
Optimising gene expression profiling using RNA-seq

Swati Parekh

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



München 2018

1. Gutachter: Prof. Wolfgang Enard

2. Gutachter: Prof. John Parsch

Tag der Abgabe: 13.02.2018

Tag der mündlichen Prüfung: 11.06.2018

Statutory declaration and statement

(Eidestattliche Versicherung und Erklärung)

Eidesstattliche Erklärung

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

München, den 14.06.2018

Parekh Swati

(Unterschrift)

Erklärung

Hiermit erkläre ich,

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

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

dass ich mich mit Erfolg der Doktorprüfung im Hauptfach und in den Nebenfächern bei der Fakultät für der unterzogen habe.

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

München, den 14.06.2018

Parekh Swati

(Unterschrift)

Table of contents

Abbreviations	5
List of publications	6
Declaration of contribution as co-author	8
Aim of the thesis	13
Summary	14
Introduction	16
Gene expression profiling	16
Scope of sequencing	19
RNA sequencing	21
RNA-seq data processing	23
Demultiplexing	24
Quality Control (QC)	24
Mapping and Quantification	25
Amplification noise	28
Single-cell RNA-sequencing (scRNA-seq)	30
Comparative transcriptomics across species	34
Computational simulation	36
Results	38
Assessing the impact of amplification noise in RNA-seq	38
The impact of amplification on differential expression analyses by RNA-seq	38
Identifying and addressing issues in single-cell RNA-seq analysis	58
Comparative Analysis of Single-Cell RNA Sequencing Methods	58
powsimR: Power analysis for bulk and single cell RNA-seq experiments	92
zUMIs: A fast and flexible pipeline to process RNA sequencing data with UMIs	138
Optimising cross species differential expression analysis	151
Strategies for RNA-seq differential expression analysis for closely related species	151
Discussion	180
Impact of amplification noise in quantitative RNA-seq	180
Identifying and addressing computational challenges in single-cell RNA-seq data analysis	183
Optimising cross species differential expression analysis	188
Conclusions and Outlook	192
References	193
List of Figures	211
List of Tables	211
Acknowledgments	212
Curriculum vitae	214

Abbreviations

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
mRNA	messenger RNA
cDNA	Complementary DNA
RPA	Ribonuclease Protection Assay
PCR	Polymerase Chain Reaction
RT-PCR	Reverse Transcription Polymerase Chain Reaction
qPCR	Quantitative Polymerase Chain Reaction
RNA-seq	RNA sequencing
bp	Base pair
HTS	High Throughput Sequencing
NGS	Next Generation Sequencing
poly-A	polyadenylated
UMI	Unique Molecular Identifier
QC	Quality Control
scRNA-seq	Single-cell RNA sequencing
PE	paired-end
SE	single-end
nh	non-human
FDR	False Discovery Rate
FPR	False Positive Rate
TPR	True Positive Rate

List of publications

Publications included in the thesis:

1. **Parekh S**, Ziegenhain C, Vieth B, Enard W, Hellmann I. *The impact of amplification on differential expression analyses by RNA-seq*. **Scientific Reports 2016**.
2. Ziegenhain C, Vieth B, **Parekh S**, Reinius B, Guillaumet-Adkins A, Smets M, et al. *Comparative Analysis of Single-Cell RNA Sequencing Methods*. **Molecular Cell 2017**.
3. Vieth B, Ziegenhain C, **Parekh S**, Enard W, Hellmann I. *powsimR: Power analysis for bulk and single cell RNA-seq experiments*. **Bioinformatics 2017**.
4. **Parekh S***, Ziegenhain C*, Vieth B, Enard W, Hellmann I. *zUMIs: A fast and flexible pipeline to process RNA sequencing data with UMIs*. **GigaScience / In revision**.
5. **Parekh S**, Vieth B, Enard W, Hellmann I. *Strategies for RNA-seq differential expression analysis for closely related species*. **Unsubmitted manuscript**.

Other publications (not included in the thesis):

6. Ebinger S*, Özdemir EZ*, Ziegenhain C*, Tiedt S*, Castro Alves C*, Grunert M, Dworzak M, Lutz C, Turati VA, Enver T, Horny HP, Sotlar K, **Parekh S**, Spiekermann K, Hiddemann W, Schepers A, Polzer B, Kirsch S, Hoffmann M, Knapp B, Hasenauer J, Pfeifer H, Panzer-Grümayer R, Enard W, Gires O, Jeremias I: *Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia. **Cancer Cell** 2016.*
7. Müller S*, Engleitner T*, Maresch R*, Zukowska M, Lange S, Konukiewitz B, Kaltenbacher T, Maximilian Zwiebel M, Öllinger R, Strong A, Yen H, Steiger K, Banerjee R, Louzada S, Fu B, Seidler B, Götzfried J, Hassan Z, Schuck K, Schönhuber N, Veltkamp C, Friedrich M, Rad L, Barenboim M, Ziegenhain C, Dovey OM, Eser S, **Parekh S**, ... , Rad R: *Evolutionary trajectories and KRAS gene dosage define pancreatic cancer phenotypes. **Nature** 2017.*
8. Ziegenhain C*, Vieth B*, **Parekh S***, Hellmann I, Enard W: *Quantitative single-cell transcriptomics. **Review; Briefings in Functional Genomics; In revision.***
9. Bagnoli JW*, Ziegenhain C*, Janjic A*, Wange LE, Vieth B, **Parekh S**, Geuder J, Hellmann I, Enard W: *mcSCRB-seq: sensitive and powerful single-cell RNA sequencing. **Nature Communication; In revision.***

Declaration of contribution as co-author

The impact of amplification on differential expression analyses by RNA-seq

I had the idea to this study and planned it with Christoph Ziegenhain and Ines Hellmann. Christoph Ziegenhain prepared the RNA-seq libraries used in the publication. I designed the analysis strategy, processed and analysed the data. Ines Hellmann and Beate Vieth assisted in simulations. The manuscript was written by Ines Hellmann, Wolfgang Enard and me.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the above contributions of Swati Parekh to this publication.

Wolfgang Enard

Comparative Analysis of Single-Cell RNA Sequencing Methods

Christoph Ziegenhain and Wolfgang Enard conceived the study. The single-cell RNA-seq protocols were established and libraries were prepared by Christoph Ziegenhain. Christoph Ziegenhain and I analysed all the data. The power simulation framework was developed by Beate Vieth. Ines Hellmann guided computational work. The manuscript was written by Wolfgang Enard and Christoph Ziegenhain with valuable input from Ines Hellmann and Björn Reinius.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the above contributions of Swati Parekh to this publication.

