 266–79.

Fisher, Simon E. 2019. "Human Genetics: The Evolving Story of FOXP2." *Current Biology: CB*.

French, Catherine A., Matthias Groszer, Christopher Preece, Anne-Marie Coupe, Klaus Rajewsky, and Simon E. Fisher. 2007. "Generation of Mice with a Conditional Foxp2 Null Allele." *Genesis* 45 (7): 440–46.

Fujita, Eriko, Yuko Tanabe, Akira Shiota, Masatsugu Ueda, Kiyotaka Suwa, Mariko Y. Momoi, and Takashi Momoi. 2008. "Ultrasonic Vocalization Impairment of Foxp2 (R552H) Knockin Mice Related to Speech-Language Disorder and Abnormality of Purkinje Cells." *Proceedings of the National Academy of Sciences of the United States of America* 105 (8): 3117–22.

Groszer, Matthias, David A. Keays, Robert M. J. Deacon, Joseph P. de Bono, Shweta Prasad-Mulcare, Simone Gaub, Muriel G. Baum, et al. 2008. "Impaired Synaptic Plasticity and Motor Learning in Mice with a Point Mutation Implicated in Human Speech Deficits." *Current Biology: CB* 18 (5): 354–62.

Guseh, J. Sawalla, Sam A. Bores, Ben Z. Stanger, Qiao Zhou, William J. Anderson, Douglas A. Melton, and Jayaraj Rajagopal. 2009. "Notch Signaling Promotes Airway Mucous Metaplasia and Inhibits Alveolar Development." *Development* 136 (10): 1751–59.

Hawkins, Finn, Philipp Kramer, Anjali Jacob, Ian Driver, Dylan C. Thomas, Katherine B. McCauley, Nicholas Skvir, et al. 2017. "Prospective Isolation of NKX2-1-Expressing Human Lung Progenitors Derived from Pluripotent Stem Cells." *The Journal of Clinical Investigation* 127 (6): 2277–94.

Hickey, Stephanie L., Stefano Berto, and Genevieve Konopka. 2019. "Chromatin Decondensation by FOXP2 Promotes Human Neuron Maturation and Expression of Neurodevelopmental Disease Genes." *Cell Reports* 27 (6): 1699–1711.e9.

## Results

Huang, Hao, Peng Teng, Junqing Du, Jun Meng, Xuemei Hu, Tao Tang, Zunyi Zhang, Yingchuan B. Qi, and Mengsheng Qiu. 2018. "Interactive Repression of MYRF Self-Cleavage and Activity in Oligodendrocyte Differentiation by TMEM98 Protein." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 38 (46): 9829–39.

Janjic, Aleksandar, Lucas E. Wange, Johannes Bagnoli, Johanna Geuder, Phong Nguyen, Daniel Richter, Christoph Ziegenhain, and Wolfgang Enard. 2018. "Prime-Seq v1 (protocols.io.S9veh66)." *Protocols.io*. ZappyLab, Inc. <https://doi.org/10.17504/protocols.io.s9veh66>.

Janjic, Aleksandar, Lucas E. Wange, Johannes W. Bagnoli, Johanna Geuder, Phong Nguyen, Daniel Richter, Beate Vieth, et al. 2021. "Prime-Seq, Efficient and Powerful Bulk RNA-Sequencing." *bioRxiv*. <https://doi.org/10.1101/2021.09.27.459575>.

Kim, Ju-Ha, Jisung Hwang, Ji Hoon Jung, Hyo-Jung Lee, Dae Young Lee, and Sung-Hoon Kim. 2019. "Molecular Networks of FOXP Family: Dual Biologic Functions, Interplay with Other Molecules and Clinical Implications in Cancer Progression." *Molecular Cancer* 18 (1): 180.

Kolde, Raivo. 2012. *Pheatmap: Pretty Heatmaps*. <https://cran.r-project.org/web/packages/pheatmap/index.html>.

Kong, Yanping, Jonathon Glickman, Meera Subramaniam, Aliakbar Shahsafaei, K. P. Alalmneni, Jon C. Aster, Jeffrey Sklar, and Mary E. Sunday. 2004. "Functional Diversity of Notch Family Genes in Fetal Lung Development." *American Journal of Physiology. Lung Cellular and Molecular Physiology* 286 (5): L1075–83.

Lai, Cecilia S. L., Simon E. Fisher, Jane A. Hurst, Faraneh Vargha-Khadem, and Anthony P. Monaco. 2001. "A Forkhead-Domain Gene Is Mutated in a Severe Speech and Language Disorder." *Nature* 413 (6855): 519–23.

Lai, Cecilia S. L., Dianne Gerrelli, Anthony P. Monaco, Simon E. Fisher, and Andrew J. Copp. 2003. "FOXP2 Expression during Brain Development Coincides with Adult Sites of Pathology in a Severe Speech and Language Disorder." *Brain: A Journal of Neurology* 126 (11): 2455–62.

Lam, Eric W-F, Jan J. Brosens, Ana R. Gomes, and Chuay-Yeng Koo. 2013. "Forkhead Box Proteins: Tuning Forks for Transcriptional Harmony." *Nature Reviews. Cancer* 13 (7): 482–95.

Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R Package for Weighted Correlation Network Analysis." *BMC Bioinformatics* 9 (December): 559.

Lefebvre, V., P. Li, and B. de Crombrughe. 1998. "A New Long Form of Sox5 (L-Sox5), Sox6 and Sox9 Are Coexpressed in Chondrogenesis and Cooperatively Activate the Type II Collagen Gene." *The EMBO Journal* 17 (19): 5718–33.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2019. "The R Package Rsubread Is Easier, Faster, Cheaper and Better for Alignment and Quantification of RNA Sequencing Reads." *Nucleic Acids Research* 47 (8): e47.

Li, Shanru, Michael Morley, Minmin Lu, Su Zhou, Kathleen Stewart, Catherine A. French, Haley O. Tucker, Simon E. Fisher, and Edward E. Morrisey. 2016. "Foxp Transcription Factors Suppress a Non-Pulmonary Gene Expression Program to Permit Proper Lung Development." *Developmental Biology* 416 (2): 338–46.

Li, Shanru, Yi Wang, Yuzhen Zhang, Min Min Lu, Francesco J. DeMayo, Joseph D. Dekker, Philip W. Tucker, and Edward E. Morrisey. 2012. "Foxp1/4 Control Epithelial Cell Fate during Lung Development and Regeneration through Regulation of Anterior Gradient 2." *Development* 139 (14): 2500–2509.

Li, Shanru, Joel Weidenfeld, and Edward E. Morrisey. 2004. "Transcriptional and DNA Binding Activity of the Foxp1/2/4 Family Is Modulated by Heterotypic and Homotypic Protein Interactions." *Molecular and Cellular Biology* 24 (2): 809–22.

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 Biology* 21 (1): 66.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.

MacDermot, Kay D., Elena Bonora, Nuala Sykes, Anne-Marie Coupe, Cecilia S. L. Lai, Sonja C. Vernes, Faraneh Vargha-Khadem, et al. 2005. "Identification of FOXP2 Truncation as a Novel Cause of Developmental Speech and Language Deficits." *American Journal of Human Genetics* 76 (6): 1074–80.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal* 17 (1): 10–12.

Mayer, Matthias P. 2013. "Hsp70 Chaperone Dynamics and Molecular Mechanism." *Trends in Biochemical Sciences* 38 (10): 507–14.

Medvedeva, Vera. 2015. "Characterization of Foxp2 Functions in the Mouse Cortex." Université Pierre et Marie Curie - Paris VI. <https://tel.archives-ouvertes.fr/tel-01192592/document>.

Moralli, Daniela, Ron Nudel, May T. M. Chan, Catherine M. Green, Emanuela V. Volpi, Antonio Benítez-Burraco, Dianne F. Newbury, and Paloma García-Bellido. 2015. "Language Impairment in a Case of a Complex Chromosomal Rearrangement with a Breakpoint Downstream of FOXP2." *Molecular Cytogenetics* 8 (1): 1–8.

Murugan, Malavika, Stephen Harward, Constance Scharff, and Richard Mooney. 2013. "Diminished FoxP2 Levels Affect Dopaminergic Modulation of Corticostriatal Signaling Important to Song Variability." *Neuron* 80 (6): 1464–76.

Ono, Masahiro. 2020. "Control of Regulatory T-Cell Differentiation and Function by T-Cell Receptor Signalling and Foxp3 Transcription Factor Complexes." *Immunology* 160 (1): 24–37.

## Results

Oswald, Franz, Patricia Klöble, André Ruland, David Rosenkranz, Bastian Hinz, Falk Butter, Sanja Ramljak, Ulrich Zechner, and Holger Herlyn. 2017. "The FOXP2-Driven Network in Developmental Disorders and Neurodegeneration." *Frontiers in Cellular Neuroscience* 11 (July): 212.

Oyama, Toshinao, Kenichi Harigaya, Nobuo Sasaki, Yoshiaki Okamura, Hiroki Kokubo, Yumiko Saga, Katsuto Hozumi, et al. 2011. "Mastermind-like 1 (MamL1) and Mastermind-like 3 (MamL3) Are Essential for Notch Signaling in Vivo." *Development* 138 (23): 5235–46.

Parekh, Swati, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. 2018. "zUMIs - A Fast and Flexible Pipeline to Process RNA Sequencing Data with UMIs." *GigaScience* 7 (6): 1–9.

Petrova, Penka, Andrei Raibekas, Jonathan Pevsner, Noel Vigo, Mordechai Anafi, Mary K. Moore, Amy E. Peaire, et al. 2003. "MANF: A New Mesencephalic, Astrocyte-Derived Neurotrophic Factor with Selectivity for Dopaminergic Neurons." *Journal of Molecular Neuroscience: MN* 20 (2): 173–88.

R Core Team. 2016. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. <https://www.r-project.org/>.

Reimers-Kipping, S., W. Hevers, S. Pääbo, and W. Enard. 2011. "Humanized Foxp2 Specifically Affects Cortico-Basal Ganglia Circuits." *Neuroscience* 175 (February): 75–84.

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 Research* 43 (7): e47.

"RNA Yields from Tissues and Cells." n.d. ThermoFisher Scientific. Accessed October 13, 2021. <https://www.thermofisher.com/de/de/home/references/ambion-tech-support/rna-isolation/general-articles/rna-yields-from-tissues-and-cells.html>.

Rock, Jason R., Xia Gao, Yan Xue, Scott H. Randell, Young-Yun Kong, and Brigid L. M. Hogan. 2011. "Notch-Dependent Differentiation of Adult Airway Basal Stem Cells." *Cell Stem Cell* 8 (6): 639–48.

Schreiweis, Christiane, Ulrich Bornschein, Eric Burguière, Cemil Kerimoglu, Sven Schreiter, Michael Dannemann, Shubhi Goyal, et al. 2014. "Humanized Foxp2 Accelerates Learning by Enhancing Transitions from Declarative to Procedural Performance." *Proceedings of the National Academy of Sciences of the United States of America* 111 (39): 14253–58.

Shu, Weiguo, Julie Y. Cho, Yuhui Jiang, Minhua Zhang, Donald Weisz, Gregory A. Elder, James Schmeidler, et al. 2005. "Altered Ultrasonic Vocalization in Mice with a Disruption in the Foxp2 Gene." *Proceedings of the National Academy of Sciences of the United States of America* 102 (27): 9643–48.

Shu, Weiguo, Min Min Lu, Yuzhen Zhang, Philip W. Tucker, Deying Zhou, and Edward E. Morrisey. 2007. "Foxp2 and Foxp1 Cooperatively Regulate Lung and Esophagus Development." *Development* 134 (10): 1991–2000.

Siles, Laura, Ester Sánchez-Tilló, Jong-Won Lim, Douglas S. Darling, Kristen L. Kroll, and Antonio Postigo. 2013. "ZEB1 Imposes a Temporary Stage-Dependent Inhibition of Muscle Gene Expression and Differentiation via CtBP-Mediated Transcriptional Repression." *Molecular and Cellular Biology* 33 (7): 1368–82.

Sin, Cora, Hongyan Li, and Dorota A. Crawford. 2015. "Transcriptional Regulation by FOXP1, FOXP2, and FOXP4 Dimerization." *Journal of Molecular Neuroscience: MN* 55 (2): 437–48.

Slowikowski, Kamil. 2018. Ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2."

Spiteri, Elizabeth, Genevieve Konopka, Giovanni Coppola, Jamee Bomar, Michael Oldham, Jing Ou, Sonja C. Vernes, Simon E. Fisher, Bing Ren, and Daniel H. Geschwind. 2007. "Identification of the Transcriptional Targets of FOXP2, a Gene Linked to Speech and Language, in Developing Human Brain." *American Journal of Human Genetics* 81 (6): 1144–57.

Stagi, Massimiliano, Zoe A. Klein, Travis J. Gould, Joerg Bewersdorf, and Stephen M. Strittmatter. 2014. "Lysosome Size, Motility and Stress Response Regulated by Fronto-Temporal Dementia Modifier TMEM106B." *Molecular and Cellular Neurosciences* 61 (July): 226–40.

Steffen Durinck, Wolfgang Huber. 2017. *biomaRt*. Bioconductor. <https://doi.org/10.18129/B9.BIOC.BIOMART>.

Stricher, François, Christophe Macri, Marc Ruff, and Sylviane Muller. 2013. "HSPA8/HSC70 Chaperone Protein: Structure, Function, and Chemical Targeting." *Autophagy* 9 (12): 1937–54.

Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators. 2018. "Single-Cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris." *Nature* 562 (7727): 367–72.

Takahashi, Kaoru, Fu-Chin Liu, Katsuiku Hirokawa, and Hiroshi Takahashi. 2003. "Expression of Foxp2, a Gene Involved in Speech and Language, in the Developing and Adult Striatum." *Journal of Neuroscience Research* 73 (1): 61–72.

Takash, W., J. Cañizares, N. Bonneaud, F. Poulat, M. G. Mattéi, P. Jay, and P. Berta. 2001. "SOX7 Transcription Factor: Sequence, Chromosomal Localisation, Expression, Transactivation and Interference with Wnt Signalling." *Nucleic Acids Research* 29 (21): 4274–83.

Team, Rstudio. 2020. "RStudio: Integrated Development for R. RStudio, PBC, Boston, MA, 2020."

Teramitsu, Ikuko, Lili C. Kudo, Sarah E. London, Daniel H. Geschwind, and Stephanie A. White. 2004. "Parallel FoxP1 and FoxP2 Expression in Songbird and Human Brain Predicts Functional Interaction." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 24 (13): 3152–63.

## Results

Tong, Xin, Hongxing Gui, Feng Jin, Bryan W. Heck, Peihui Lin, Jianjie Ma, Joseph D. Fondell, and Chih-Cheng Tsai. 2011. "Ataxin-1 and Brother of Ataxin-1 Are Components of the Notch Signalling Pathway." *EMBO Reports* 12 (5): 428–35.

Tsao, Po-Nien, Chisa Matsuoka, Shu-Chen Wei, Atsuyasu Sato, Susumu Sato, Koichi Hasegawa, Hung-Kuan Chen, et al. 2016. "Epithelial Notch Signaling Regulates Lung Alveolar Morphogenesis and Airway Epithelial Integrity." *Proceedings of the National Academy of Sciences of the United States of America* 113 (29): 8242–47.

Tsao, Po-Nien, Michelle Vasconcelos, Konstantin I. Izvolsky, Jun Qian, Jining Lu, and Wellington V. Cardoso. 2009. "Notch Signaling Controls the Balance of Ciliated and Secretory Cell Fates in Developing Airways." *Development* 136 (13): 2297–2307.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. "Proteomics. Tissue-Based Map of the Human Proteome." *Science* 347 (6220): 1260419.

Vargha-Khadem, Faraneh, David G. Gadian, Andrew Copp, and Mortimer Mishkin. 2005. "FOXP2 and the Neuroanatomy of Speech and Language." *Nature Reviews. Neuroscience* 6 (2): 131–38.

Vernes, Sonja C., Peter L. Oliver, Elizabeth Spiteri, Helen E. Lockstone, Rathu Puliyadi, Jennifer M. Taylor, Joses Ho, et al. 2011. "Foxp2 Regulates Gene Networks Implicated in Neurite Outgrowth in the Developing Brain." *PLoS Genetics* 7 (7): e1002145.

Vernes, Sonja C., Elizabeth Spiteri, Jérôme Nicod, Matthias Groszer, Jennifer M. Taylor, Kay E. Davies, Daniel H. Geschwind, and Simon E. Fisher. 2007. "High-Throughput Analysis of Promoter Occupancy Reveals Direct Neural Targets of FOXP2, a Gene Mutated in Speech and Language Disorders." *American Journal of Human Genetics* 81 (6): 1232–50.

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): 3486–88.

Walker, Douglas G., Alexis M. Whetzel, Geidy Serrano, Lucia I. Sue, Lih-Fen Lue, and Thomas G. Beach. 2016. "Characterization of RNA Isolated from Eighteen Different Human Tissues: Results from a Rapid Human Autopsy Program." *Cell and Tissue Banking* 17 (3): 361–75.