Wolfgang Enard

powsimR: Power analysis for bulk and single cell RNA-seq experiments

Beate Vieth and Ines Hellmann conceived the study. The idea to this work emerged from power simulations for “The impact of amplification on differential expression analyses by RNA-seq” and “Comparative Analysis of Single-Cell RNA Sequencing Methods”. Beate Vieth developed and programmed *powsimR*. I helped in data processing and testing the program with Christoph Ziegenhain and evaluated its performance relative to empirical bulk and scRNA-seq data. Beate Vieth, Ines Hellmann and Wolfgang Enard wrote the manuscript.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the above contributions of Swati Parekh to this publication.

Wolfgang Enard

zUMIs: A fast and flexible pipeline to process RNA sequencing data with UMIs

Christoph Ziegenhain and I conceived the idea and implemented individual modules. I designed the backbone of the pipeline and wrapped all the modules in one. Beate Vieth tested the code and performed power simulations to evaluate intron mappings. Christoph Ziegenhain, Ines Hellmann, Wolfgang Enard and I wrote the manuscript.

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

Swati Parekh

Christoph Ziegenhain

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the above contributions of Swati Parekh to this publication.

Wolfgang Enard

Strategies for RNA-seq differential expression analysis for closely related species

This study was conceived by Ines Hellmann and me. I designed and performed all the simulations. Ines Hellmann and I developed divergence estimation framework. I analysed all the data and Beate Vieth assisted in statistical methods. The manuscript was written by Ines Hellmann, Wolfgang Enard and me.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the above contributions of Swati Parekh to this publication.

Wolfgang Enard

Aim of the thesis

High-throughput sequencing technology has enabled us to explore the whole transcriptome of biological systems even at single cell resolution by producing billions of short reads. There are many applications of RNA-sequencing (RNA-seq), such as expression profiling to compare genetically modified systems, studying evolution of traits across species or understanding disease mechanisms in an individual (Shendure and Lieberman Aiden 2012). The aim of this work is to develop and optimise computational strategies to minimize unwanted technical noise and thus improve relative quantification using bulk or single-cell RNA-seq within or across species.

Summary

The generation of cDNA libraries from RNA transcripts and its subsequent sequencing using high throughput sequencers is called RNA-seq. While it has become the dominant technology to quantify expression levels genome-wide, its experimental and computational methods are still rapidly evolving. To make optimal use of this data, it is necessary to quantify sources of technical noise, benchmark different experimental and computational methods and generate new tools where necessary.

In the first study of this thesis, we investigated the technical noise introduced by PCR amplification of cDNA during library generation when handling small amounts of starting material. To address this question, we analysed datasets generated from Universal Human Reference RNA (UHRR) using different library preparation protocols and a publicly available single-cell dataset. We find that read duplicates emerging during amplification can not be correctly identified computationally. However, if 4-10bp random barcodes (Unique Molecular Identifiers - UMIs) are used to tag each cDNA molecule before amplification, it enables correct identification of duplicate reads. Additionally, early pooling of samples before amplification using sample barcoding helps to overcome variable amplification rates across samples. Using simulations I show that the power to detect differential expression is negatively correlated with the number of PCR cycles used. Furthermore, pooling samples prior to PCR, increases the power while controlling the False Discovery Rate (FDR).

We confirm this finding in the second study of this thesis where we compared six scRNA-seq library preparation methods for their sensitivity to detect genes, accuracy to estimate expression levels, precision of measurements, power to detect differential expression and cost efficiency. The power simulation framework designed to evaluate the impact of amplification and comparison of various scRNA-seq protocols for detecting

differential expression was published in a third study describing our software package, *powsimR*. When benchmarking different scRNA-seq protocols, we found that available UMI- processing pipelines were lacking desirable features for the analysis of scRNA-seq experiments implemented in *zUMIs* include: adaptive downsampling, automatically identifying intact cell barcodes and additionally also counts and collapses intron-mapping reads. Therefore, in the fourth study of this thesis- I developed *zUMIs*, a fast and flexible pipeline that incorporates all the above features.

In the fifth study, we investigated on a set of challenges that arise when using RNA-seq to study expression profiling across diverged species. To this end, I developed a simulation framework to model mammalian whole genome sequence evolution together with gene expression profiles. Using simulations we could show that a common well resolved genomic reference can be used 1) to compare gene expression changes between closely related species 2) to compare expression changes between conditions within non-reference species and 3) to compare expression changes between species relative to conditions. Moreover, by simulating expression profiles in different sequencing layouts, we could show that longer reads increase sensitivity and that a single-end (SE) sequencing layout is sufficient for quantitative gene expression studies. This study allows to improve gene expression quantification among species and also shows that species-specific genome annotations become crucial at divergence levels above ~10%.

All in all, we have shown that, for quantitative RNA-seq, the utility of noise reduction due to UMIs is indeed a function of the amount of amplification and is especially important for low input applications such as scRNA-seq. Moreover, using simulations, we could measure the impact of varying genomic resources quality on relative quantification between diverged species. In conclusion, I used simulations to identify and correct for possible sources of bias through computational methods to improve gene expression quantification using RNA-seq.

Introduction

Gene expression profiling

The instructions for a biological system are ultimately contained in DNA. The cell is the basic unit of any living system and cells within an organism can have different functions, although they contain the same DNA. This is achieved because cells with different functions transcribe different parts of their DNA into RNA, which in turn serves as a template for proteins. This process of information transfer from DNA to RNA to protein is known as the central dogma of molecular biology (Crick 1970; Strasser 2006) and puts transcription and the abundance of RNAs at the root of cellular processes. Hence, determining the abundance of all transcripts in a cell or group of cells is highly informative. Such measurements of gene expression profiles have provided numerous insights into biological systems ranging from the evolution of phenotypic traits (Sousa et al. 2017; Brawand et al. 2011), cellular identity (GTEx Consortium et al. 2017; Lonsdale et al. 2013), functioning of various tissues and organs in disease condition (Emilsson et al. 2008; Delgado and León 2006), and population scale studies to understand variation in expression profiles (Stranger et al. 2007), to name just a few recent examples.

Various methods have been developed and evolved for mRNA quantification over time. About 40 years ago, quantification of transcripts was done using northern blotting, where RNA molecules are size separated using gel electrophoresis, transferred onto a nylon membrane and detected by hybridisation to radioactively labelled complementary probes specific to the gene of interest (Alwine, Kemp, and Stark 1977). Later, a more sensitive method called ribonuclease protection assay (RPA) came into light (Azrolan and Breslow 1990; Sambrook and Russell 2001). In RPA, the hybridisation of mRNA with the probe

takes place in solution; after enzymatic degradation of unspecific hybrids, the remaining product is electrophoresed on a polyacrylamide gel and visualised by phosphorimaging or radiography. Gradually, two polymerase chain reaction (PCR) based techniques were employed to characterize and quantify mRNA levels, namely reverse transcription polymerase chain reaction (RT-PCR) (Chelly et al. 1988; Rappolee et al. 1988) and quantitative or real-time polymerase chain reaction (qPCR) (Becker-André and Hahlbrock 1989; Weis et al. 1992; A. M. Wang, Doyle, and Mark 1989). Quantification by qPCR is done by measuring signals of fluorescent dye incorporated into a complementary DNA (cDNA) molecule reverse transcribed from mRNA during amplification in real time. However, the above mentioned techniques are time consuming and limited only to selected genes.

The idea of randomly sequencing cDNA clones to discover genes was first coined in 1982 (Putney, Herlihy, and Schimmel 1983; Sutcliffe et al. 1982) and such sequences were later called expressed sequence tags (EST) (Adams et al. 1991); however, it was not a quantitative measure of expression levels. In 1995, the first sequencing-based quantitative gene expression profiling was attempted using serial analysis of gene expression (SAGE) (Velculescu et al. 1995), where restriction enzyme digested 11bp short tags of cDNA are sanger sequenced. Many variants of SAGE have been developed to overcome the ambiguity issue due to short tags (Saha et al. 2002; Matsumura et al. 2005; Gowda et al. 2004). However, this technique is dependent on the presence of restriction enzyme sites preventing global profiling of whole transcriptomes. Eventually, the development moved towards on array hybridisation of the entire transcriptome using predetermined probes called “DNA microarrays” (Schena et al. 1995). Microarrays are solid surface chips with microscopic spots containing oligonucleotide probes that are designed complementary to cDNA sequences of known genes. A set of probes for each gene are designed taking into account the properties for optimum hybridisation like GC content, melting temperature, self hybridisation, and cross hybridisation with other targets in the genome (Liu, Bebu, and Li

2010). After the probes are immobilised on the array surface, fluorescently labelled cDNA of samples are added onto the array to hybridise with their complementary probes. A washing step is performed to remove non-specifically bound cDNA and the fluorescent signal is measured. This signal corresponds to the amount of cDNA molecules for each gene enabling relative quantification of gene expression across samples (Duggan et al. 1999; Schulze and Downward 2001). With the increasing popularity of microarray technologies, there was a pressing need for more stringent quality controls. Thus, a MicroArray Quality Control (MAQC) was initiated to address concerns of technical performance by assessing various methods on the same reference RNA performed in different labs (MAQC Consortium et al. 2006; Shippy et al. 2006). A relatively high level of inter and intra-platform concordance in differential gene expression measures were observed. The second phase of MAQC project (MAQC-II) (Shi et al. 2010) focused on generating and benchmarking predictive models to reliably anticipate the clinically relevant outcome from patient data. The conclusions from MAQC-II project assumed that integrating other types of biological data at the DNA, micro-RNA and protein levels would increase the prediction accuracy of clinical data. Moreover, MAQC data sets provided a platform for benchmarking newly developed protocols and analysis pipelines (Kerr 2007; Bullard et al. 2010). Nevertheless, using microarrays for gene expression profiling comes with certain disadvantages: 1) cross hybridisation of probes with multiple targets; 2) the probes are designed based on predetermined sequences, limiting the usability of microarrays only to the species with well resolved sequence and gene models; 3) with the fluorescence intensity-based measures, genes with low expression levels are affected by the presence of background noise and high expression levels are affected by signal saturation. Sequencing-based gene expression profiling using RNA sequencing (RNA-seq) (Mortazavi et al. 2008) overcomes these limitations by measuring the expression levels via digital counting of sequenced reads per gene. With the increased sensitivity, specificity and the ability to detect novel genes and

isoforms, RNA-seq provides practical solutions for a broad range of experimental designs, making it the state-of-the-art method for gene expression studies.

Scope of sequencing

There have been phenomenal advancements in the field of sequencing technologies since the first draft of the human genome sequence (Lander et al. 2001; Venter et al. 2001). Historically, studying a biological system by decoding the nucleotide base order began with the chain termination method proposed by Frederick Sanger (Sanger, Nicklen, and Coulson 1977). Gradually, technological variations in this method brought about many automated DNA sequencers leading to the birth of the first commercial first-generation DNA sequencer (Hunkapiller et al. 1991). Methods like PCR and recombinant DNA technology slowly led to further development in the sequencing technology leading to parallel sequencing of hundreds of samples by ABI PRISM, which aided in the completion of the first draft of human genome sequence (C.-Y. Chen 2014; Ansorge 2009). The 454 GS 20 was the first machine that brought massively parallel sequencing easily available to users in 2005 (Margulies et al. 2005) which was later upgraded to 454 GS FLX with better per base sequencing quality (Voelkerding, Dames, and Durtschi 2009). This was a major breakthrough in the field of High-Throughput Sequencing (HTS).

Ever since the National Human Genome Research Institute (NHGRI) started a race to achieve a 1000\$ human genome (Schloss 2008), a boom in massively parallel sequencing technology development emerged with different read length, throughput and base quality (Kircher and Kelso 2010). These technologies were then overruled by Illumina sequencing with its remarkably reduced cost, higher throughput and increasing speed with rapid improvements in sequencing quality (Zimmerman 2014). Illumina sequencing works on the principle of “sequencing by synthesis” using cyclic reversible termination (Bentley et al. 2008). In this technology, the sequencing templates are immobilized on the surface of a

flow-cell by incorporating adapters into fragmented DNA. These templates are then amplified into several copies on the flow-cell by “bridge amplification” (Fedurco et al. 2006) followed by adding all four nucleotide bases. The bases are modified for a single fluorophore and a chain terminator incorporating only one base at a time. The fluorescence signal is recorded and base-calling is done from the images generated at every cycle (Kircher, Stenzel, and Kelso 2009).