Webb, D. M., and J. Zhang. 2005. "FoxP2 in Song-Learning Birds and Vocal-Learning Mammals." *The Journal of Heredity* 96 (3): 212–16.

Wickham, Hadley. 2010. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation* (version v1.0.7). <https://github.com/tidyverse/dplyr>.

Wickham, Hadley, and Lionel Henry. 2020. "Tidyr: Tidy Messy Data." *R Package Version 1* (2): 397.

Wilke, Claus O. 2019. Cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2.”

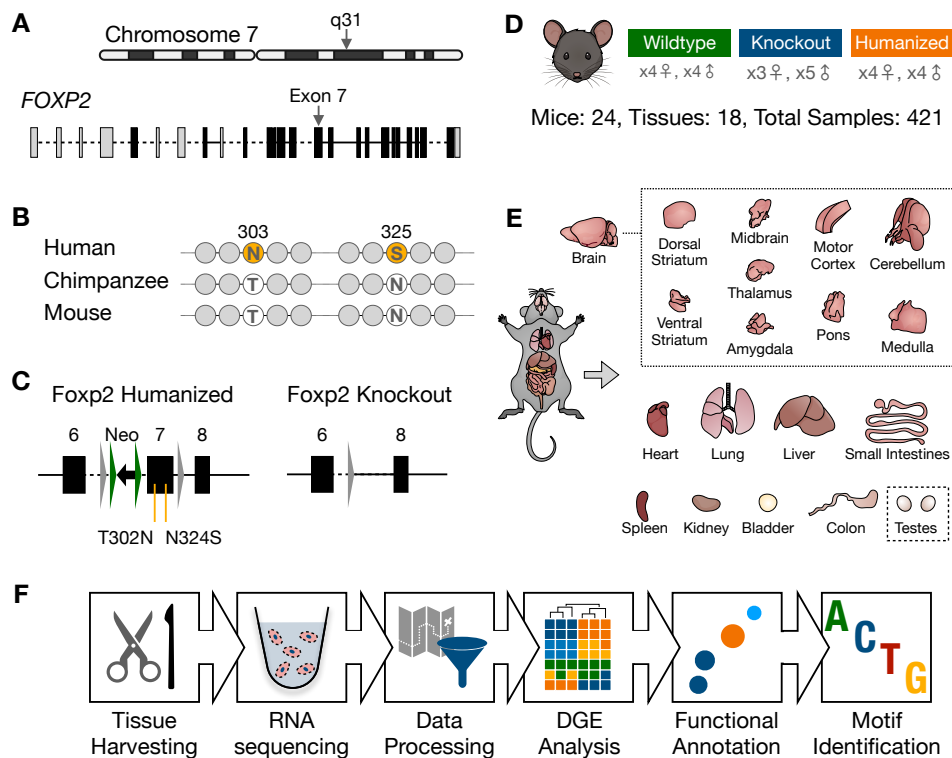
Xu, Shuqin, Pei Liu, Yuanxing Chen, Yi Chen, Wei Zhang, Haixia Zhao, Yiwei Cao, et al. 2018. “Foxp2 Regulates Anatomical Features That May Be Relevant for Vocal Behaviors and Bipedal Locomotion.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (35): 8799–8804.

Yang, Zhi, Keisuke Hikosaka, Mohammad T. K. Sharkar, Tomoki Tamakoshi, Abhishek Chandra, Bo Wang, Tatsuo Itakura, et al. 2010. “The Mouse Forkhead Gene Foxp2 Modulates Expression of the Lung Genes.” *Life Sciences* 87 (1-2): 17–25.

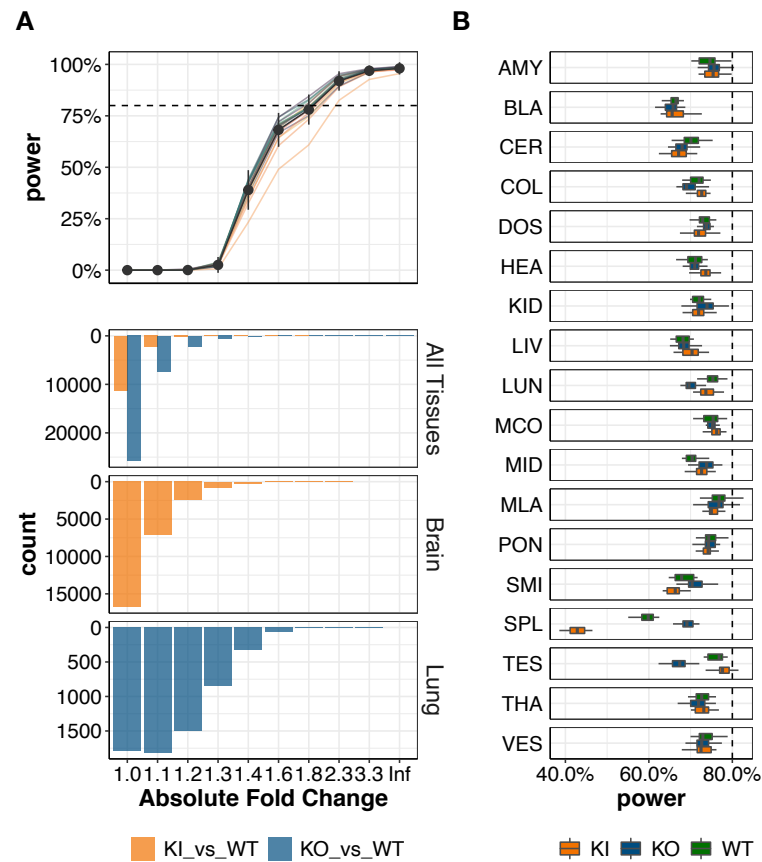
Yu, Guangchuang, and Qing-Yu He. 2016. “ReactomePA: An R/Bioconductor Package for Reactome Pathway Analysis and Visualization.” *Molecular bioSystems* 12 (2): 477–79.

Zhang, Jianzhi, David M. Webb, and Ondrej Podlaha. 2002. “Accelerated Protein Evolution and Origins of Human-Specific Features: Foxp2 as an Example.” *Genetics* 162 (4): 1825–35.

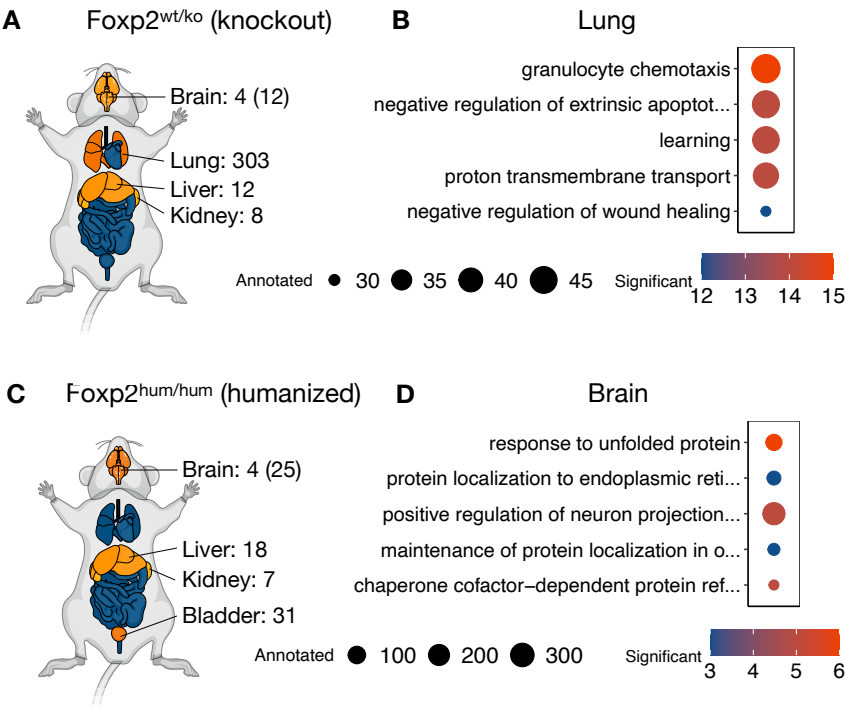
## Figures



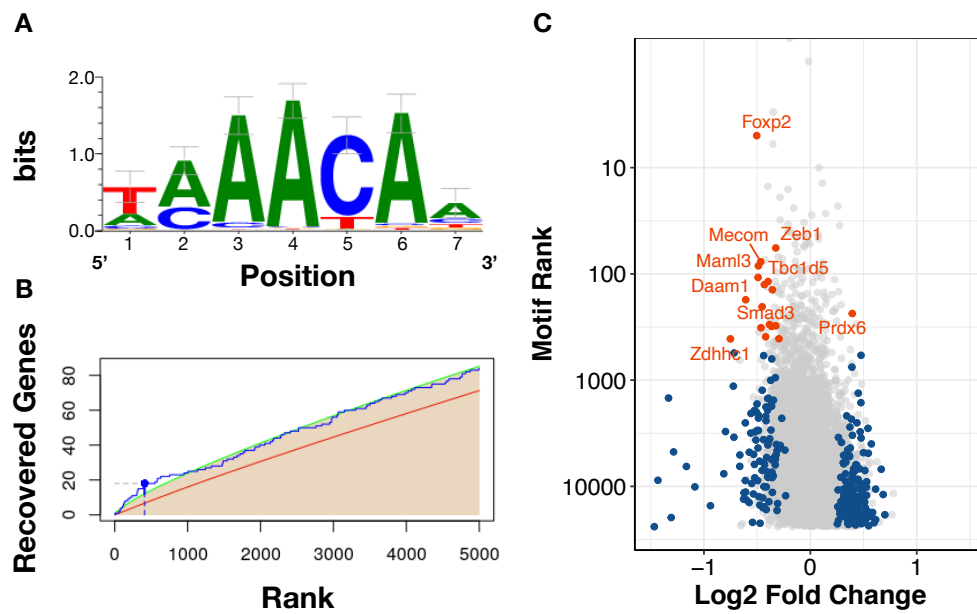
**Figure 1. *FOXP2* overview and experimental design.** (A) *FOXP2* is located on chromosome 7q31. The exons are represented with boxes (black are protein coding) and the introns with a line (adapted from Fisher 2019). (B) *FOXP2* has two human specific amino acid substitutions at position 303 and 325. (C) *Foxp2* humanized (*Foxp2*<sup>hum/hum</sup>) and knockout (*Foxp2*<sup>wt/ko</sup>) mouse models, along with wild-type mice were used in (D) the experiment to generate a total of 421 samples. (E) Eighteen tissues were harvested, including nine from the brain, along with heart, lung, liver, small intestines, spleen, kidney, bladder, colon, and testes from male mice. (F) Following tissue harvesting, RNA sequencing was performed on all samples simultaneously and then the data was processed and analyzed, including differential gene expression analysis, functional annotation, and motif identification.



**Figure 2. Power to detect differentially expressed genes is overall relatively similar, but higher in certain tissues.** (A) Across all tissues, an average of 80% power is reached at an absolute log fold change of 1.8. When investigating differences between the genotypes with all samples, we observe low fold changes among a majority of the genes. However, within certain tissues the fold changes are higher. In the humanized-wild-type comparison in the brain and the knockout-wild-type comparison in the lung, for example, we have higher fold changes and thus more power to detect differentially expressed genes. (B) With our experimental design (i.e.  $n = 8$  per genotype, per tissue) we generally do not reach a marginal power of 80% in any tissue; however, with the exception of the spleen, we observe similar power in most tissues.

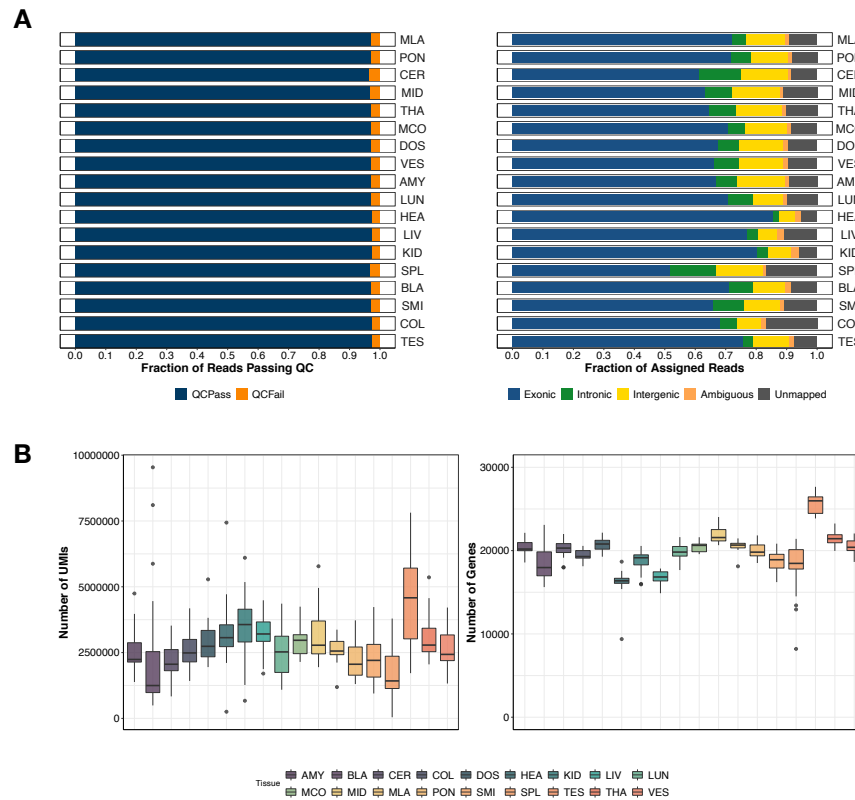


**Figure 3. Differentially expressed genes in knockout- and humanized-wild-type comparisons.** (A) In knockout-wild-type comparisons we detected 303 differentially expressed genes in the lung. Among all brain samples, we detected only 4 differentially expressed genes (thalamus = 2, motor cortex = 1, and medulla = 1). However when analyzing all brain samples together and factoring the tissue into the model, we detected 12 differentially expressed genes. (B) Gene ontology of the lung tissue identified terms relating to cell migration, metabolism, and learning. (C) In humanized-wild-type comparisons we detected 31 differentially expressed genes in the bladder and 25 when using all brain tissue samples (a total of 4 in individual tissue comparisons of the brain, thalamus = 2, motor cortex = 1, and ventral striatum = 1). (D) Terms relating to regulation of proteins and neuron projections were detected in the brain samples.

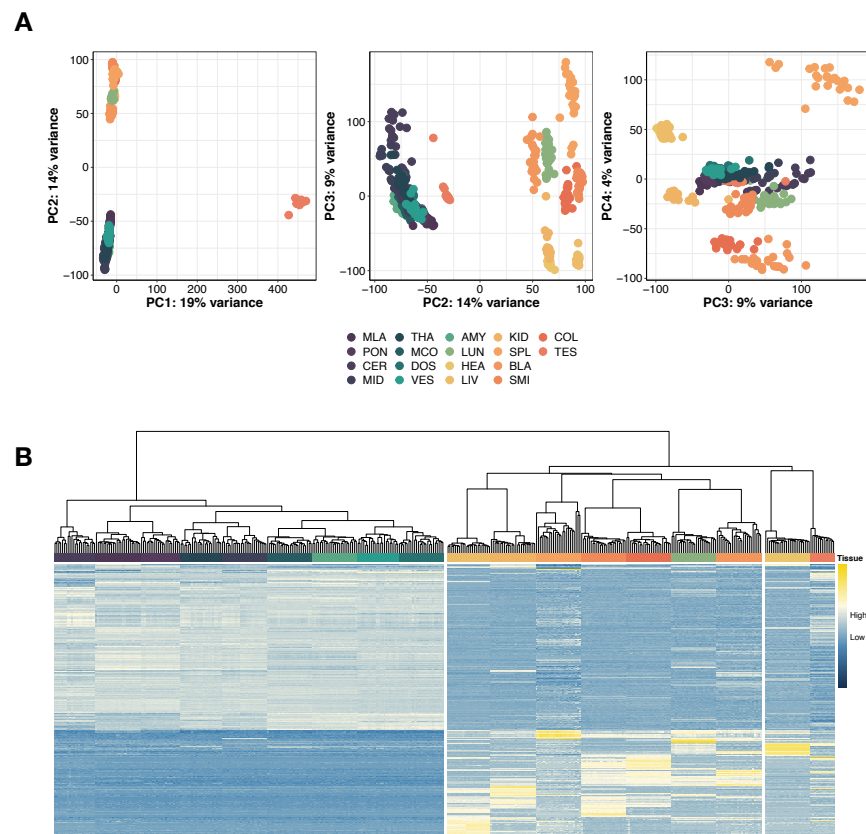


**Figure 4. Motif identification for *FOXP2* finds target genes in lung samples.** (A) The sequence of the enriched motif identified among the differentially expressed genes in the knockout-wild-type comparison, which matches the *FOXP2* motif. (B) Gene ranking of all genes within the identified motif determines 18 genes with high scores. (C) Volcano plot of all genes in the knockout-wild-type lung comparison, with blue genes representing differentially expressed genes (adj. P value > 0.05) and orange genes representing differentially expressed genes enriched in motif identification.