Apart from costs, features like sequencing layout, depth and read length should be taken into consideration for different applications (Metzker 2010; Reuter, Spacek, and Snyder 2015; Kircher and Kelso 2010). DNA fragments can be sequenced from one end (single-end) or both (paired-end). Choosing the ideal sequencing layout depends on the goal of the study. Depending on the machine used, up to 600bp (300bp paired-end) sequencing reads can be generated using Illumina technology (Goodwin, McPherson, and McCombie 2016). Paired-end sequencing layout is preferable for detecting genomic rearrangements, de novo transcript detection and differential isoform expression analysis (Garber et al. 2011; Katz et al. 2010). For quantitative gene expression studies of model organisms, cost efficient single-end sequencing is sufficient (Conesa et al. 2016). On the other hand, longer reads improve mappability, novel splice-site detection (Łabaj et al. 2011) and the comparative analysis of diverged species with poorly resolved gene models.

Rapid advancements in the technologies bring new challenges like sources of noise, sequencing quality, and the development of computational methods to tackle application-related issues (Kircher, Stenzel, and Kelso 2009; Stegle, Teichmann, and Marioni 2015). In this work, we focus on investigating potential sources of noise and discuss about possible optimisation for gene expression profiling using RNA-seq generated using the Illumina platform.

RNA sequencing

RNA-seq enables us to characterize and quantify the whole transcriptome of a system. Apart from mRNA quantification, RNA-seq can be used for novel gene discovery, fusion transcript detection, detecting alternative splicing events, variant discovery and allele specific expression (Z. Wang, Gerstein, and Snyder 2009; Ozsolak and Milos 2011). There are various library preparation methods established with distinct features (van Dijk, Jaszczyszyn, and Thermes 2014). The major steps involved in library preparation are RNA extraction, reverse transcription to cDNA, fragmentation, adapter ligation and sequencing (Figure 1).

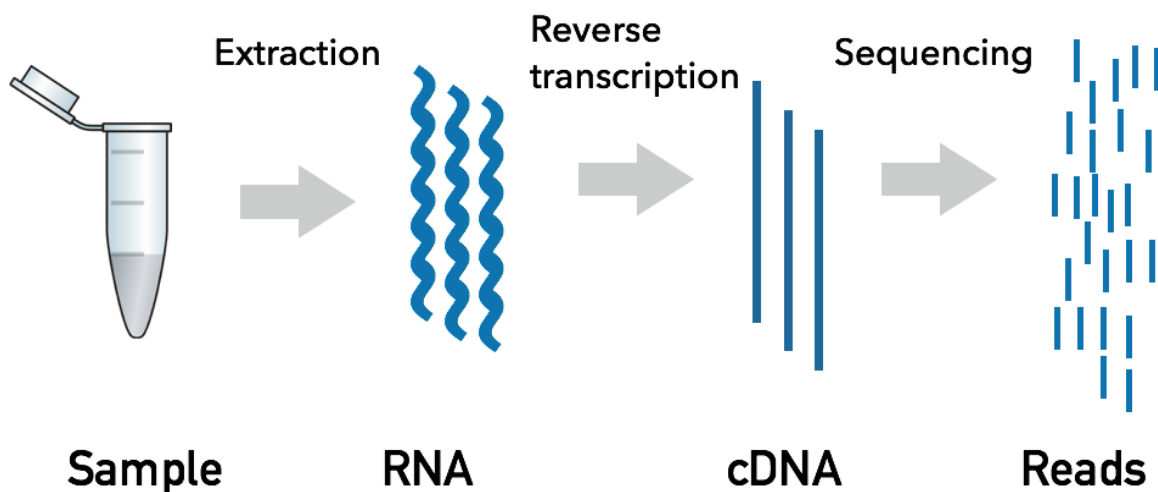


Figure 1: Basic steps of RNA-sequencing. Depicted here are the basic steps from sample to short read sequencing, going from left to right.

Different protocols feature distinct ways of fragmentation, required amount of starting material, number of PCR cycles, strand specificity, mRNA enrichment and sample pooling (Levin et al. 2010; Ziegenhain et al. 2017; Parekh et al. 2016). The first step of any RNA-seq protocol is to extract RNA. The major fraction (>80%) of a cells' RNA is ribosomal (O'Neil,

Glowatz, and Schlumpberger 2013); hence, in order to be able to quantify mRNA levels an enrichment step is necessary. Therefore, most of the protocols have rRNA depletion using magnetic beads or nuclease, or mRNA enrichment using Oligo-dT particles to pull down poly-A+ mRNAs (Choy et al. 2015). Depending on the protocol, RNA is fragmented by either heat or chemical hydrolysis as in TruSeq protocol (Mortazavi et al. 2008), or enzymatic digestion or sonication after being reverse transcribed into cDNA (Adey et al. 2010; Picelli et al. 2013). In some protocols, the next step is adapter ligation and adding sample specific barcodes to multiplex them into one tube to facilitate sequencing multiple samples on the same run and reduce batch effects (Marioni et al. 2008).

In RNA-seq experiments, known synthetic spike-in molecules (typically ERCCs) are also added (External RNA Controls Consortium 2005; Jiang et al. 2011) during the library preparation. These spike-in molecules have a wide range of expression levels with varying length and GC content. They provide a ground truth to calculate standard curve to measure the accuracy of quantification and technical biases during library preparation. However, these spike-in molecules are often criticised for being influenced by biological signals and thus not suggested to use for normalisation across samples (Risso et al. 2014; Tung et al. 2017).

The final libraries are amplified before pooling for sequencing. The required number of PCR cycles varies for each protocol depending on the amount of starting material and requirements of the protocol (Parekh et al. 2016; van Dijk, Jaszczyszyn, and Thermes 2014). These libraries are loaded onto the sequencer and a massive amount of short reads are generated. The choice of sequencing layout and read length typically depends on the application in question (Conesa et al. 2016; X. Zhou and Rokas 2014; Mortazavi et al. 2008).

RNA-seq data processing

After sequencing, the basic data processing for any RNA-seq experiment involves demultiplexing, mapping and counting. Typically, when the genomic resources are available, short reads are mapped to the genome using a splice aware aligner. Mapped reads are then converted into a count matrix with samples in columns and genes as rows (Figure 2).

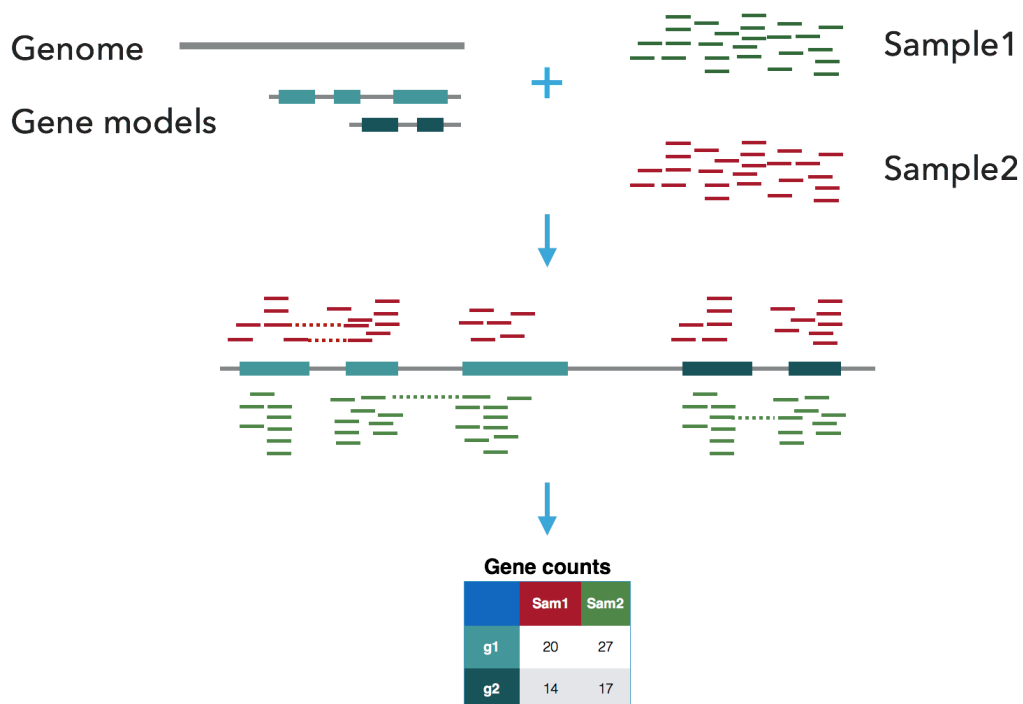


Figure 2: RNA-seq data processing. The genome is shown as a straight grey line and exons on the gene are depicted as coloured boxes. Reads are coloured in red and green to distinguish two samples whereas two genes are the dark and light shades of seagreen color. The dashed line in the read mapping shows reads spanning over intron-exon junctions. After mapping, reads assigned to features are counted per gene per sample and tabulated as a gene count table.

Demultiplexing

In a RNA-seq experiment, multiple libraries are pooled together using sample specific barcodes before sequencing (Craig et al. 2008; Meyer et al. 2007). Pooling as many samples as possible in the same sequencing run not only reduces experimental costs but also prevents sequencing bias (Auer and Doerge 2010). During the library preparation for Illumina sequencing technology, sample barcodes are added in one of the library adapters and sequenced as index reads usually 6-8bp long. With the rapid increase in throughput, Illumina introduced dual barcoding by incorporating barcodes in both library adapters. Depending on the sequence length, up to 27 different barcodes were possible with single indexing approach, whereas with the double indexing method, up to 384 different samples could be multiplexed (Kircher, Sawyer, and Meyer 2012). Thus, RNA-seq data analysis begins with demultiplexing the raw reads from pooled sequences based on their sample barcodes sequenced as index reads (Renaud et al. 2015; Galanti, Shasha, and Gunsalus 2017).

Quality Control (QC)

Various Quality Control (QC) metrics are applied at every stage of data analysis. The QC of raw reads informs about the issues with library preparation and sequencing. To this end the per base sequence quality, GC content, the presence of undesirable sequences (primers or adapters) and overrepresentation of certain fragments or contamination are evaluated. Some of the QC metrics are observed after mapping. For a typical bulk RNA-seq experiment, more than 65% reads are uniquely assigned to exonic regions for a species with a well resolved genome assembly and gene models. Uniformity of read distribution across gene body is another measure of quality check for full length RNA-seq libraries. Several tools are available for deriving these QC metrics: FastQC (Bioinformatics 2011), kraken (Davis et al. 2013), Qualimap (García-Alcalde et al. 2012; Okonechnikov, Conesa, and

García-Alcalde 2016), NGS-QC toolkit (Patel and Jain 2012) and RNA-SeQC (DeLuca et al. 2012).

Mapping and Quantification

When it comes to mapping of RNA-seq reads, the researcher is confronted with many choices that will influence all downstream analyses, starting with a decision about which gene/transcript annotation and which mapper to use, ending with the fine-tuning of mapping parameters (Baruzzo et al. 2016; Fonseca et al. 2012; Conesa et al. 2016; Engström et al. 2013).

First, there appears to be two major mapping strategies: 1) mapping to genome and 2) mapping to transcriptome. Second, choosing which annotations to use: 1) focus on well-known genes using the RefSeq (O'Leary et al. 2016) or HAVANA (Hancock, Hancock, and Zvelebil 2004) curated annotation or 2) be inclusive and accept some level of incorrect annotation and thus always use the most recent ENSEMBL (Zerbino et al. 2017) release and map. The choice of reference and annotations strongly impacts the downstream analysis in RNA-seq (Zhao and Zhang 2015; Nellore et al. 2016; Garber et al. 2011; Mortazavi et al. 2008). Generally, genome sequences are better resolved compared to gene models. In addition, if the reads from unannotated regions of the genome are forced to map to the reference transcriptome, it can lead to spurious mapping (Zhao 2014). Moreover, if the goal is to identify novel splice junctions and fusion transcripts, mapping to a reference genome is mandatory.

On average, 12% of the reads mapped to a genome span exon-intron junctions. Hence, it is important to use a splice-aware mapper for RNA-seq reads (Engström et al. 2013) such as STAR (Dobin et al. 2013) or HISAT (Kim, Langmead, and Salzberg 2015). To obtain the optimal alignment, it is necessary to adjust mapping parameters based on the nature of the input data. For instance: 1) allowed number of mismatches per read depends on quality of

sequencing; 2) in the case of paired-end sequencing, the range of insert size between two reads of a fragment is important to set; 3) intron size should be adjusted based on the species sequenced; 4) the number of bases a read is allowed to span over intron-exon junction site should be adjusted according to the read length. In the case of reference-based RNA-seq quantification, trimming and filtering of low quality reads and adapter contamination is not necessary, because low quality reads are unlikely to map at all (Dobin et al. 2013; Baruzzo et al. 2016; Engström et al. 2013).