## Supplementary Figures

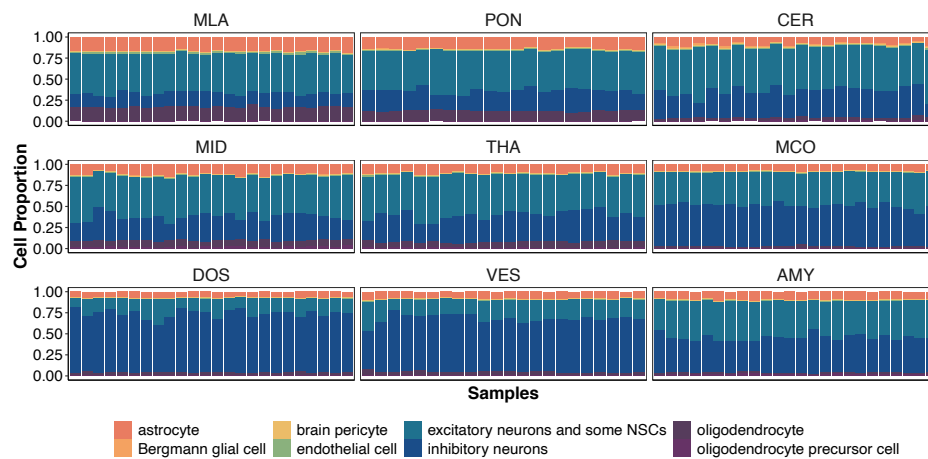


**Supplementary Figure 1. Sequencing quality and sensitivity of libraries.** (A) Sequencing quality was very high with more than 95% of reads passing quality filtering and similar between tissues (Phred quality score threshold of 20 for 3 BC bases and 4 UMI bases were filtered). Additionally, feature distribution between tissues was fairly similar with the exception of spleen samples which had a higher fraction of unmapped reads. (B) An approximate average of 2.5 million UMIs and 20,000 genes were detected among all samples.

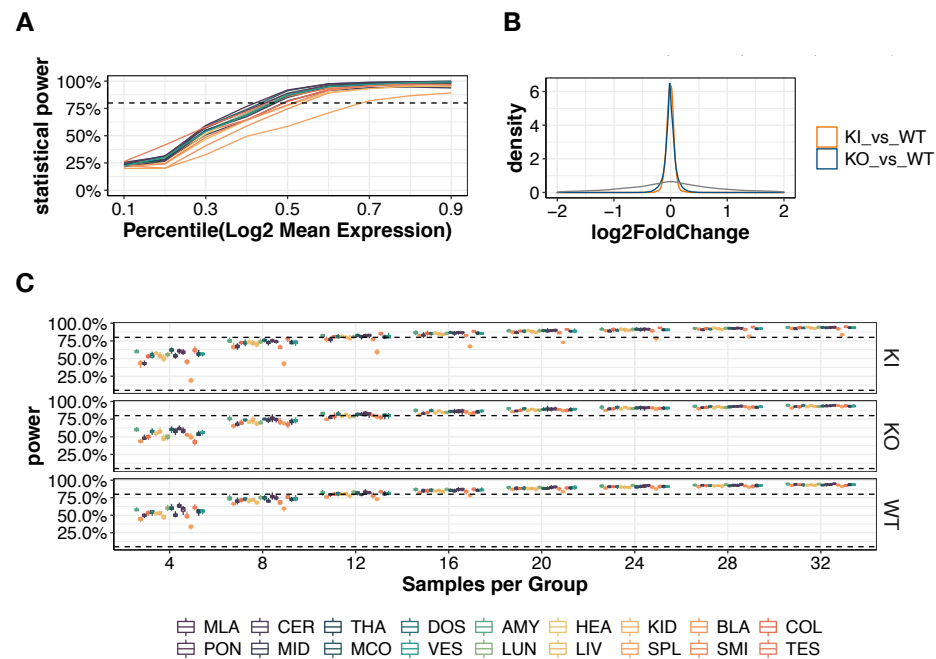


**Supplementary Figure 2. Clustering of samples.** (A) Principal component analysis, showing PC1-PC2, PC2-PC3, and PC3-PC4. The largest variance is explained by tissues, with PC1 showing the separation of testes from all other samples and PC2 the brain samples from the rest of the body. (B) Hierarchical clustering of samples divides the samples by brain and all other samples.

## Results

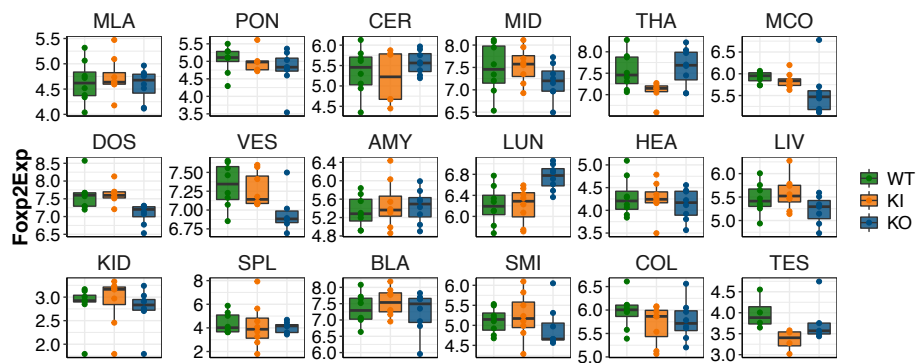


**Supplementary Figure 3. Deconvolution of brain samples.** Cell proportions were determined using deconvolution of the bulk RNA-seq libraries for each brain tissue with a single-cell RNA-seq reference from the Tabula Muris. Although some samples vary from one another, the samples are homogenous (chi-square test of homogeneity = 1).



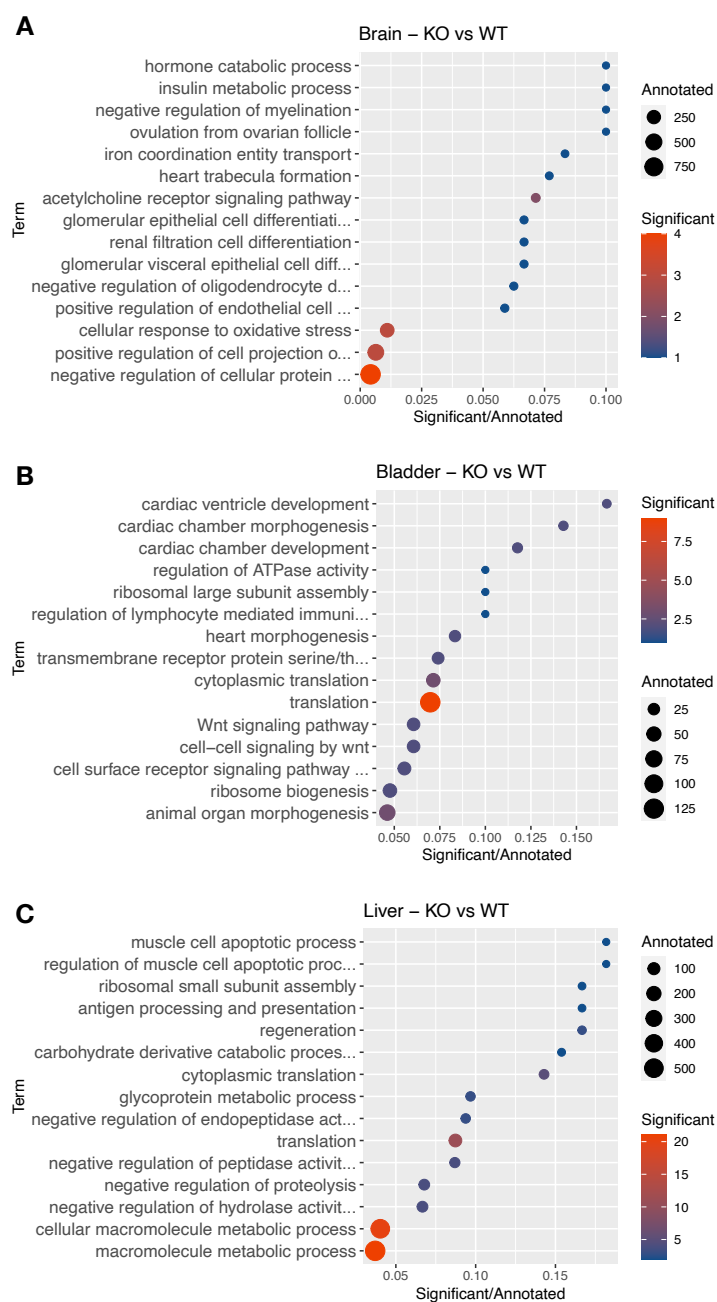
**Supplementary Figure 4. Sufficient power is reached with stronger expression or more samples.** (A) Marginal power of each tissue at different expression levels shows insufficient power to detect differential expression among lowly expressed genes. (B) Density plot of observed  $\log_2$  fold change in the knockout (blue line) and humanized (orange line) comparisons to wild-type mice, compared to  $\log_2$  fold changes observed in previously published experiments utilizing prime-seq (gray line). (C) Observed power in each tissue for each genotype at various numbers of samples shows most tissues pass 80% power at 16 samples.

## Results

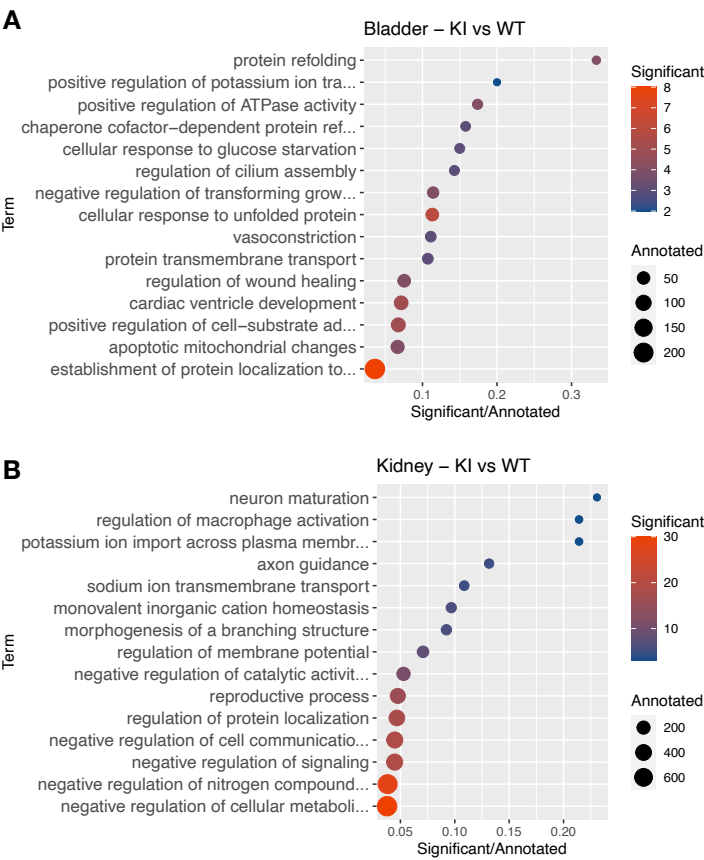


**Supplementary Figure 5.** Rlog transformed UMI-counts for *Foxp2* in various tissues, contrasting the knockout, humanized, and wild-type samples.

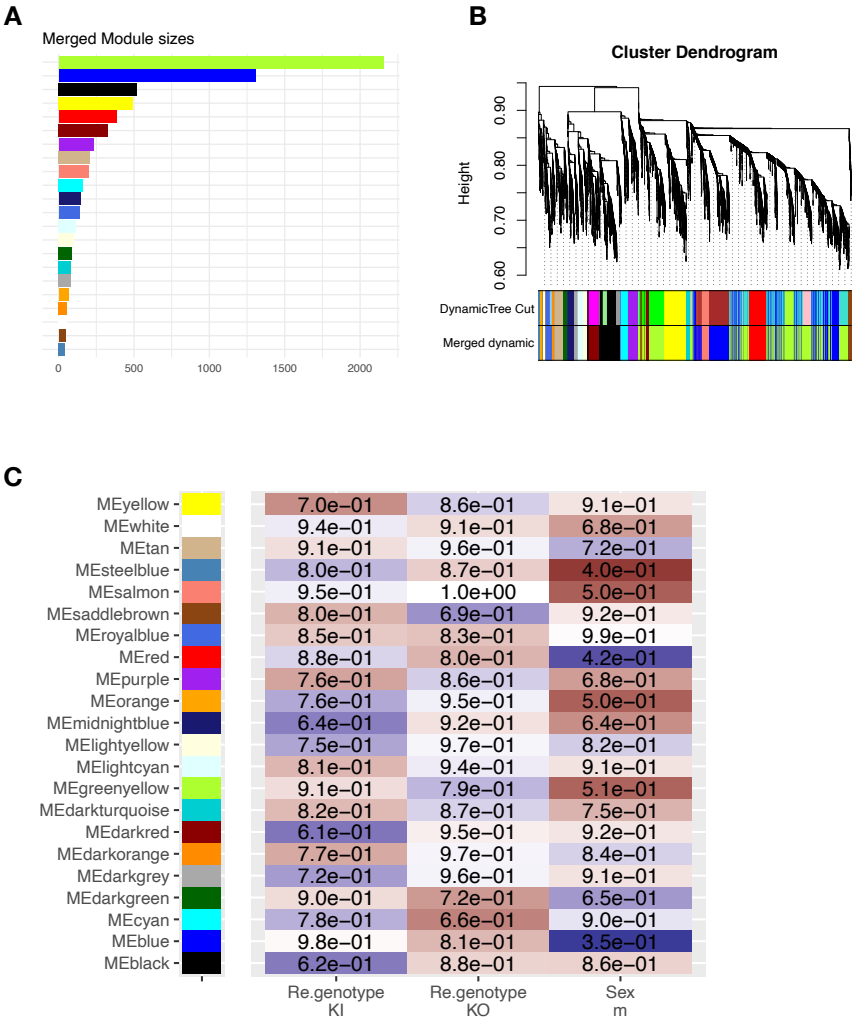
# Investigating loss-of-function and human evolution of FOXP2 using global transcriptomic analysis



**Supplementary Figure 6.** Gene ontology in knockout compared to wildtype tissues for (A) brain, (B) bladder, and (C) liver.



**Supplementary Figure 7.** Gene ontology in humanized compared to wildtype tissues for (A) bladder and (B) kidney.



**Supplementary Figure 8. WGCNA Analysis.** (A) Identified and (B) merged modules in network analysis from all brain samples. (C) All identified modules were not explained by the genotypic difference (knockout-wild-type and humanized-wild-type comparison) or sex of mouse.