With the increasing throughput, speed has become an important criterion for data processing. Mapping free quantification known as “pseudo-alignment” methods have come into light with more appealing run-time compared to full alignment methods (Patro et al. 2017; Bray et al. 2016; Srivastava et al. 2016). However, these methods utilise transcriptome as a reference, which poses a problem for species where gene models are not fully resolved and also for novel transcript detection. Table 1 lists important features of different mapping and quantification tools frequently used for reference based RNA-seq analysis.

Tool	Alignment	Quantification	Reference type	Multimapping	Input format
STAR	Full alignment	Yes	Genome Transcriptome	UD,A,N,B	fastq SAM BAM
HiSAT	Full alignment	No	Genome Transcriptome	UD,A,B	fastq
GSNAP	Full alignment	No	Genome Transcriptome	UD,A,N,B	fastq
Mapsplice	Full alignment	No	Genome Transcriptome	B	fastq
RapMap	Pseudo-alignment	Yes	Transcriptome	H (<=200)	fastq
Kallisto	Pseudo-alignment	Yes	Transcriptome	A	fastq
Salmon	Pseudo-alignment pre-aligned BAM	Yes	Transcriptome	A	Fastq SAM BAM
RSEM	Full alignment by Bowtie Bowtie2 STAR	Yes	Transcriptome	A	SAM BAM fastq

Express	Pre-aligned BAM	Yes	Transcriptome	A	SAM BAM
featureCounts	Pre-aligned BAM	Yes	Genome	UD,N,A	SAM BAM
HTSeq-count	Pre-aligned BAM	Yes	Genome	UD,N,A	SAM BAM
ESAT	Pre-aligned BAM	Yes	Genome	N,A,ES	SAM BAM

Table 1: Utilities of commonly used RNA-seq mapping and quantification tools. The tools with light blue background are primarily used for mapping, light green background are used for quantification and white background perform mapping free quantification. The type of alignment algorithm a tool supports is given in “Alignment” column. The “Read assignment” column represents if the tool uses Expectation Maximization (EM) algorithm to resolve reads assigned to more than one isoforms of a gene to execute isoform level quantification. The data in column “Multi-mapping” shows how the alignments are reported for multi-mapping reads: UD- user defined, A- all, N- none (only uniquely mapped reads reported), B- randomly chosen one hit, ES- a read is assigned unique if one of the best hits is within a transcript, H- hard cutoff for number of multi-mapping hits to report.

After mapping, expression estimates for each gene are calculated as the sum of reads mapped to the exons using quantification tools such as featureCounts (Liao, Smyth, and Shi 2014) or HT-Seq (Anders, Pyl, and Huber 2014). By default, these tools count every read with at least 1 base overlap with an exon, while reads mapped to the positions where two genes from different strands overlap are not counted. The reads mapping to more than one loci are called multi-mapping reads. Different quantification tools employ different strategies to deal with multi-mapping reads: 1) discard all the multi-mapping reads, 2) randomly choose one loci, or 3) distribute equal weight among all the loci (Table 1). Dealing with the reads mapping to multiple locations on the reference (multi-mapping reads) is an open question of RNA-seq data analysis. Various strategies have been implemented to “handle” multi-mapping reads, but a comprehensive assessment of their impact on quantification remains unresolved.

Amplification noise

To accurately measure absolute or relative expression of genes, we require unbiased quantification of their expression levels. One of the major concerns in RNA-seq is the over amplification of certain transcript molecules that do not contribute to the actual expression level estimation of genes. Amplification of transcript molecules generates duplicated sequencing reads. Read duplicates in RNA-seq can be classified into three types: 1) read duplicates arising from different RNA molecules fragmented at the same site, known as “natural duplicates”, 2) PCR duplicates that originated from PCR amplification, and 3) optical duplicates generated by the same cluster on a flow cell misread as a separate cluster by the software (van Dijk, Jaszczyszyn, and Thermes 2014; Kozarewa et al. 2009; Mamanova et al. 2010; X. Zhou and Rokas 2014). Studies have shown the effect of PCR duplicates in the case of variant detection and ChIP-seq analysis and methods have been proposed to computationally correct PCR duplicates (Baumann and Doerge 2014; Mezlini et al. 2013; Ebbert et al. 2016; W. Zhou et al. 2014). However, the impact of presence of PCR duplicates had not been thoroughly studied in the context of quantification and differential expression by RNA-seq. If PCR amplification of fragments were uniform, all fragments should be amplified with the same efficiency. In reality, certain fragments are over-amplified leading to non-uniformity of reads along the transcript (Figure 3). Given that PCR is exponential, such variability in amplification rate propagates with more PCR cycles and can ultimately distort expression profiles. Especially in single-cell RNA-seq experiments, the low starting amount of mRNA is unavoidable and it leads to an increasing number of PCR cycles during library preparation. Thus, it has been suggested to carefully optimize the number of PCR cycles during library preparation (Picelli et al. 2014; Kolodziejczyk, Kim, Svensson, et al.

2015).

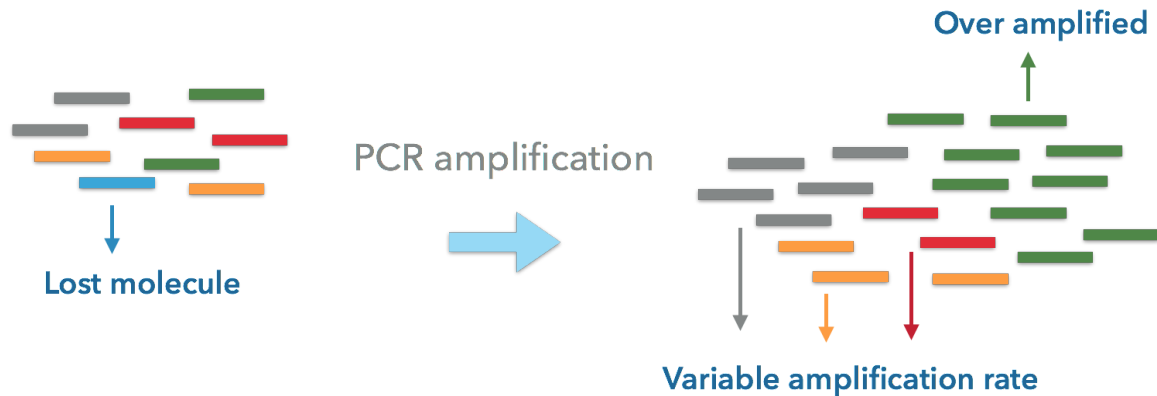


Figure 3: Schematic of variable PCR amplification rate. Depicted here are different cDNA molecules in different colors (grey, green, red, orange and blue). Certain fragments may amplify very efficiently (green), while others may be underrepresented (red) or lost (blue).

Recently, molecular tagging technologies have enabled us to track real PCR duplicates experimentally (Islam et al. 2014). Unique molecular identifiers (UMI) are random 4-10bp oligos incorporated to each cDNA molecule during reverse transcription (Macosko et al. 2015; Hashimshony et al. 2012; Bagnoli et al. 2017; Zilionis et al. 2017). Since each cDNA molecule most likely has a sequence of UMI, during amplification, several copies of each cDNA molecule are generated with the same UMI sequence. Thus, after sequencing, each initial cDNA molecule can be counted by collapsing the same UMIs per gene in each cell (Figure 4).

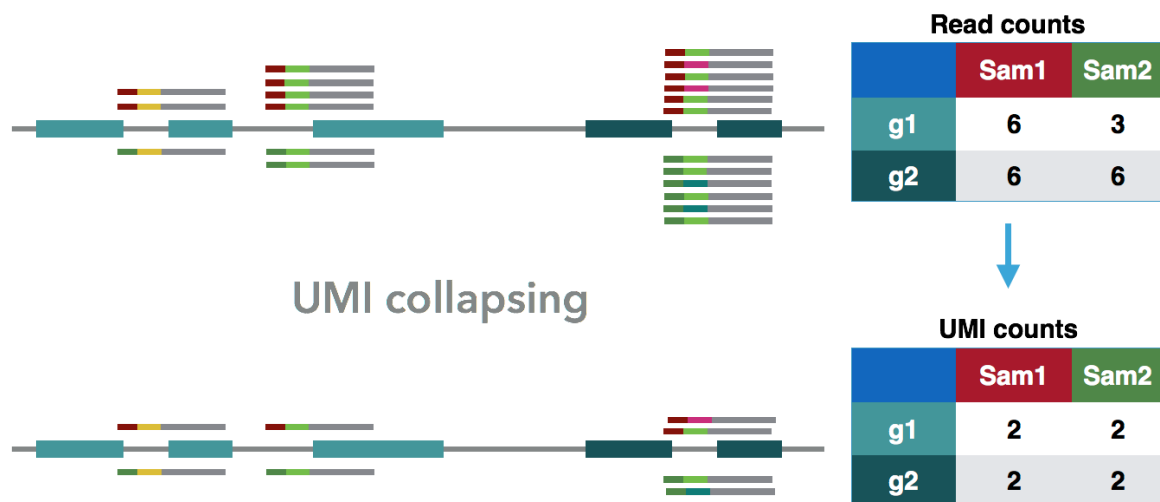


Figure 4: Unique Molecular Identifiers. Two genes are shown here with their exonic parts as blocks (dark and light cyan) on a genome (grey line). Short piled fragments are the cDNA reads (grey) starting with cell barcodes (magenta and dark green) and molecular barcodes (UMI) (light green and yellow). Upper panel shows all the assigned cDNA reads to genes in two samples (above and below the genome). Right side of the genome depiction is a count table with samples as columns and genes as rows with the number of reads assigned to per gene per sample. Lower panel shows only uniquely retained reads based on their UMI sequences and a corresponding count table.

The major challenge is to distinguish them from natural duplicates generated due to fragmentation preference at similar sites to prevent overcorrection of reads (Adey et al. 2010; Baumann and Doerge 2014). A comprehensive analysis of read duplication introduced by preferential fragmentation and/or PCR amplification is needed to understand the impact of amplification on differential expression analysis.

Single-cell RNA-sequencing (scRNA-seq)

By measuring gene expression levels in the basic unit of biology, the cell, single-cell RNA-seq allows to see information pertaining to every cell previously hidden in averages of bulk RNA-seq (Wills et al. 2013). This higher resolution has been leveraged to understand allele-specific transcription (Deng et al. 2014; Reinius and Sandberg 2015) and gene

expression heterogeneity within tissues (Kolodziejczyk, Kim, Tsang, et al. 2015; Martinez-Jimenez et al. 2017). Furthermore, it has become possible to discover new cell types (Trapnell 2015) in various tissues, such as blood (Björklund et al. 2016; Villani et al. 2017), spleen (Jaitin et al. 2014), brain (Poulin et al. 2016; Zeisel et al. 2015; La Manno et al. 2016; Gokce et al. 2016; Tasic et al. 2016) and others (Grün et al. 2015; Muraro et al. 2016; Macosko et al. 2015). Fueled by this new technology, the Human Cell Atlas initiative (Regev et al. 2017) has begun with the goal of achieving a reference map for all human cell types. The basic steps in a scRNA-seq experiment involve tissue dissociation, cell isolation and lysis, reverse transcription of RNA to cDNA, amplification and preparation of libraries of cDNA fragments. Despite the fact that basic steps in scRNA-seq library preparation methods remain the same, plethora of scRNA-seq protocols have been published in recent years for specific applications (Kolodziejczyk, Kim, Svensson, et al. 2015). Most methods use well plates or microfluidic droplets to encapsulate cells. An ideal scRNA-seq method features 1) sensitivity to detect transcript molecules, 2) accuracy to measure expression levels, 3) precision of measured expression levels, 4) cost efficiency. Two major studies have shown comprehensive assessment of these protocols (Ziegenhain et al. 2017; Svensson et al. 2017). Based on this evaluation, various variants of existing protocols are being developed to optimise the performance of scRNA-seq (Bagnoli et al. 2017).

scRNA-seq also comes with new computational challenges (Stegle, Teichmann, and Marioni 2015). First, with tens of thousands of cells being sequenced, minimizing cross cell contamination has become more demanding. In most of the scRNA-seq protocols, cell specific barcodes are added at the reverse transcription step to be able to pool more cells in a single reaction (Jaitin et al. 2014; Soumillon et al. 2014; Hashimshony et al. 2016; Macosko et al. 2015; Klein et al. 2015; Picelli et al. 2014). There are various sources of errors coming from PCR, sequencing or spill-over from other samples. For droplet-based methods like Drop-seq and DroNc-seq, cell barcodes are not known a priori but they are determined from

the sequenced reads with the notion that intact cells have more reads compared to broken or dead cells (Macosko et al. 2015; Zilionis et al. 2017; Habib et al. 2017). Filtering of low quality cells and reads prior to analysis is essential to avoid misinterpretations (Ilicic et al. 2016; Bacher and Kendzioriski 2016; Guo et al. 2015; Finak et al. 2015; McCarthy et al. 2017).