Supplementary Table 1

Sample	Sex	Cage-ID	Animal	Strain	DoB	Batch	Order	Foxp2	Re-genotype	Snca	Weight
1 f	6086	1492	FOXP2_ko_5H11	1/8/2020	1	1.1	wt/ko	wt/ko	wt/ko	wt/ko	23
2 m	6085	1488	FOXP2_ko_5H11	1/8/2020	1	1.2	wt/ko	wt/ko	wt/ko	wt/ko	30.08
3 f	6088	1499	FOXP2_ki_5H11_deltaNeo	1/9/2020	1	1.3	hum/hum	hum/hum	wt/ko	wt/ko	26.9
4 f	6086	1490	FOXP2_ko_5H11	1/8/2020	1	1.4	wt/ko	wt/ko	wt/ko	wt/ko	21.82
5 m	6091	1510	FOXP2_ki_5H11_deltaNeo	1/10/2020	1	1.5	wt/ko	wt/ko	wt/ko	wt/ko	32.1
6 m	6089	1503	FOXP2_ki_5H11_deltaNeo	1/10/2020	1	1.6	hum/hum	hum/hum	wt/ko	wt/ko	34.6
7 f	6086	1493	FOXP2_ko_5H11	1/8/2020	2	2.1	wt/ko	wt/ko	wt/ko	wt/ko	22.4
8 m	6188	1519	FOXP2_ko_5H11	2/2/2020	2	2.2	wt/ko	wt/ko	wt/ko	wt/ko	30.22
9 m	6188	1521	FOXP2_ko_5H11	2/2/2020	2	2.3	wt/ko	wt/ko	wt/ko	wt/ko	30.3
10 f	6092	1507	FOXP2_ki_5H11_deltaNeo	1/10/2020	2	2.4	hum/hum	hum/hum	wt/ko	wt/ko	23.1
11 f	6092	1508	FOXP2_ki_5H11_deltaNeo	1/10/2020	2	2.5	hum/hum	hum/hum	wt/ko	wt/ko	24.13
12 m	6089	1504	FOXP2_ki_5H11_deltaNeo	1/10/2020	2	2.6	hum/hum	hum/hum	wt/ko	wt/ko	28.5
13 m	6356	1594	FOXP2_ko_5H11	2/27/2020	3	3.1	wt/ko	wt/ko	wt/ko	wt/ko	27
14 m	6341	1534	FOXP2_ko_5H11	2/25/2020	3	3.2	wt/ko	wt/ko	wt/ko	wt/ko	27.8
15 f	6092	1513	FOXP2_ki_5H11_deltaNeo	1/10/2020	3	3.3	wt/ko	wt/ko	wt/ko	wt/ko	23.85
16 m	6089	1505	FOXP2_ki_5H11_deltaNeo	1/10/2020	3	3.4	hum/hum	hum/hum	wt/ko	wt/ko	33.5
17 m	6091	1511	FOXP2_ki_5H11_deltaNeo	1/10/2020	3	3.5	wt/ko	wt/ko	wt/ko	wt/ko	28.65
18 m	6085	1489	FOXP2_ko_5H11	1/8/2020	3	3.6	wt/ko	wt/ko	wt/ko	wt/ko	30.4
19 f	6086	1491	FOXP2_ko_5H11	1/8/2020	4	4.1	wt/ko	wt/ko	wt/ko	wt/ko	24.4
20 f	6345	1551	FOXP2_ki_5H11_deltaNeo	2/25/2020	4	4.2	hum/hum	hum/hum	wt/ko	wt/ko	23.6
21 f	6424	1540	FOXP2_ko_5H11	2/25/2020	4	4.3	wt/ko	wt/ko	wt/ko	wt/ko	23.6
22 m	6298	1524	FOXP2_ki_5H11_deltaNeo	2/24/2020	4	4.4	hum/hum	hum/hum	wt/ko	wt/ko	28.03
23 m	6343	1543	FOXP2_ko_5H11	2/25/2020	4	4.5	wt/ko	wt/ko	wt/ko	wt/ko	25.53
24 f	6092	1512	FOXP2_ki_5H11_deltaNeo	1/10/2020	4	4.6	wt/ko	wt/ko	wt/ko	wt/ko	26.95

**Supplementary Table 2**

<b>Tissue</b>	<b>KI</b>	<b>KO</b>	<b>Grouping</b>	<b>Model</b>
<b>Lung</b>	0	303	Individual	Re.genotype + Sex
Brain	25	12	Group	Re.genotype + Tissue + Sex + Batch
Bladder	31	8	Individual	Re.genotype + Sex + Batch
Liver	18	12	Individual	Re.genotype + Sex + Batch
Kidney	7	0	Individual	Re.genotype + Sex + Batch
Thalamus	2	2	Individual	Re.genotype + Sex + Batch
Amygdala-Motor Cortex	0	0	Group	Re.genotype + Tissue + Sex + Batch
Motor Cortex	1	1	Individual	Re.genotype + Sex
Ventral-Dorsal Striatum	0	0	Group	Re.genotype + Tissue + Sex + Batch
Medulla	0	1	Individual	Re.genotype + Sex + Batch
Ventral Striatum	1	0	Individual	Re.genotype + Sex + Batch
Colon	0	1	Individual	Re.genotype + Sex + Batch
Colon-Small Intestines	0	1	Group	Re.genotype + Tissue + Sex + Batch
Pons	0	0	Individual	Re.genotype + Sex
Cerebellum	0	0	Individual	Re.genotype + Sex
Midbrain	0	0	Individual	Re.genotype + Sex
Dorsal Striatum	0	0	Individual	Re.genotype + Sex
Amygdala	0	0	Individual	Re.genotype + Sex + Batch
Heart	0	0	Individual	Re.genotype + Sex + Batch
Spleen	0	0	Individual	Re.genotype + Sex
Small Intestines	0	0	Individual	Re.genotype + Sex + Batch
Testes	0	0	Individual	Re.genotype
Pons-Medulla	0	0	Group	Re.genotype + Tissue + Sex + Batch
Thalamus-Midbrain	0	0	Group	Re.genotype + Tissue + Sex + Batch

## Results

### Supplementary Table 3

Lung - KO vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
Ank3	ENSMUSG00000069601	-0.6624	716.9361	0.0000	0.0172	-4.1461
Atp6v0e	ENSMUSG00000015575	0.4064	383.4040	0.0001	0.0172	4.0485
C1qc	ENSMUSG00000036896	0.5498	70.6770	0.0000	0.0172	4.0739
Ccdc162	ENSMUSG00000075225	-0.7975	77.5854	0.0000	0.0172	-4.1785
Clasp2	ENSMUSG00000033392	-0.5010	102.3633	0.0000	0.0172	-4.1785
Cracr2a	ENSMUSG00000061414	-0.6278	57.2060	0.0000	0.0172	-4.5468
Daam1	ENSMUSG00000034574	-0.6075	330.0859	0.0000	0.0172	-4.2177
Dapk1	ENSMUSG00000021559	-0.4916	292.4238	0.0000	0.0172	-4.1318
Ehbp1	ENSMUSG00000042302	-0.6011	92.7849	0.0001	0.0172	-4.0247
Gas5	ENSMUSG00000053332	0.3596	338.5875	0.0000	0.0172	4.1364
Il18r1	ENSMUSG00000026070	-0.7236	170.1691	0.0000	0.0172	-4.5123
Lars2	ENSMUSG00000035202	-1.4329	1199.7091	0.0001	0.0172	-4.0447
Mob3b	ENSMUSG00000073910	-0.5350	179.2044	0.0001	0.0172	-4.0230
NA	ENSMUSG00000076258	-1.2241	1082.5271	0.0001	0.0172	-4.0340
Naaladl2	ENSMUSG00000102758	-0.7167	52.0632	0.0000	0.0172	-4.1268
Rpl3	ENSMUSG00000060036	0.4280	967.7488	0.0000	0.0172	4.4530
Sfta3-ps	ENSMUSG00000112343	-0.4037	468.5588	0.0000	0.0172	-4.0751
Tcf20	ENSMUSG00000041852	-0.6634	165.6625	0.0000	0.0172	-4.1471
Tnik	ENSMUSG00000027692	-0.5513	88.6232	0.0001	0.0172	-4.0157
Torn1l2	ENSMUSG00000000538	-0.6079	127.3803	0.0000	0.0172	-4.0832
Wfdc1	ENSMUSG00000023336	0.6847	83.3406	0.0000	0.0172	4.2156
Ypel3	ENSMUSG00000042675	0.4691	318.0896	0.0000	0.0172	4.2921
Apoe	ENSMUSG00000002985	0.5071	1540.3955	0.0001	0.0174	3.9445
Atp8a1	ENSMUSG00000037685	-0.3968	718.5875	0.0001	0.0174	-3.9437
Camk1d	ENSMUSG00000039145	-1.5975	2353.2863	0.0001	0.0174	-3.9870
Gm15564	ENSMUSG00000086324	-1.5417	80.2453	0.0001	0.0174	-3.9382
Igfbp6	ENSMUSG00000023046	0.4410	1537.1557	0.0001	0.0174	3.9575
Map4k4	ENSMUSG00000026074	-0.3948	169.0006	0.0001	0.0174	-3.9567
Nav2	ENSMUSG00000052512	-0.4637	370.0114	0.0001	0.0174	-3.9638
Ptprj	ENSMUSG00000025314	-0.5737	196.1521	0.0001	0.0174	-3.9690
Zdhhc1	ENSMUSG00000039199	-0.7513	184.5520	0.0001	0.0175	-3.9296
Ap2m1	ENSMUSG00000022841	0.5455	224.1995	0.0001	0.0183	3.9104
Cnbp	ENSMUSG00000030057	0.3967	356.5766	0.0001	0.0184	3.8686
Eif3k	ENSMUSG00000053565	0.3436	322.5013	0.0001	0.0184	3.8678
Fth1	ENSMUSG00000024661	0.4636	6602.6905	0.0001	0.0184	3.8968
Msi2	ENSMUSG00000069769	-0.3635	344.3072	0.0001	0.0184	-3.8871
Rps4x-ps	ENSMUSG00000104699	0.4226	333.9342	0.0001	0.0184	3.8793
Unc50	ENSMUSG00000026111	0.5852	101.7412	0.0001	0.0184	3.8781
Psmb2	ENSMUSG00000028837	0.3494	227.7201	0.0001	0.0190	3.8541
Bag1	ENSMUSG00000028416	0.3662	411.5751	0.0001	0.0198	3.8314
Mir6240	ENSMUSG000000098343	-1.4666	439.6207	0.0001	0.0198	-3.8336
Lyz2	ENSMUSG00000069516	0.4581	9698.7999	0.0001	0.0205	3.7998
Nfia	ENSMUSG00000028565	-0.2945	511.8846	0.0001	0.0205	-3.8013
Rps4x	ENSMUSG00000031320	0.6138	3213.7050	0.0001	0.0205	3.8052
Selenok	ENSMUSG00000042682	0.4491	458.8064	0.0001	0.0205	3.8056
Arpc1b	ENSMUSG00000029622	0.4348	415.3112	0.0002	0.0222	3.7376
Myo1e	ENSMUSG00000032220	-0.4158	212.0609	0.0002	0.0222	-3.7386

# Investigating loss-of-function and human evolution of FOXP2 using global transcriptomic analysis

Lung - KO vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
NA	ENSMUSG00000106106	-1.2456	142300.4380	0.0002	0.0222	-3.7360
Rngtt	ENSMUSG00000028274	-0.5435	59.6186	0.0002	0.0222	-3.7650
Rps5	ENSMUSG00000012848	0.4442	3372.2264	0.0002	0.0222	3.7377
Sap18b	ENSMUSG000000061104	0.3845	315.5333	0.0002	0.0222	3.7531
Sh3bgrl3	ENSMUSG00000028843	0.4974	456.4192	0.0002	0.0222	3.7521
Tmem98	ENSMUSG00000035413	0.5734	61.4240	0.0002	0.0222	3.7435
Trib1	ENSMUSG00000032501	-0.8136	107.2716	0.0002	0.0222	-3.7349
Gm11808	ENSMUSG00000068240	0.3447	808.5280	0.0002	0.0226	3.6995
Lyve1	ENSMUSG00000030787	0.5110	341.0209	0.0002	0.0226	3.6988
Ms4a4d	ENSMUSG00000024678	0.5772	63.5036	0.0002	0.0226	3.7197
Ncoa1	ENSMUSG00000020647	-0.4773	191.6886	0.0002	0.0226	-3.7117
Rbm25	ENSMUSG00000010608	-0.3628	399.0284	0.0002	0.0226	-3.7033
Tanc2	ENSMUSG00000053580	-0.4876	206.6690	0.0002	0.0226	-3.7017
Zfpn2	ENSMUSG00000022306	-0.6112	82.6036	0.0002	0.0226	-3.7224
Enah	ENSMUSG00000022995	-0.5010	108.1903	0.0002	0.0227	-3.6938
Acad9	ENSMUSG00000027710	-1.5228	72.7332	0.0002	0.0230	-3.6824
Ppib	ENSMUSG00000032383	0.3438	488.1362	0.0002	0.0230	3.6849
Dguok	ENSMUSG00000014554	0.4882	71.4985	0.0002	0.0232	3.6764
Acot13	ENSMUSG00000006717	-1.2856	2229.2878	0.0003	0.0253	-3.6396
Atp13a3	ENSMUSG00000022533	-0.3286	216.1337	0.0003	0.0253	-3.6316
Cdk2ap2	ENSMUSG00000024856	0.4959	263.8988	0.0003	0.0253	3.6377
H2-Aa	ENSMUSG00000036594	0.3899	1095.2057	0.0003	0.0253	3.6490
Pkig	ENSMUSG00000035268	0.3097	180.3868	0.0003	0.0253	3.6320
Samd4	ENSMUSG00000021838	-0.4201	249.2895	0.0003	0.0253	-3.6379
Anxa1	ENSMUSG00000024659	0.3926	400.9595	0.0003	0.0268	3.6090
Bcap31	ENSMUSG00000002015	0.3947	342.5886	0.0003	0.0268	3.6113
Slc25a5	ENSMUSG00000016319	0.5127	462.2646	0.0003	0.0272	3.6023
Noc2l	ENSMUSG00000009567	-1.3082	4665.2477	0.0003	0.0275	-3.5954
Ccl6	ENSMUSG00000018927	0.7004	290.4406	0.0003	0.0278	3.5867
Mecom	ENSMUSG00000027684	-0.4664	491.4670	0.0003	0.0278	-3.5857
Ube2e2	ENSMUSG00000058317	-0.4909	146.9681	0.0003	0.0278	-3.5822
Cfap61	ENSMUSG00000037143	-0.6232	57.6847	0.0004	0.0280	-3.5729
H2-T23	ENSMUSG00000067212	0.5988	317.5095	0.0004	0.0280	3.5703
Ly6c1	ENSMUSG00000079018	0.5482	549.9834	0.0004	0.0280	3.5722
Krtcap2	ENSMUSG00000042747	0.3524	423.4658	0.0004	0.0282	3.5590
Pcx	ENSMUSG00000024892	-0.4875	66.9321	0.0004	0.0282	-3.5600
Txndc9	ENSMUSG00000058407	0.4364	102.5719	0.0004	0.0282	3.5608
Cnpy2	ENSMUSG00000025381	0.3846	176.6228	0.0004	0.0283	3.5532
Eif1	ENSMUSG00000035530	0.3393	1348.9105	0.0004	0.0283	3.5495
Laptn4a	ENSMUSG00000020585	0.3995	894.4383	0.0004	0.0283	3.5488
Calm2	ENSMUSG00000036438	0.3526	956.9940	0.0004	0.0284	3.5388
H2-Ab1	ENSMUSG00000073421	0.4534	1519.5689	0.0004	0.0284	3.5443
Rps6ka3	ENSMUSG00000031309	-0.4320	160.5855	0.0004	0.0284	-3.5364
Sec13	ENSMUSG00000030298	0.4507	128.2125	0.0004	0.0284	3.5390
Ndufb11	ENSMUSG00000031059	0.3204	399.3083	0.0004	0.0290	3.5247
Stox2	ENSMUSG00000038143	-0.5333	148.9467	0.0004	0.0290	-3.5272
Maml3	ENSMUSG00000061143	-0.4880	269.8201	0.0004	0.0295	-3.5149

## Results

Lung - KO vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
Ngdn	ENSMUSG00000022204	0.4770	90.6976	0.0004	0.0295	3.5148
Atp5h	ENSMUSG00000034566	0.3724	909.3391	0.0005	0.0299	3.5028
Hydin	ENSMUSG00000059854	-0.5583	84.5975	0.0005	0.0299	-3.5049
Tmed3	ENSMUSG00000032353	0.3317	245.0071	0.0005	0.0299	3.5034
Ctss	ENSMUSG00000038642	0.3956	235.8164	0.0005	0.0301	3.4939
Tm2d2	ENSMUSG00000031556	0.4595	188.1214	0.0005	0.0301	3.4938
Tspan4	ENSMUSG00000025511	0.6667	78.9732	0.0005	0.0301	3.4933
Gstp-ps	ENSMUSG000000103653	0.4421	705.0631	0.0005	0.0306	3.4860
H2bc4	ENSMUSG000000018102	0.3800	414.8133	0.0005	0.0310	3.4733
mt-Rnr1	ENSMUSG000000064337	-0.4871	22459.7887	0.0005	0.0310	-3.4719
Ppp6r2	ENSMUSG00000036561	-0.4725	63.5954	0.0005	0.0310	-3.4738
Prkn	ENSMUSG00000023826	-0.5614	97.8103	0.0005	0.0310	-3.4698
Rbx1	ENSMUSG00000022400	0.3502	484.9683	0.0005	0.0310	3.4777
Id4	ENSMUSG00000021379	-0.5114	66.8711	0.0005	0.0313	-3.4647
Kctd8	ENSMUSG00000037653	-0.7181	79.9562	0.0005	0.0314	-3.4616
Grpel1	ENSMUSG00000029198	0.5302	99.0460	0.0005	0.0315	3.4577
Rpl18a	ENSMUSG00000045128	0.2897	1532.8566	0.0006	0.0318	3.4534
Gm28438	ENSMUSG000000101939	-0.5005	1061.0990	0.0006	0.0325	-3.4450
Sod3	ENSMUSG000000072941	0.4996	572.5166	0.0006	0.0332	3.4341
Uqcrb	ENSMUSG00000021520	0.2875	446.9968	0.0006	0.0332	3.4344
Ccl21a	ENSMUSG000000094686	0.5276	295.1750	0.0007	0.0355	3.3981
Cuedc2	ENSMUSG00000036748	0.3866	178.7083	0.0007	0.0355	3.3914
Fnbp4	ENSMUSG000000008200	-0.4360	67.1089	0.0007	0.0355	-3.3961
Gdi1	ENSMUSG000000015291	0.4354	92.8743	0.0007	0.0355	3.3985
Mtx2	ENSMUSG000000027099	0.5418	62.0956	0.0007	0.0355	3.4035
Notch2	ENSMUSG000000027878	-0.4534	90.9144	0.0007	0.0355	-3.4069
Piezo2	ENSMUSG000000041482	-0.3993	156.8368	0.0007	0.0355	-3.4065
Psma1	ENSMUSG000000030751	0.4651	147.9523	0.0007	0.0355	3.3927
Rpl13	ENSMUSG00000000740	0.2543	2375.0391	0.0007	0.0355	3.3913
Rps6-ps4	ENSMUSG000000081406	0.3279	515.6587	0.0007	0.0355	3.4010
Snrbp	ENSMUSG000000027404	0.3167	464.1474	0.0007	0.0355	3.4089
NA	ENSMUSG0000000048191	-1.2629	56.6004	0.0007	0.0360	-3.3853
mt-Nd2	ENSMUSG000000064345	-0.3903	6962.8638	0.0007	0.0362	-3.3751
Ninl	ENSMUSG000000068115	-1.1648	91.2728	0.0007	0.0362	-3.3767
Selenof	ENSMUSG000000037072	0.3143	473.7159	0.0007	0.0362	3.3797
Zeb1	ENSMUSG000000024238	-0.3251	259.9838	0.0007	0.0362	-3.3767
Nmt2	ENSMUSG000000026643	-0.3630	92.2687	0.0008	0.0367	-3.3692
9030622O22Rik	ENSMUSG000000086141	-0.6649	81.4883	0.0008	0.0370	-3.3598
Bcl6	ENSMUSG000000022508	-0.4518	68.9973	0.0008	0.0370	-3.3634
Gphn	ENSMUSG000000047454	-1.0852	1528.1559	0.0008	0.0370	-3.3584
Ift57	ENSMUSG000000032965	0.4359	51.1620	0.0008	0.0370	3.3603
Ascc2	ENSMUSG000000020412	-0.4366	102.8956	0.0008	0.0373	-3.3406
Dda1	ENSMUSG000000074247	0.4154	94.0647	0.0008	0.0373	3.3405
Gm47720	ENSMUSG000000112468	-0.4981	50.9504	0.0008	0.0373	-3.3515
Gngt2	ENSMUSG000000038811	0.4123	154.6201	0.0008	0.0373	3.3427
Id3	ENSMUSG000000007872	0.4177	478.6391	0.0008	0.0373	3.3400
Ier3ip1	ENSMUSG000000090000	0.3309	200.4830	0.0008	0.0373	3.3492

# Investigating loss-of-function and human evolution of FOXP2 using global transcriptomic analysis

Lung - KO vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
Met	ENSMUSG00000009376	-0.4107	113.7850	0.0008	0.0373	-3.3463
Napsa	ENSMUSG00000002204	0.3583	689.1693	0.0008	0.0373	3.3460
Dock10	ENSMUSG000000038608	-0.4012	169.7263	0.0008	0.0373	-3.3380
Ckap4	ENSMUSG000000046841	-0.4005	97.7862	0.0009	0.0375	-3.3214
Eps8	ENSMUSG000000015766	-0.3941	97.5515	0.0009	0.0375	-3.3176
Fcf1	ENSMUSG000000021243	0.3543	101.5301	0.0009	0.0375	3.3232
H3f3b	ENSMUSG000000016559	0.3508	854.4374	0.0009	0.0375	3.3287
Mustn1	ENSMUSG000000042485	0.3789	97.4748	0.0009	0.0375	3.3207
Ndufb9	ENSMUSG000000022354	0.3330	267.8412	0.0009	0.0375	3.3256
Pbx3	ENSMUSG000000038718	-0.3839	119.2549	0.0009	0.0375	-3.3297
Ppp1r14c	ENSMUSG000000040653	-0.3425	206.1488	0.0009	0.0375	-3.3174
Snx4	ENSMUSG000000022808	-0.3125	151.4581	0.0009	0.0375	-3.3162
Sra1	ENSMUSG000000006050	0.4888	152.1371	0.0009	0.0375	3.3271
Ubb	ENSMUSG0000000019505	0.4505	3332.2574	0.0009	0.0375	3.3289
Foxp2	ENSMUSG000000029563	-0.5040	81.3699	0.0009	0.0383	-3.3084
Fkbp1a	ENSMUSG000000032966	0.4016	710.1044	0.0010	0.0389	3.3025
Dnpep	ENSMUSG000000026209	0.4254	79.0524	0.0010	0.0395	3.2965
Tjp2	ENSMUSG000000024812	-0.3117	250.1215	0.0010	0.0397	-3.2933
Gm6136	ENSMUSG0000000084106	0.2450	316.6764	0.0010	0.0401	3.2865
Nutl2-ps1	ENSMUSG000000071497	0.5453	81.5365	0.0010	0.0401	3.2869
Celf1	ENSMUSG000000005506	-0.3365	198.0165	0.0010	0.0409	-3.2794
Celf2	ENSMUSG000000002107	-0.3242	424.1604	0.0011	0.0410	-3.2739
Chka	ENSMUSG0000000024843	-0.4148	146.7547	0.0011	0.0410	-3.2698
Crip2	ENSMUSG000000006356	0.3949	1796.4586	0.0011	0.0410	3.2668
Gbf1	ENSMUSG0000000025224	-0.3714	145.6088	0.0011	0.0410	-3.2762
Hsd17b11	ENSMUSG0000000029311	0.4788	1532.7880	0.0011	0.0410	3.2729
Osbp16	ENSMUSG000000042359	-0.3637	261.9234	0.0011	0.0410	-3.2664
Pde6d	ENSMUSG0000000026239	0.4143	63.1096	0.0011	0.0410	3.2653
Prdx6	ENSMUSG0000000026701	0.3941	2208.1859	0.0011	0.0410	3.2706
Arap2	ENSMUSG0000000037999	-0.3583	158.8623	0.0012	0.0414	-3.2375
Atxn1	ENSMUSG0000000046876	-0.3565	241.3713	0.0011	0.0414	-3.2532
Cyba	ENSMUSG0000000006519	0.3530	324.2341	0.0011	0.0414	3.2571
Dnmt3b	ENSMUSG0000000027478	-1.3329	215.9276	0.0012	0.0414	-3.2354
Fam172a	ENSMUSG0000000064138	-0.4463	166.8010	0.0012	0.0414	-3.2477
Fbxo6	ENSMUSG0000000055401	0.4340	94.2086	0.0011	0.0414	3.2558
H2-D1	ENSMUSG0000000073411	0.4650	2218.1044	0.0012	0.0414	3.2423
Klf12	ENSMUSG0000000072294	-0.5601	67.1372	0.0012	0.0414	-3.2380
Lamc1	ENSMUSG0000000026478	-0.3219	154.2713	0.0012	0.0414	-3.2499
Myl12b	ENSMUSG0000000034868	0.4156	574.3002	0.0012	0.0414	3.2426
Pbx1	ENSMUSG0000000052534	-0.3625	631.9271	0.0012	0.0414	-3.2467
Psmc8	ENSMUSG0000000030591	0.3416	187.6158	0.0012	0.0414	3.2377
Runx1	ENSMUSG0000000022952	-0.4013	199.0616	0.0012	0.0414	-3.2494
Sf3b4	ENSMUSG0000000068856	0.4929	83.5663	0.0011	0.0414	3.2603
Snapin	ENSMUSG0000000001018	0.3485	104.9860	0.0012	0.0414	3.2363
Taldo1	ENSMUSG0000000025503	0.3644	325.6788	0.0011	0.0414	3.2532
Tns3	ENSMUSG0000000020422	-0.2388	362.3492	0.0012	0.0414	-3.2403
Acsf4	ENSMUSG0000000031278	-0.3464	270.6728	0.0013	0.0415	-3.2098

## Results

Lung - KO vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
Ak7	ENSMUSG00000041323	-0.5754	67.9270	0.0014	0.0415	-3.2005
Arpp19	ENSMUSG00000007656	0.3152	410.6029	0.0013	0.0415	3.2130
Cd9	ENSMUSG00000030342	0.3272	904.0043	0.0014	0.0415	3.2043
Clic1	ENSMUSG00000007041	0.3546	406.6101	0.0013	0.0415	3.2065
Cnn3	ENSMUSG00000053931	0.3717	171.8837	0.0013	0.0415	3.2195
Cope	ENSMUSG00000055681	0.3857	281.6502	0.0013	0.0415	3.2223
Creld2	ENSMUSG00000023272	0.5643	141.2916	0.0014	0.0415	3.1999
Cystm1	ENSMUSG00000046727	0.3912	549.2247	0.0012	0.0415	3.2282
Grap	ENSMUSG00000004837	0.4995	136.1207	0.0013	0.0415	3.2132
Ifitm3	ENSMUSG00000025492	0.4794	1405.5454	0.0014	0.0415	3.2023
Il31ra	ENSMUSG00000050377	-1.6045	116.5980	0.0013	0.0415	-3.2136
Mif2	ENSMUSG00000030120	0.3395	380.6566	0.0014	0.0415	3.2034
Ndufa9	ENSMUSG00000000399	0.3651	95.1971	0.0014	0.0415	3.2005
Plin2	ENSMUSG00000028494	0.5172	151.1860	0.0013	0.0415	3.2180
Plscr3	ENSMUSG00000019461	0.3316	132.9067	0.0014	0.0415	3.2039
Rarres2	ENSMUSG00000009281	0.3266	398.8039	0.0014	0.0415	3.1998
Rpl17-ps3	ENSMUSG000000113948	0.5523	970.9388	0.0013	0.0415	3.2180
Shisa5	ENSMUSG000000025647	0.4115	336.6412	0.0013	0.0415	3.2263
Slc24a3	ENSMUSG000000063873	-0.4838	73.0318	0.0014	0.0415	-3.2001
Smdt1	ENSMUSG00000022452	0.2656	611.0300	0.0013	0.0415	3.2141
Snx3	ENSMUSG00000019804	0.3948	373.2844	0.0014	0.0415	3.2031
Tmbim4	ENSMUSG00000020225	0.3342	177.2000	0.0014	0.0415	3.2010
Traf5	ENSMUSG00000026637	0.2929	249.8770	0.0014	0.0415	3.2036
Kcnq1ot1	ENSMUSG000000101609	-0.4752	198.3792	0.0014	0.0417	-3.1965
Tbc1d5	ENSMUSG00000023923	-0.3974	106.8723	0.0014	0.0417	-3.1959
Smad3	ENSMUSG00000032402	-0.4530	75.7129	0.0014	0.0425	-3.1892
Dnah9	ENSMUSG00000056752	-0.6114	68.5705	0.0014	0.0425	-3.1874
Aldoa	ENSMUSG00000030695	0.3563	912.7928	0.0015	0.0426	3.1845
Emc2	ENSMUSG00000022337	0.3499	83.3376	0.0014	0.0426	3.1846
Dnah5	ENSMUSG00000022262	-0.5086	90.4668	0.0015	0.0426	-3.1814
Snx24	ENSMUSG00000024535	-0.3566	81.0251	0.0015	0.0426	-3.1815
Ccser1	ENSMUSG000000039578	-0.5262	96.2589	0.0015	0.0426	-3.1800
Slc25a4	ENSMUSG00000031633	0.3655	468.7883	0.0015	0.0431	3.1757
Ctsd	ENSMUSG00000007891	0.3679	538.4892	0.0015	0.0440	3.1661
Myo5c	ENSMUSG00000033590	-0.3861	167.3213	0.0015	0.0440	-3.1659
Rbfa	ENSMUSG00000024570	0.4673	55.3300	0.0015	0.0440	3.1676
Vkorc1	ENSMUSG000000096145	0.3723	105.2682	0.0016	0.0441	3.1637
Irak1	ENSMUSG00000031392	-0.3893	80.4784	0.0016	0.0443	-3.1611
Gm11966	ENSMUSG000000080904	0.4064	202.6342	0.0016	0.0443	3.1576
Kmt2d	ENSMUSG00000048154	-0.4462	72.1458	0.0016	0.0443	-3.1539
Mrpl42	ENSMUSG000000062981	0.4385	140.6506	0.0016	0.0443	3.1549
Ninj1	ENSMUSG00000037966	0.4629	70.3945	0.0016	0.0443	3.1535
Slc35b1	ENSMUSG00000020873	0.3465	107.4993	0.0016	0.0443	3.1538
Tmem9b	ENSMUSG00000031021	0.4755	142.1908	0.0016	0.0443	3.1552
Hipk2	ENSMUSG000000061436	-0.3724	171.8743	0.0016	0.0443	-3.1522
Pi4k2b	ENSMUSG00000029186	-0.3078	156.2578	0.0016	0.0443	-3.1508
Dram1	ENSMUSG00000020057	-0.2321	376.1375	0.0017	0.0447	-3.1458

# Investigating loss-of-function and human evolution of FOXP2 using global transcriptomic analysis

Lung - KO vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
Fkbp2	ENSMUSG00000056629	0.3779	560.7082	0.0016	0.0447	3.1470
Mtcl1	ENSMUSG00000052105	-0.2824	117.6835	0.0017	0.0447	-3.1446
Fam193b	ENSMUSG00000021495	-0.4408	67.5404	0.0017	0.0448	-3.1427
Mib1	ENSMUSG00000024294	-0.3486	119.9445	0.0017	0.0448	-3.1417
Cita	ENSMUSG00000028478	0.2902	293.1037	0.0017	0.0454	3.1352
Rbms3	ENSMUSG00000039607	-0.3293	461.4998	0.0017	0.0454	-3.1352
Rps6-ps3	ENSMUSG00000082465	0.3231	1038.3828	0.0017	0.0459	3.1310
Atp6v1f	ENSMUSG00000004285	0.3382	370.8551	0.0018	0.0462	3.1277
Ctnnd2	ENSMUSG00000022240	-0.5253	72.4038	0.0018	0.0463	-3.1257
Hsbp1	ENSMUSG00000031839	0.3300	430.9725	0.0018	0.0467	3.1221
Braf	ENSMUSG00000002413	-0.3199	160.9480	0.0019	0.0472	-3.1081
Eef1d	ENSMUSG00000055762	0.2579	227.4234	0.0019	0.0472	3.1092
Ifitm2	ENSMUSG00000060591	0.5183	828.1919	0.0018	0.0472	3.1149
Map1lc3b	ENSMUSG00000031812	0.3657	323.4928	0.0019	0.0472	3.1093
Ppp1r11	ENSMUSG00000036398	0.3409	198.7234	0.0019	0.0472	3.1113
Rpl22l1	ENSMUSG00000039221	0.5288	255.8185	0.0019	0.0472	3.1133
Rplp0	ENSMUSG00000067274	0.2611	1084.1652	0.0019	0.0472	3.1099
Tbc1d22a	ENSMUSG00000051864	-0.9391	404.5362	0.0019	0.0472	-3.1094
Tcta	ENSMUSG00000039461	0.4247	60.0653	0.0019	0.0472	3.1115
Txn2	ENSMUSG00000005354	0.3123	157.3918	0.0019	0.0472	3.1071
Grip1	ENSMUSG00000034813	-0.5221	83.6498	0.0019	0.0472	-3.1059
Hint1	ENSMUSG00000020267	0.2800	458.6511	0.0019	0.0472	3.1046
Cpeb1	ENSMUSG00000025586	0.3698	97.7914	0.0019	0.0474	3.1013
Pdgfd	ENSMUSG00000032006	-0.4098	94.4343	0.0019	0.0474	-3.0999
Ppm1h	ENSMUSG00000034613	-0.4172	91.8815	0.0019	0.0474	-3.1001
Cd74	ENSMUSG00000024610	0.4562	4425.9254	0.0020	0.0475	3.0960
Gnptg	ENSMUSG00000035521	0.4107	90.3715	0.0019	0.0475	3.0980
St3gal4	ENSMUSG00000032038	-0.4307	279.4560	0.0020	0.0475	-3.0968
Kit	ENSMUSG00000005672	0.4177	111.8898	0.0020	0.0476	3.0942
Apip	ENSMUSG00000010911	0.4765	70.0419	0.0020	0.0477	3.0923
Cd63	ENSMUSG00000025351	0.3370	634.3366	0.0020	0.0477	3.0902
Psma2	ENSMUSG00000015671	0.3278	277.1633	0.0020	0.0477	3.0891
Vamp8	ENSMUSG00000050732	0.3070	1151.2613	0.0020	0.0477	3.0902
A330023F24Rik	ENSMUSG00000096929	-0.5189	62.8227	0.0021	0.0484	-3.0819
Ubr3	ENSMUSG00000044308	-0.3604	125.3579	0.0021	0.0484	-3.0814
Yif1a	ENSMUSG00000024875	0.3907	85.0364	0.0020	0.0484	3.0832
NA	ENSMUSG00000076281	-0.7577	122.2396	0.0021	0.0487	-3.0789
Ndufa6	ENSMUSG00000022450	0.2792	491.7786	0.0021	0.0489	3.0751
Sftpc	ENSMUSG00000022097	0.3921	120531.4152	0.0021	0.0489	3.0757
Txn1	ENSMUSG00000028367	0.3827	3195.3860	0.0021	0.0489	3.0745
Lriq1	ENSMUSG00000019892	-0.4953	58.6445	0.0021	0.0493	-3.0694
Spes1	ENSMUSG00000021917	0.3601	293.2779	0.0021	0.0493	3.0702
Lrg1	ENSMUSG00000037095	0.6140	242.7823	0.0022	0.0497	3.0664
Atraid	ENSMUSG00000013622	0.4147	166.5313	0.0023	0.0497	3.0543
Aurkaip1	ENSMUSG00000065990	0.3709	173.8403	0.0022	0.0497	3.0593
Cst3	ENSMUSG00000027447	0.3420	2197.7048	0.0022	0.0497	3.0560
Dennd4a	ENSMUSG00000053641	-0.4875	104.4937	0.0022	0.0497	-3.0585

## Results

Lung - KO vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
Gm10123	ENSMUSG00000062933	0.3773	313.2929	0.0022	0.0497	3.0628
Gm49207	ENSMUSG000000115497	0.4769	147.2938	0.0022	0.0497	3.0612
Rrp7a	ENSMUSG00000018040	0.4701	58.4621	0.0022	0.0497	3.0586
Syf2	ENSMUSG000000028821	0.3638	228.6364	0.0022	0.0497	3.0602
Tbrg1	ENSMUSG00000011114	0.2992	181.9947	0.0022	0.0497	3.0628
Tecr	ENSMUSG000000031708	0.3465	393.0362	0.0023	0.0497	3.0548
Tm2d3	ENSMUSG000000078681	0.5317	57.4060	0.0022	0.0497	3.0570
Acaca	ENSMUSG000000020532	-0.4142	133.3009	0.0023	0.0498	-3.0438
Cdk14	ENSMUSG000000028926	-0.3814	445.5196	0.0023	0.0498	-3.0453
Chd2	ENSMUSG000000078671	-0.3276	184.1310	0.0023	0.0498	-3.0463
Eif2b4	ENSMUSG000000029145	0.4394	59.6950	0.0023	0.0498	3.0497
Naca	ENSMUSG000000061315	0.2642	361.9947	0.0023	0.0498	3.0450
Nbea	ENSMUSG000000027799	-0.4388	143.0261	0.0023	0.0498	-3.0485
Pdia3	ENSMUSG000000027248	0.2864	512.9539	0.0024	0.0498	3.0416
Phf14	ENSMUSG000000029629	-0.2686	136.5957	0.0024	0.0498	-3.0418
Plet1	ENSMUSG000000032068	0.3630	86.7600	0.0023	0.0498	3.0472
Pttg1	ENSMUSG000000020415	0.3485	266.7501	0.0023	0.0498	3.0442
Thrb	ENSMUSG000000021779	-0.4384	75.1761	0.0023	0.0498	-3.0430
Uvrag	ENSMUSG000000035354	-0.3698	147.7398	0.0023	0.0498	-3.0493
Arih1	ENSMUSG000000025234	-0.4183	164.4054	0.0024	0.0498	-3.0395
Itm2b	ENSMUSG000000022108	0.3752	2330.3192	0.0024	0.0498	3.0404

# Investigating loss-of-function and human evolution of FOXP2 using global transcriptomic analysis

## Supplementary Table 4

Brain - KO vs WT							
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	t	B
Ide	ENSMUSG00000056999	-0.492	4.985	0.000	0.000	-7.340	14.856
Btaf1	ENSMUSG00000040565	-0.436	5.178	0.000	0.000	-6.297	9.960
Gm47283	ENSMUSG00000096768	-0.566	6.971	0.000	0.000	-6.193	10.159
Gm15417	ENSMUSG00000074466	0.407	4.196	0.000	0.001	5.386	5.977
Klf15	ENSMUSG00000030087	-0.306	4.361	0.000	0.020	-4.740	3.532
Hspb1	ENSMUSG00000004951	-0.288	3.832	0.000	0.020	-4.724	3.461
A230059L01Rik	ENSMUSG00000087627	-0.607	0.798	0.000	0.030	-4.595	0.065
Fgfbp3	ENSMUSG00000047632	-0.374	3.446	0.000	0.032	-4.548	2.330
D130009I18Rik	ENSMUSG000000115432	-0.365	7.352	0.000	0.044	-4.448	2.845
Tmtc2	ENSMUSG00000036019	-0.265	8.153	0.000	0.045	-4.378	2.613
Ly6a	ENSMUSG00000075602	-0.208	4.935	0.000	0.045	-4.379	2.394
Tmem98	ENSMUSG00000035413	-0.290	4.177	0.000	0.045	-4.371	2.249

## Supplementary Table 5

Liver - KO vs WT							
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat	
Rpl19	ENSMUSG00000017404	0.318	325.068	0.000	0.038	3.759	
Ctisl	ENSMUSG00000021477	0.489	1203.651	0.000	0.038	3.781	
Ambp	ENSMUSG00000028356	0.368	5050.710	0.000	0.038	4.150	
Rpl18a	ENSMUSG00000045128	0.330	1034.308	0.000	0.038	3.771	
Rps27	ENSMUSG00000090733	0.338	1172.496	0.000	0.038	3.997	
Rps13	ENSMUSG00000090862	0.407	483.340	0.000	0.038	3.757	
Rpl13	ENSMUSG00000000740	0.295	1686.991	0.000	0.040	3.566	
Rps3a1	ENSMUSG00000028081	0.355	270.583	0.000	0.040	3.588	
Igfbp2	ENSMUSG00000039323	0.762	886.639	0.000	0.040	3.635	
Rps27a-ps3	ENSMUSG00000055093	0.416	622.039	0.000	0.040	3.605	
Gm4149	ENSMUSG00000074800	0.350	285.554	0.000	0.040	3.610	
Rps18-ps5	ENSMUSG000000113061	0.304	956.930	0.000	0.040	3.687	

## Results

### Supplementary Table 6

Bladder - KO vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
Rpl12	ENSMUSG00000038900	-0.423	825.022	0.000	0.005	-4.574
Rps21	ENSMUSG00000039001	-0.871	1169.658	0.000	0.009	-4.320
Rplp1	ENSMUSG00000007892	-0.370	5515.563	0.000	0.010	-4.124
Rps12	ENSMUSG00000061983	-0.556	815.127	0.000	0.010	-4.134
Rps20	ENSMUSG00000028234	-0.338	1554.513	0.000	0.018	-3.941
Tnnt2	ENSMUSG00000026414	0.518	792.399	0.000	0.025	3.810
Rplp2	ENSMUSG00000025508	-0.340	1868.914	0.000	0.033	-3.694
Tpt1	ENSMUSG000000060126	-0.233	717.672	0.000	0.033	-3.675

### Supplementary Table 7

Bladder - KI vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
Hspa8	ENSMUSG00000015656	0.505	1147.880	0.000	0.000	5.679
Gadd45g	ENSMUSG00000021453	1.587	155.739	0.000	0.000	5.480
Pkig	ENSMUSG00000035268	0.737	191.544	0.000	0.000	5.453
Gm15542	ENSMUSG00000083396	1.047	72.339	0.000	0.000	5.311
Txnip	ENSMUSG00000038393	-0.462	540.657	0.000	0.001	-4.883
Tnnt2	ENSMUSG00000026414	0.634	792.399	0.000	0.003	4.694
Cryab	ENSMUSG00000032060	0.980	196.590	0.000	0.005	4.521
Cpe	ENSMUSG00000037852	0.528	282.700	0.000	0.005	4.520
Tnfrsf12a	ENSMUSG00000023905	1.457	81.855	0.000	0.005	4.431
Rasl11a	ENSMUSG00000029641	1.086	41.929	0.000	0.005	4.423
Bdh1	ENSMUSG00000046598	0.755	243.706	0.000	0.005	4.436
Dnm2	ENSMUSG00000033335	-0.316	196.799	0.000	0.006	-4.398
Hspb7	ENSMUSG00000006221	0.887	100.355	0.000	0.008	4.311
Tmem63b	ENSMUSG00000036026	-0.414	98.853	0.000	0.009	-4.266
Abhd2	ENSMUSG00000039202	-0.488	89.940	0.000	0.013	-4.171
Nkd1	ENSMUSG00000031661	0.549	127.575	0.000	0.013	4.143
Gm8355	ENSMUSG000000093798	0.595	291.845	0.000	0.019	4.031
Tmem181b-ps	ENSMUSG00000096780	-0.568	43.242	0.000	0.019	-4.045
Raly	ENSMUSG000000027593	0.442	105.669	0.000	0.025	3.959
Wfdc1	ENSMUSG00000023336	0.742	144.931	0.000	0.031	3.882
Selenos	ENSMUSG00000075701	0.572	87.053	0.000	0.031	3.887
Cnn1	ENSMUSG00000001349	0.763	4886.336	0.000	0.031	3.866
2900026A02Rik	ENSMUSG00000051339	-0.448	85.610	0.000	0.031	-3.859
Ccn1	ENSMUSG000000028195	2.111	144.330	0.000	0.032	3.838
Slc25a4	ENSMUSG00000031633	0.532	932.281	0.000	0.032	3.819
Higd1a	ENSMUSG00000038412	0.618	93.547	0.000	0.032	3.825
Grik5	ENSMUSG00000003378	-0.549	56.828	0.000	0.035	-3.790
Hspa1a	ENSMUSG000000091971	0.862	102.142	0.000	0.041	3.737
Gm10108	ENSMUSG00000062038	0.786	117.837	0.000	0.041	3.728
Gm10053	ENSMUSG00000058927	0.821	64.645	0.000	0.046	3.694
Gtf2e2	ENSMUSG00000031585	0.537	54.966	0.000	0.049	3.670

# Investigating loss-of-function and human evolution of FOXP2 using global transcriptomic analysis

## Supplementary Table 8

Brain - KI vs WT							
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	t	B
Tmem106b	ENSMUSG000000029571	0.223	5.553	0.000	0.000	6.503	12.250
Hspb1	ENSMUSG00000004951	-0.374	3.832	0.000	0.000	-6.116	10.215
Hspa1a	ENSMUSG000000091971	-0.469	4.409	0.000	0.000	-6.001	9.676
mt-Ti	ENSMUSG000000064342	-0.271	5.778	0.000	0.000	-5.662	8.114
Ica1	ENSMUSG000000062995	-0.191	5.869	0.000	0.000	-5.603	7.840
P4ha1	ENSMUSG000000019916	-0.240	5.210	0.000	0.001	-5.453	7.132
Xlr3b	ENSMUSG000000073125	0.922	0.512	0.000	0.001	5.370	4.743
Ahsa2	ENSMUSG000000020288	-0.239	4.507	0.000	0.002	-5.165	5.835
Dnajb1	ENSMUSG000000005483	-0.271	5.799	0.000	0.002	-5.123	5.669
Manf	ENSMUSG000000032575	-0.305	6.556	0.000	0.006	-4.850	4.483
Hspa1b	ENSMUSG000000090877	-0.658	1.534	0.000	0.006	-4.858	3.551
Gm11769	ENSMUSG000000085636	-0.356	4.332	0.000	0.007	-4.778	4.209
Banp	ENSMUSG000000025316	-0.309	4.502	0.000	0.007	-4.776	4.183
Hspa5	ENSMUSG000000026864	-0.208	7.809	0.000	0.008	-4.740	3.986
Srsf5	ENSMUSG000000021134	0.173	6.670	0.000	0.010	4.667	3.726
Tmem125	ENSMUSG000000050854	0.261	3.355	0.000	0.011	4.644	3.649
Creld2	ENSMUSG000000023272	-0.327	5.057	0.000	0.011	-4.628	3.602
Xlr4a	ENSMUSG000000079845	0.775	0.656	0.000	0.012	4.583	2.323
Sdf211	ENSMUSG000000022769	-0.358	4.834	0.000	0.014	-4.535	3.233
Adamts1	ENSMUSG000000022893	0.319	2.866	0.000	0.014	4.532	3.125
Gm15417	ENSMUSG000000074466	-0.342	4.196	0.000	0.024	-4.392	2.683
Aph1b	ENSMUSG000000032375	-0.223	3.760	0.000	0.029	-4.332	2.423
Cdkn1a	ENSMUSG000000023067	-0.307	4.581	0.000	0.030	-4.313	2.378
Insig1	ENSMUSG000000045294	0.177	5.202	0.000	0.034	4.270	2.196
Slc35d3	ENSMUSG000000050473	-0.273	5.166	0.000	0.041	-4.215	2.004

## Supplementary Table 9

Liver - KI vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
Hsp90aa1	ENSMUSG000000021270	0.721	344.123	0.000	0.004	4.624
Serpina1a	ENSMUSG000000066366	0.285	19325.660	0.000	0.004	4.645
Hspa8	ENSMUSG000000015656	0.795	1673.285	0.000	0.024	4.077
Rpl19	ENSMUSG000000017404	0.312	325.068	0.000	0.024	4.088
Serpina1c	ENSMUSG000000079015	0.365	16048.036	0.000	0.024	4.025
Msrb1	ENSMUSG000000075705	0.314	1089.305	0.000	0.028	3.926
Gm6136	ENSMUSG000000084106	0.408	264.433	0.000	0.028	3.905
Azgp1	ENSMUSG000000037053	0.238	2446.620	0.000	0.029	3.849
Rpl19-ps11	ENSMUSG000000081094	0.302	382.692	0.000	0.029	3.838
Rpl6	ENSMUSG000000029614	0.381	605.157	0.000	0.038	3.716
Rpl19-ps7	ENSMUSG0000000117405	0.296	408.624	0.000	0.038	3.736
Vtn	ENSMUSG000000017344	0.458	5502.928	0.000	0.049	3.589
Slc35b1	ENSMUSG000000020873	0.437	163.370	0.000	0.049	3.526
C8b	ENSMUSG000000029656	0.533	441.326	0.000	0.049	3.550
Cd302	ENSMUSG000000060703	0.378	1173.736	0.000	0.049	3.531
Serpina10	ENSMUSG000000061947	0.485	330.265	0.000	0.049	3.604
Kdelr2	ENSMUSG000000079111	0.329	212.946	0.000	0.049	3.525
Rps27	ENSMUSG000000090733	0.271	1172.496	0.000	0.049	3.527

## Results

### Supplementary Table 10

Kidney - KI vs WT						
Gene Name	Gene ID	Log2 Fold Change	Average Expression	P value	Adjusted P Value	stat
Hspa1a	ENSMUSG000000091971	1.271	80.658	0.000	0.001	5.154
Nudt19	ENSMUSG000000034875	0.487	1501.572	0.000	0.012	4.499
Reep6	ENSMUSG000000035504	0.768	203.998	0.000	0.012	4.556
Alas1	ENSMUSG000000032786	0.848	299.841	0.000	0.020	4.329
Car4	ENSMUSG000000000805	0.423	702.905	0.000	0.036	4.146
Igf1r	ENSMUSG000000005533	-0.584	303.737	0.000	0.046	-4.036
Rora	ENSMUSG000000032238	-0.559	112.523	0.000	0.046	-4.009

# DISCUSSION

## SINGLE CELL TRANSCRIPTOMICS AND THE RACE TO THE TOP

**S**ingle cell RNA-seq has undoubtedly altered the field of genomics and will continue to do so. Unlike other technologies, it has had optimization and development built into its core by a community dedicated to mapping every cell within the human body. However, unlike other technologies, such as NGS for example, the utilities of scRNA-seq remain far more varied, and as such, future optimizations and developments will likely occur and will likely be necessary. Or at the very least, once a few exceptional methods have established themselves, they will find a stand-out place within the spectrum of protocols.

There are numerous points of division across the single cell transcriptomics field. One could divide protocols based on how the cells are processed (i.e. plate-based, microfluidics, or droplet), the structure of the libraries (i.e. full length versus counting approaches), the cost of the protocol (i.e. low-cost, in-house protocols versus high-cost, commercial kits), and other various differences. Each of these distinctions has its own set of advantages and disadvantages. Additionally, the perspective of the researcher must also be carefully considered, since an advantage to one may be a disadvantage to another.

With mcSCRB-seq specifically, I aimed to and achieved in developing a highly sensitive, powerful, and cost-efficient plate-based scRNA-seq method. The choices for such a method stemmed primarily from our own motivation, which necessitated a protocol that could integrate fluorescence-activated cell sorting (FACS) and was above all else very cost-efficient. Additionally, the protocol has been cited extensively, likely because the optimizations to sensitivity, which stem from the addition of PEG, and the decreases in amplification biases, due to the use of Terra polymerase, can be easily adapted to other protocols. Therefore, if a research group

already has an established method in place, then simply adapting the improvements to increase sensitivity could suffice. For some researchers, however, mcSCRB-seq might not be an appropriate choice, as plate-based methods tend to be less high-throughput in nature compared to droplet based methods. Additionally, the questions of the experimental study might be more appropriately answered with other methods.

It is precisely this varied use of scRNA-seq that make determining a superior protocol an ongoing challenge. There is no one-size fits all protocol, even though numerous studies have attempted to determine the best candidate. However, as each of these studies has focused on different goals or parameters, they often draw different conclusions. Ziegenhain et al., for example, investigated several hundred cells at sequencing depths in the range of 250,000 to 1 million reads, a design that would be appropriate when investigating specific cells in greater detail. Mereu et al., on the other hand, investigated thousands of cells at ten- to fifty-fold lower sequencing depths. Such a study design is far more appropriate for example, when developing an atlas of various cell types with strong expression differences. And as these goals effectively oppose one another, it is a possible explanation for why mcSCRB-seq, or more specifically its predecessor SCR-seq, performed strongly in the initial comparative study (Ziegenhain et al., 2017) and poorly in a more recent study (Mereu et al., 2020). Comparison studies can also draw opposing conclusions, not only based on their goals, but also based on their design. Although Meru et al. reflects the real-world situation of a consortium, this creates a potential for more technical variables, for example, where and how the sequencing is performed. For some methods the sequencing was carried out by the project organizers, whereas for other methods the sequencing was performed by the library preparers. And in such a case, this means that a variety of equipment and quality filtering were introduced into the study.

Thus, whether further optimizations to mcSCRB-seq should be made to improve its ability in cell atlas creation remains unclear. In some way, this would move the protocol more in line with the current trend of the single cell community. Such protocols (e.g. 10X Genomics Chromium) already exist, perform very well, and fill this role, as shown in Mereu et al. But once such highly detailed atlases are publicly available, protocols that are primarily designed to process many cells at low sequencing depth may fall out of fashion. The focus of the genomics community may move towards characterizing few cells in greater detail, and using atlases to support such findings. The direction could also progress towards a dual-model system, where high-throughput, low depth methods are used as an initial study, and then lower-throughput, high depth methods are used to

investigate a certain subset of cells in far greater detail. Lastly, now that numerous, high quality methods exist, a large focus of the community has transferred to developing spatial transcriptomic protocols as well as the necessary computational tools to carry out the analysis. Therefore, rather than additional optimizations to mcSCR-seq, time may be better spent contributing to the field of spatial transcriptomics.

What is clear, however, is that although a complete overhaul of mcSCR-seq would likely not benefit the greater scientific community, small optimizations and changes, especially those utilized in the development of prime-seq, could be beneficial. For example, utilizing longer barcoded oligo(d)T primers would allow the protocol to be more easily multiplexed with 10X Genomics 3' libraries, a current industry leader in gene expression analysis. Such multiplexing is especially important as it allows one to utilize the newest high-throughput Illumina sequencers (e.g. Novaseq) which decreases sequencing costs. Additionally, the recently developed Smart-seq3 underwent extensive testing for the lysis buffer (Hagemann-Jensen et al., 2020), and such findings could be adapted to mcSCR-seq.

## **BULK RNA-SEQ IN A SINGLE CELL WORLD**

**I**n just fifteen years we went from characterizing gene expression in just a couple samples to now studying the transcriptome of hundreds of thousands of cells from many samples. And while improved technologies to answer biological questions relating to gene expression have been developed and utilized, often the latest and greatest technologies can be excessive depending on the goal at hand. Therefore, it is important to understand the space these new methods occupy.

Firstly, let us examine the medical community. Generally, direct research is not a primary focus, rather most physicians prioritize patient care, diagnostics, and treatment. That said, there are occasional instances of overlap between the medical and biomedical/biological communities, especially in terms of technological advances. Sequencing, both Sanger and NGS, have already proven themselves as essential tools for physicians, capable of diagnosing genetic disorders not previously seen with lower-resolution patient karyotyping. Microarrays have also become a staple, capable of providing chromosomal analysis, oncological panels, as well as expression profiles for well known target genes. However, RNA-seq has yet to integrate fully into the repertoire of diagnostic tools.

Secondly, within the context of biomedical or biological research, especially in academic settings, RNA-seq has become a staple, but single-cell, spatial, and long-read methods have yet to be utilized across the board. For example, the power, as well as the cost efficiency, of RNA-seq compared to qPCR has been shown many times over (Alpern et al., 2019); yet in the last five years, almost the same number of published studies are available on PubMed that use qPCR compared to RNA-seq (55.6 and 59.8 thousand, respectively). This highlights that even after a decade and a half, an outdated and more expensive technology is still being substantially utilized. The explanation for this is surely multi-faceted, however, a likely reason is that implementation of new technology takes time as well as resources. This is even more apparent in the context of research, where resources may be a limiting factor. Additionally, the way bulk RNA-seq continues to be useful even when scRNA-seq has become more standardized, qPCR will still have a place among molecular biologists, especially for single-target applications or quality control.

Lastly, there is genomic research, a small subset of the greater picture where innovation has been at the forefront, and the newest techniques are heralded and praised. This perspective is unusual in that method development is itself a large component of the research conducted.

Thus, even with the development of new genomic methods, the context of where and how they will be utilized must be taken into account. For example, a genomic researcher may claim that the future is single cell, and bulk no longer has a place. But a physician that is analyzing patient samples may be unable to perform diagnostic single cell studies, either due to limited resources or lack of appropriate samples. Bulk RNA-seq then could likely be sufficient and a substantial improvement on the microarray panels currently available. Such use is even supported by single-cell biologists, who provide datasets, in part, to be used references and enable less comprehensive and less resource intensive future studies.

With this in mind, it is easy to see that although prime-seq is a bulk protocol, it may very well be one of the most useful in the field of genomics. Bulk RNA-seq is generally more affordable due to the smaller scale of the experiment, more successful with difficult samples or unverified tissues, and can provide detailed data capturing almost all expressed genes in the sample. These characteristics of affordability, robustness, and sensitivity were all prioritized in the development of prime-seq, in order to create a protocol that could benefit not only the genomics community, but also biomedical and biological researchers, and one day perhaps even medical professionals.

## CHARACTERIZING THE EFFECT OF *FOXP2* ON DOPAMINERGIC STRIATAL NEURONS

**A**ddressing a phenotype as complex as human language, specifically in relation to *FOXP2*, requires a multi-experiment approach from various perspectives. Previous studies have examined *FOXP2* in numerous organisms including humans (Fisher et al., 1998), mice (Enard et al., 2002), songbirds (Wohlgemuth et al., 2014; Fee and Scharff, 2010), and bats (Li et al., 2007). However, many of these studies have not been thorough, such as those conducted in humans due to ethical reasons. The underlying mechanisms responsible for the observed effects of *FOXP2* knockouts or the effects of introducing a human variant into a model organism, have yet to be well understood. Within this work, I aimed to build a comprehensive dataset of various tissues assessing the effect on the transcriptome in both cases. And, although this *FOXP2* dataset is one of the largest, as with any experiment more questions may be raised than answered.

The human specific substitutions in *FOXP2* have been implicated in learning by modulating CBG circuitry (Enard et al., 2009). Specifically, dopamine levels and synaptic plasticity are affected (Co et al., 2020; Schreiweis et al., 2014), which is supported by my own data as well. However, what appears to be clear is that to further elucidate the mechanism, specifically the targets and the affected genes in each mouse model, additional experiments will be required. Future studies will have to investigate the loss-of-function effect of *FOXP2* (knockout) and the gain-of-function effect in human *FOXP2* (humanized) by manipulating learning in these mouse models. As *FOXP2* is involved in CBG circuitry, understanding its role in the transition from goal-directed learning to habitual behavior will be essential. Testing the hypothesis that the human variant is responsible for a faster transition to behavioral automatization, and therefore responsible for the evolution of human speech, will be the next step in answering the question: “Why are we able to speak and our closest relatives are not?”.



# REFERENCES

- Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* 2010;11(12):R119. doi: 10.1186/gb-2010-11-12-r119.
- Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature.* 2003 Mar 13;422(6928):198-207. doi: 10.1038/nature01511.
- Allfrey VG, Faulkner R, Mirsky AE . Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc Natl Acad Sci U S A.* 1964 May;51(5):786-94. doi: 10.1073/pnas.51.5.786.
- Alpern D, Gardeux V, Russeil J, Mangeat B, Meireles-Filho ACA, Breyse R, Hacker D, Deplancke B. BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol.* 2019 Apr 19;20(1):71. doi: 10.1186/s13059-019-1671-x.
- Alwine JC, Kemp DJ, Stark GR. 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.* 1977 Dec;74(12):5350-4. doi: 10.1073/pnas.74.12.5350.
- Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 2020 Feb 7;21(1):30. doi: 10.1186/s13059-020-1935-5.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000 May;25(1):25-9. doi: 10.1038/75556.
- Avery OT, Macleod CM, McCarty M. 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 iiii. *J Exp Med.* 1944 Feb 1;79(2):137-58. doi: 10.1084/jem.79.2.137.
- Bagnoli JW, Ziegenhain C, Janjic A, Wange LE, Vieth B, Parekh S, Geuder J, Hellmann I, Enard W. Mcscrb-Seq protocol [Internet]. protocols.io; 2018 [cited 2021 Sep 20]. Available from: <https://dx.doi.org/10.17504/protocols.io.p9kdr4w>
- Bagnoli JW, Ziegenhain C, Janjic A, Wange LE, Vieth B, Parekh S, Geuder J, Hellmann I, Enard W. Sensitive and powerful single-cell RNA sequencing using mcSCRb-seq. *Nat Commun.* 2018 Jul 26;9(1):2937. doi: 10.1038/s41467-018-05347-6.
- Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, Mardis ER, Sadar MD, Siddiqui AS, Marra MA, Jones SJ. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics.* 2006 Sep 29;7:246. doi: 10.1186/1471-2164-7-246.

## References

- Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E, Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott-Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschield CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurler ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008 Nov 6;456(7218):53-9. doi: 10.1038/nature07517.
- Blanchet S, Cornu D, Hatin I, Grosjean H, Bertin P, Namy O. Deciphering the reading of the genetic code by near-cognate tRNA. *Proc Natl Acad Sci U S A*. 2018 Mar 20;115(12):3018-3023. doi: 10.1073/pnas.1715578115.
- Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. High-resolution mapping and characterization of open chromatin across the genome. *Cell*. 2008 Jan 25;132(2):311-22. doi: 10.1016/j.cell.2007.12.014.
- Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 2013 Dec;10(12):1213-8. doi: 10.1038/nmeth.2688.
- Burnette WN. "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate--polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Anal Biochem*. 1981 Apr;112(2):195-203. doi: 10.1016/0003-2697(81)90281-5.
- Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schneeberger K. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. *PLoS One*. 2016 Jun 21;11(6):e0157779. doi: 10.1371/journal.pone.0157779.
- Carvin CD, Dhasarathy A, Friesenhahn LB, Jessen WJ, Kladde MP. Targeted cytosine methylation for in vivo detection of protein-DNA interactions. *Proc Natl Acad Sci U S A*. 2003 Jun 24;100(13):7743-8. doi: 10.1073/pnas.1332672100.

- Casneuf T, Van de Peer Y, Huber W. In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics*. 2007 Nov 26;8:461. doi: 10.1186/1471-2105-8-461.
- Cheung F, Haas BJ, Goldberg SM, May GD, Xiao Y, Town CD. Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics*. 2006 Oct 24;7:272. doi: 10.1186/1471-2164-7-272.
- Chiang PW, Song WJ, Wu KY, Korenberg JR, Fogel EJ, Van Keuren ML, Lashkari D, Kurnit DM. Use of a fluorescent-PCR reaction to detect genomic sequence copy number and transcriptional abundance. *Genome Res*. 1996 Oct;6(10):1013-26. doi: 10.1101/gr.6.10.1013.
- Co M, Hickey SL, Kulkarni A, Harper M, Konopka G. Cortical Foxp2 Supports Behavioral Flexibility and Developmental Dopamine D1 Receptor Expression. *Cereb Cortex*. 2020 Mar 14;30(3):1855-1870. doi: 10.1093/cercor/bhz209.
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczeni-ak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016 Jan 26;17:13. doi: 10.1186/s13059-016-0881-8.
- Crick FH. On protein synthesis. *Symp Soc Exp Biol*. 1958;12:138-63.
- Darst RP, Nabitsi NH, Pardo CE, Riva A, Kladde MP. DNA methyltransferase accessibility protocol for individual templates by deep sequencing. *Methods Enzymol*. 2012;513:185-204. doi: 10.1016/B978-0-12-391938-0.00008-2.
- Edfors F, Danielsson F, Hallström BM, Käll L, Lundberg E, Pontén F, Forsström B, Uhlén M. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol*. 2016 Oct 20;12(10):883. doi: 10.15252/msb.20167144.
- Eisenstein, M. Westward expansion. *Nat Methods*. 2005 Aug;2(8):796. doi: 10.1038/nmeth1005-796.
- Emrich SJ, Barbazuk WB, Li L, Schnable PS. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res*. 2007 Jan;17(1):69-73. doi: 10.1101/gr.5145806.
- Enard W, Gehre S, Hammerschmidt K, Höltner SM, Blass T, Somel M, Brückner MK, Schreiweis C, Winter C, Sohr R, Becker L, Wiebe V, Nickel B, Giger T, Müller U, Groszer M, Adler T, Aguilar A, Bolle I, Calzada-Wack J, Dalke C, Ehrhardt N, Favor J, Fuchs H, Gailus-Durner V, Hans W, Hölzlwimmer G, Javaheri A, Kalaydjiev S, Kallnik M, Kling E, Kunder S, Mossbrugger I, Naton B, Racz I, Rathkolb B, Rozman J, Schrewe A, Busch DH, Graw J, Ivandic B, Klingenspor M, Klopstock T, Ollert M, Quintanilla-Martinez L, Schulz H, Wolf E, Wurst W, Zimmer A, Fisher SE, Morgenstern R, Arendt T, de Angelis MH, Fischer J, Schwarz J, Pääbo S. A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell*. 2009 May 29;137(5):961-71. doi: 10.1016/j.cell.2009.03.041.
- Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Pääbo S. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*. 2002 Aug 22;418(6900):869-72. doi: 10.1038/nature01025.
- Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*. 1994 Nov;5(11):976-89. doi: 10.1016/1044-0305(94)80016-2.
- Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 1998 Mar;8(3):186-94.
- Fee MS, Scharff C. The songbird as a model for the generation and learning of complex sequential behaviors. *ILAR J*. 2010;51(4):362-77. doi: 10.1093/ilar.51.4.362.

## References

- Fisher SE, Vargha-Khadem F, Watkins KE, Monaco AP, Pembrey ME. Localisation of a gene implicated in a severe speech and language disorder. *Nat Genet.* 1998 Feb;18(2):168-70. doi: 10.1038/ng0298-168.
- Galalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S, McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D, Juul S, Clarke J, Heron AJ, Turner DJ. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods.* 2018 Mar;15(3):201-206. doi: 10.1038/nmeth.4577.
- Gibson UE, Heid CA, Williams PM. A novel method for real time quantitative RT-PCR. *Genome Res.* 1996 Oct;6(10):995-1001. doi: 10.1101/gr.6.10.995.
- Gilbert W. Why genes in pieces? *Nature.* 1978 Feb 9;271(5645):501. doi: 10.1038/271501a0.
- Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.* 2016 May 17;17(6):333-51. doi: 10.1038/nrg.2016.49.
- Goryshin IY, Reznikoff WS. Tn5 in vitro transposition. *J Biol Chem.* 1998 Mar 27;273(13):7367-74. doi: 10.1074/jbc.273.13.7367.
- Hagemann-Jensen M, Ziegenhain C, Chen P, Ramsköld D, Hendriks GJ, Larsson AJM, Faridani OR, Sandberg R. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat Biotechnol.* 2020 Jun;38(6):708-714. doi: 10.1038/s41587-020-0497-0.
- Hansen KD, Wu Z, Irizarry RA, Leek JT. Sequencing technology does not eliminate biological variability. *Nat Biotechnol.* 2011 Jul 11;29(7):572-3. doi: 10.1038/nbt.1910.
- Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. *Nat Rev Genet.* 2017 Aug;18(8):473-484. doi: 10.1038/nrg.2017.44.
- Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, Dor Y, Regev A, Yanai I. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 2016 Apr 28;17:77. doi: 10.1186/s13059-016-0938-8.
- Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development. *Science.* 1975 Jan 24;187(4173):226-32.
- Hotchkiss RD. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem.* 1948 Aug;175(1):315-32.
- Hurst JA, Baraitser M, Auger E, Graham F, Norell S. An extended family with a dominantly inherited speech disorder. *Dev Med Child Neurol.* 1990 Apr;32(4):352-5. doi: 10.1111/j.1469-8749.1990.tb16948.x.
- Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, Ziegenhain C, Enard W. prime-seq protocol [Internet]. protocols.io; 2020 [cited 2021 Sep 20]. Available from: <https://dx.doi.org/10.17504/protocols.io.s9veh66>
- Jenuwein T, Allis CD. Translating the histone code. *Science.* 2001 Aug 10;293(5532):1074-80. doi: 10.1126/science.1063127.
- Jessen WJ, Dhasarathy A, Hoose SA, Carvin CD, Risinger AL, Kladde MP. Mapping chromatin structure in vivo using DNA methyltransferases. *Methods.* 2004 May;33(1):68-80. doi: 10.1016/j.ymeth.2003.10.025.
- Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* 2012 Dec;22(12):2497-506. doi: 10.1101/gr.143008.112.

- Kim JH, Hwang J, Jung JH, Lee HJ, Lee DY, Kim SH. Molecular networks of FOXP family: dual biologic functions, interplay with other molecules and clinical implications in cancer progression. *Mol Cancer*. 2019 Dec 9;18(1):180. doi: 10.1186/s12943-019-1110-3.
- Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet*. 2019 May;20(5):273-282. doi: 10.1038/s41576-018-0088-9.
- Köhler A, Hurt E. Exporting RNA from the nucleus to the cytoplasm. *Nat Rev Mol Cell Biol*. 2007 Oct;8(10):761-73. doi: 10.1038/nrm2255.
- Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. *Science*. 1974 May 24;184(4139):868-71. doi: 10.1126/science.184.4139.868.
- Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. *Cold Spring Harb Protoc*. 2015 Apr 13;2015(11):951-69. doi: 10.1101/pdb.top084970.
- Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*. 2001 Oct 4;413(6855):519-23. doi: 10.1038/35097076.
- Lam EW, Brosens JJ, Gomes AR, Koo CY. Forkhead box proteins: tuning forks for transcriptional harmony. *Nat Rev Cancer*. 2013 Jul;13(7):482-95. doi: 10.1038/nrc3539.
- Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglu S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shores N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*. 2012 Sep;22(9):1813-31. doi: 10.1101/gr.136184.111.
- Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014 Feb 3;15(2):R29. doi: 10.1186/gb-2014-15-2-r29.
- Li G, Wang J, Rossiter SJ, Jones G, Zhang S. Accelerated FoxP2 evolution in echolocating bats. *PLoS One*. 2007 Sep 19;2(9):e900. doi: 10.1371/journal.pone.0000900.
- Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR 3rd. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol*. 1999 Jul;17(7):676-82. doi: 10.1038/10890.
- Livingstone M, Atas E, Meller A, Sonenberg N. Mechanisms governing the control of mRNA translation. *Phys Biol*. 2010 May 12;7(2):021001. doi: 10.1088/1478-3975/7/2/021001.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8.
- Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*. 1997 Sep 18;389(6648):251-60. doi: 10.1038/38444.
- Mardis ER. Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)*. 2013;6:287-303. doi: 10.1146/annurev-anchem-062012-092628.
- Marx V. Method of the Year: spatially resolved transcriptomics. *Nat Methods*. 2021 Jan;18(1):9-14. doi: 10.1038/s41592-020-01033-y.

## References

- Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, Batlle E, Sagar, Grün D, Lau JK, Boutet SC, Sanada C, Ooi A, Jones RC, Kaihara K, Brampton C, Talaga Y, Sasagawa Y, Tanaka K, Hayashi T, Braeuning C, Fischer C, Sauer S, Trefzer T, Conrad C, Adiconis X, Nguyen LT, Regev A, Levin JZ, Parekh S, Janjic A, Wange LE, Bagnoli JW, Enard W, Gut M, Sandberg R, Nikaido I, Gut I, Stegle O, Heyn H. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol.* 2020 Jun;38(6):747-755. doi: 10.1038/s41587-020-0469-4.
- Minnoye, L, Marinov, G K, Krausgruber, T, Pan L, Marand A P, Secchia S, Greenleaf W J, Furlong E E M, Zhao K, Schmitz R J, Bock C, Aerts S. Chromatin accessibility profiling methods. *Nat Rev Methods Primers.* 2021;1,10. doi: 10.1038/s43586-020-00008-9
- Moor AE, Itzkovitz S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr Opin Biotechnol.* 2017 Aug;46:126-133. doi: 10.1016/j.copbio.2017.02.004.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 2008 Jul;5(7):621-8. doi: 10.1038/nmeth.1226.
- O'Neil D, Glowatz H, Schlumpberger M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Curr Protoc Mol Biol.* 2013 Jul;Chapter 4:Unit 4.19. doi: 10.1002/0471142727.mb0419s103.
- Pardo CE, Nabils NH, Darst RP, Kladde MP. Integrated DNA methylation and chromatin structural analysis at single-molecule resolution. *Methods Mol Biol.* 2015;1288:123-41. doi: 10.1007/978-1-4939-2474-5\_9.
- Perry RP. Processing of RNA. *Annu Rev Biochem.* 1976;45:605-29. doi: 10.1146/annurev.bi.45.070176.003133.
- 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 Nov;10(11):1096-8. doi: 10.1038/nmeth.2639.
- Pine PS, Munro SA, Parsons JR, McDaniel J, Lucas AB, Lozach J, Myers TG, Su Q, Jacobs-Helber SM, Salit M. Evaluation of the External RNA Controls Consortium (ERCC) reference material using a modified Latin square design. *BMC Biotechnol.* 2016 Jun 24;16(1):54. doi: 10.1186/s12896-016-0281-x.
- Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Göttgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klennerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundberg J, Majumder P, Marioni JC, Merad M, Mhlanga M, Nawijn M, Netea M, Nolan G, Pe'er D, Phillipakis A, Ponting CP, Quake S, Reik W, Rozenblatt-Rosen O, Sanes J, Satija R, Schumacher TN, Shalek A, Shapiro E, Sharma P, Shin JW, Stegle O, Stratton M, Stubbington MJT, Theis FJ, Uhlen M, van Oudenaarden A, Wagner A, Watt F, Weissman J, Wold B, Xavier R, Yosef N; Human Cell Atlas Meeting Participants. The Human Cell Atlas. *Elife.* 2017 Dec 5;6:e27041. doi: 10.7554/eLife.27041.
- Renart J, Reiser J, Stark GR. Transfer of proteins from gels to diazobenzylxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure. *Proc Natl Acad Sci U S A.* 1979 Jul;76(7):3116-20. doi: 10.1073/pnas.76.7.3116.
- Riba A, Di Nanni N, Mittal N, Arhné E, Schmidt A, Zavolan M. Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. *Proc Natl Acad Sci U S A.* 2019 Jul 23;116(30):15023-15032. doi: 10.1073/pnas.1817299116.

- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015 Apr 20;43(7):e47. doi: 10.1093/nar/gkv007.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010 Jan 1;26(1):139-40. doi: 10.1093/bioinformatics/btp616.
- Roeder RG, Rutter WJ. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature.* 1969 Oct 18;224(5216):234-7. doi: 10.1038/224234a0.
- Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, Bani Asadi N, Gerstein MB, Wong WH, Snyder MP, Schadt E, Lam HYK. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat Commun.* 2017 Jul 5;8(1):59. doi: 10.1038/s41467-017-00050-4.
- Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* 1977 Dec;74(12):5463-7. doi: 10.1073/pnas.74.12.5463.
- Sasagawa Y, Danno H, Takada H, Ebisawa M, Tanaka K, Hayashi T, Kurisaki A, Nikaido I. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* 2018 Mar 9;19(1):29. doi: 10.1186/s13059-018-1407-3.
- Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* 1995 Oct 20;270(5235):467-70. doi: 10.1126/science.270.5235.467.
- Schreiweis C, Bornschein U, Burguière E, Kerimoglu C, Schreiter S, Dannemann M, Goyal S, Rea E, French CA, Puliadi R, Groszer M, Fisher SE, Mundry R, Winter C, Hevers W, Pääbo S, Enard W, Graybiel AM. Humanized Foxp2 accelerates learning by enhancing transitions from declarative to procedural performance. *Proc Natl Acad Sci U S A.* 2014 Sep 30;111(39):14253-8. doi: 10.1073/pnas.1414542111.
- SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol.* 2014 Sep;32(9):903-14. doi: 10.1038/nbt.2957.
- Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform.* 2015 Jan;16(1):59-70. doi: 10.1093/bib/bbt086.
- Shatkin AJ, Manley JL. The ends of the affair: capping and polyadenylation. *Nat Struct Biol.* 2000 Oct;7(10):838-42. doi: 10.1038/79583.
- Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. DNA sequencing at 40: past, present and future. *Nature.* 2017 Oct 19;550(7676):345-353. doi: 10.1038/nature24286.
- Shendure J. The beginning of the end for microarrays? *Nat Methods.* 2008 Jul;5(7):585-7. doi: 10.1038/nmeth0708-585.
- Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc.* 2010 Feb;2010(2):pdb.prot5384. doi: 10.1101/pdb.prot5384.
- Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *Biorxiv.* 2014 Mar 5;Preprint. doi: 10.1101/003236 (2014).
- Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet.* 2019 Nov;20(11):631-656. doi: 10.1038/s41576-019-0150-2.

## References

- Starks RR, Biswas A, Jain A, Tuteja G. Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics Chromatin*. 2019 Feb 22;12(1):16. doi: 10.1186/s13072-019-0260-2.
- Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*. 2015 Mar;16(3):133-45. doi: 10.1038/nrg3833.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005 Oct 25;102(43):15545-50. doi: 10.1073/pnas.0506580102.
- Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*. 2009 May;6(5):377-82. doi: 10.1038/nmeth.1315.
- Tavares L, Alves PM, Ferreira RB, Santos CN. Comparison of different methods for DNA-free RNA isolation from SK-N-MC neuroblastoma. *BMC Res Notes*. 2011;4:3. doi: 10.1186/1756-0500-4-3.
- Thomson JJ. XL. Cathode Rays. *Phil Mag*. 1897 Oct;5(44):269,293-316. doi: 10.1080/14786449708621070.
- Towbin H, Staehelin T, Gordon J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc Natl Acad Sci U S A*. 1979 Sep;76(9):4350-4. doi: 10.1073/pnas.76.9.4350.
- Vargha-Khadem F, Watkins K, Alcock K, Fletcher P, Passingham R. Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. *Proc Natl Acad Sci U S A*. 1995 Jan 31;92(3):930-3. doi: 10.1073/pnas.92.3.930.
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science*. 1995 Oct 20;270(5235):484-7. doi: 10.1126/science.270.5235.484.
- Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun*. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.
- Wang AM, Doyle MV, Mark DF. Quantitation of mRNA by the polymerase chain reaction. *Proc Natl Acad Sci U S A*. 1989 Dec;86(24):9717-21. doi: 10.1073/pnas.86.24.9717.
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009 Jan;10(1):57-63. doi: 10.1038/nrg2484.
- Warner JR. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci*. 1999 Nov;24(11):437-40. doi: 10.1016/s0968-0004(99)01460-7.
- Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol*. 2007 May;144(1):32-42. doi: 10.1104/pp.107.096677.
- Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*. 2017 Jan 17;18(1):38. doi: 10.1186/s12859-016-1457-z.
- Wohlgemuth S, Adam I, Scharff C. FoxP2 in songbirds. *Curr Opin Neurobiol*. 2014 Oct;28:86-93. doi: 10.1016/j.conb.2014.06.009.
- Yates III, J. A century of mass spectrometry: from atoms to proteomes. *Nat Methods*. 2011 Jul;8:633-637. doi: 10.1038/nmeth.1659.

- Yi H, Cho YJ, Won S, Lee JE, Jin Yu H, Kim S, Schroth GP, Luo S, Chun J. Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic Acids Res.* 2011 Nov 1;39(20):e140. doi: 10.1093/nar/gkr617.
- Zaret K. Micrococcal nuclease analysis of chromatin structure. *Curr Protoc Mol Biol.* 2005 Feb;Chapter 21:Unit 21.1. doi: 10.1002/0471142727.mb2101s69.
- Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep.* 2018 Mar 19;8(1):4781. doi: 10.1038/s41598-018-23226-4.
- Ziegenhain C, Vieth B, Parekh S, Hellmann I, Enard W. Quantitative single-cell transcriptomics. *Brief Funct Genomics.* 2018 Jul 1;17(4):220-232. doi: 10.1093/bfpg/ely009.
- Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell.* 2017 Feb 16;65(4):631-643.e4. doi: 10.1016/j.molcel.2017.01.023.

# FIGURE LIST

<b>Figure 1</b>	A general overview of molecular biology.....	<b>2</b>
<b>Figure 2</b>	Illumina Sequencing by Synthesis .....	<b>12</b>
<b>Figure 3</b>	General RNA Sequencing Workflow .....	<b>15</b>
<b>Figure 4</b>	FOPX2 Overview .....	<b>20</b>

# PUBLICATION LIST

## Included in this Work

**Janjic A\***, Wange LE\*, Bagnoli JW, Geuder J, et al. Prime-Seq, Efficient and Powerful Bulk RNA-Sequencing. *Preprint*. bioRxiv. 2021 Sep. doi: 10.1101/2021.09.27.459575

Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, Batlle E, Sagar, Grün D, Lau JK, Boutet SC, Sanada C, Ooi A, Jones RC, Kaihara K, Brampton C, Talaga Y, Sasagawa Y, Tanaka K, Hayashi T, Braeuning C, Fischer C, Sauer S, Trefzer T, Conrad C, Adiconis X, Nguyen LT, Regev A, Levin JZ, Parekh S, **Janjic A**, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol*. 2020;38(6):747-755. doi:10.1038/s41587-020-0469-4

Bagnoli JW\*, Ziegenhain C\*, **Janjic A\***, et al. Sensitive and powerful single-cell RNA sequencing using mcSCR-seq. *Nat Commun*. 2018;9(1):2937. doi:10.1038/s41467-018-05347-6

\*contributed equally

## Not Included in this Work

Chavarria-Pizarro T, Resl P, **Janjic A**, Werth S. Gene expression responses to thermal shifts in the endangered lichen *Lobaria pulmonaria*. *Mol Ecol*. 2021. doi: 10.1111/mec.16281.

Kempf J, Kneller K, Hersbach BA, Petrik D, Riedemann T, Bednarova V, **Janjic A**, et al. Heterogeneity of neurons reprogrammed from spinal cord astrocytes by the proneural factors *Ascl1* and *Neurogenin2*. *Cell Rep*. 2021;36(7):109571. doi: 10.1016/j.celrep.2021.109571

Pekayvaz A, 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. *Preprint*. 2021. doi: 10.1101/2021.02.03.429351

Geuder J, Ohnuki M, Wange LE, **Janjic A**, et al. A non-invasive method to generate induced pluripotent stem cells from primate urine. *Sci Rep*. 2021;11(1):3516. doi: 10.1038/s41598-021-82883-0

Shami A, Atzler D, Bosmans LA, Winkels H, Meiler S, Lacy M, van Tiel C, Ta Megens R, Nitz K, Baardman J, Kusters P, Seijkens T, Buerger C, **Janjic A**, 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(31):2938-2948. doi: 10.1093/eurheartj/ehaa484

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(2):121-142. doi:10.1007/s00401-020-02176-0

Kozak EL, Palit S, Miranda-Rodríguez JR, Janjic A, Böttcher A, Lickert H, Enard W, Theis FJ, López-Schier H. Epithelial Planar Bipolarity Emerges from Notch-Mediated Asymmetric Inhibition of Emx2. *Curr Biol*. 2020;30(6):1142-1151.e6. doi: 10.1016/j.cub.2020.01.027.

Bagnoli JW, Wange LE, **Janjic A**, Enard W. Studying Cancer Heterogeneity by Single-Cell RNA Sequencing. *Methods Mol Biol*. 2019;1956:305–319. doi:10.1007/978-1-4939-9151-8\_14

Emdad L, **Janjic A**, et al. Suppression of miR-184 in malignant gliomas up-regulates SND1 and promotes tumor aggressiveness. *Neuro Oncol*. 2015;17(3):419–429. doi:10.1093/neuonc/nou220

Dasgupta S, Menezes ME, Das SK, Emdad L, **Janjic A**, et al. Novel role of MDA-9/syntenin in regulating urothelial cell proliferation by modulating EGFR signaling. *Clin Cancer Res*. 2013;19(17):4621–4633. doi:10.1158/1078-0432.CCR-13-0585

# PUBLICATION COPY RIGHTS

## SENSITIVE AND POWERFUL SINGLE-CELL RNA SEQUENC- ING USING MCSCRB-SEQ

**Author:** Johannes W. Bagnoli et al.

**Publication:** Nature Communications

**Publisher:** Springer Nature

**Date:** Jul 26, 2018

**Copyright** © 2018, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

## **BENCHMARKING SINGLE-CELL RNA-SEQUENCING PROTOCOLS FOR CELL ATLAS PROJECTS**

**Author:** Elisabetta Mereu et al

**Publication:** Nature Biotechnology

**Publisher:** Springer Nature

**Date:** Apr 6, 2020

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

### Author Request

Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:

To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).

The author and any academic institution, where they work, at the time may reproduce the contribution for the purpose of course teaching.

To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them.

To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website.

Authors wishing to use the published version of their article for promotional use or on a web site must request in the normal way.

# PRIME-SEQ, EFFICIENT AND POWERFUL BULK RNA-SEQUENCING

**Author:** Aleksandar Janjic et al

**Publication:** bioRxiv

**Publisher:** Cold Spring Harbor Laboratory

**Date:** Sept 28, 2021

**Copyright** © 2021, The Author(s), CC-BY-NC-ND 4.0

Author Request

The copyright holder for this preprint is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.