The variance in quality amongst cells of a single cell RNA-seq experiment is much larger than in bulk data and it is highly recommended to filter non-informative cells prior to downstream analysis (Ilicic et al. 2016; Bacher and Kendzioriski 2016; Guo et al. 2015; Finak et al. 2015; McCarthy et al. 2017). Reads with the same cell barcode that exhibit a high abundance of adapters, polyAs, overrepresented sequences and low quality reads may indicate dead cell, debris, or low or degraded RNA content, and they should be marked for removal from the downstream analysis (Ilicic et al. 2016). With the accelerated increase in throughput and methods where visual inspection is not possible, computational methods to identify such cells are inevitable (Macosko et al. 2015; Klein et al. 2015; Zheng et al. 2017). Based on mapping statistics, we can gain a more detailed view on the problems that have occurred during the processing of individual cells. Statistically the most sound method is to examine the distribution of the maximal pairwise correlation coefficient of counts. Assuming that each cell-type occurs at least twice and breaking cells is a random process, bad cells are expected to have a lower correlation with other cells (Petropoulos et al. 2016; Ziegenhain et al. 2017).

Nowadays, quantitative scRNA-seq methods are pushed towards molecular counting where each transcript molecule is tagged with a short (4-10 bases) barcode (UMI). In most of the early pooling end-sequencing methods (Jaitin et al. 2014; Soumillon et al. 2014; Hashimshony et al. 2016; Macosko et al. 2015; Klein et al. 2015; Picelli et al. 2014), the cellular and molecular barcodes (UMI - Unique Molecular Identifier) are sequenced in the

same read. Depending on the protocol, read length of cellular and molecular barcodes vary (Table 2).

Method	UMI / barcode read	Illumina index	Cell barcode length	UMI length
SCRB-seq	read1	i7	6	10
MARS-seq	read2	i7	6	8
CEL-seq2	read1	i7	6	6
Drop-seq	read1	i7	12	8
inDrops	read2	i7	18	6
10x Genomics	read2 / i7	i5	14	10
Smart-seq2	NA	i5 & i7	-	-
Smart-seq C1	NA	i5 & i7	-	-

Table 2: Method specific barcode read information.

For the cell barcodes, it is suggested to apply hard cutoff for filtering (Macosko et al. 2015) discarding reads where the barcode contains n (default $n=1$) low quality bases (default < 30 phred). However, for molecular barcodes, Islam et al (Islam et al. 2014) proposes to remove UMIs with read counts under 1% of the mean of all UMIs at a given locus. UMI-tools (Smith, Heger, and Sudbery 2017) implements network-based adjacency and directional adjacency methods which considers both edit distance and the relative counts of similar UMIs to identify PCR/sequencing errors and group them together. Obviously, for deeper sequencing longer UMIs are necessary for higher complexity measurements to capture different transcript molecules of the same gene. The data processing of UMI-based protocols is different compared to that of full length scRNA-seq especially accounting for UMI

information during quantification (Figure 4). While existing quantification tools, such as Kallisto (Bray et al. 2016) and ESAT (Derr et al. 2016) have implemented UMI-based quantification, new pipelines such as the Drop-seq pipeline (Macosko et al. 2015), scPipe (Tian et al. 2017) and Drop-est (Petukhov et al. 2017) have implemented a generic approach to process the data from raw reads to the UMI/reads count tables. These pipelines are either specific to certain methods, published as separate modules. In summary, there is a need for a flexible pipeline accommodating the vast choice of new sequencing protocols and providing useful features for various applications.

Comparative transcriptomics across species

In 1975, King and Wilson observed in their comparison of Human and Chimpanzee that the two primates are highly similar at sequence level despite having major differences at phenotypic levels (King and Wilson 1975). It is believed that the variations in gene expression and regulation programs are more amenable to phenotypic differences among species than the sequence itself (Cáceres et al. 2003; Carroll 2005; Gilad et al. 2006; Stern and Orgogozo 2008; Romero, Ruvinsky, and Gilad 2012; King and Wilson 1975). Changes in gene expression between species are thus of potential interest in understanding the evolution of phenotypic traits. A comprehensive analysis of gene expression and regulation patterns in diverged species gives insights into several applications: 1) the evolution of gene expression in different organs (Enard et al. 2002; Brawand et al. 2011); 2) the effect of diseases or food and nutrition in different animal models (Segal et al. 2005; Sweet-Cordero et al. 2005; Rasche, Al-Hasani, and Herwig 2008; Ellis et al. 2013); 3) the process of ageing in different animals (McCarroll et al. 2004; de Magalhães, Curado, and Church 2009).

Microarray technologies have been commonly used for the comparison of gene expression dynamics across species (Uddin et al. 2004; Enard et al. 2002; Cáceres et al. 2003; Khaitovich et al. 2005, 2004; Gilad et al. 2006, 2005; Fortna et al. 2004; Khaitovich et al.

2006). However, microarrays are available only for a limited number of species. This limitation can be overcome by using available microarray from one species and restrict the expression analysis only to orthologous genes (Enard et al. 2002). This approach raises an issue of variability in affinity of probes to hybridise with their respective targets in different species. To avoid such bias due to sequence divergence, only the probes with identical sequences between species were used for the analysis (Khaitovich et al. 2005). However, with distantly diverged species, this approach limits the search space to a handful of genes to compare between species. In 2005, Gilad et al. showed that this issue can be handled by using multi-species cDNA array designed by using probes from all the species in comparison (Gilad et al. 2005). Later an approach to correct for species differences by modelling probe binding affinity came into light (Dannemann et al. 2009).

Advancements in the Next generation sequencing technologies have opened exciting opportunities and increased power to study evolution through gene expression profiling using RNA-seq (Brawand et al. 2011; Khaitovich et al. 2006; Bakken et al. 2016; Blekhan et al. 2010; Wunderlich et al. 2014). The vital part of an RNA-seq study is the good quality genome sequence and well resolved gene models (Pipes et al. 2013; Benjamin et al. 2014). In reality, for non-model organisms, the genome assembly and gene models often lack good resolution. For instance, compared to human, the gene models of non-human primates have various quality issues such as shorter 3' UTRs, poorly resolved genic features or an absence of a gene (Figure 5). Such technical issues can cause systematic bias in quantitative gene expression analysis. Generally, the comparison across diverged species has been carried out by stringent filtering of orthologous regions while mapping to the genome of origin (Brawand et al. 2011; Villar et al. 2015; Pipes et al. 2013; Warnefors and Kaessmann 2013; Zhu et al. 2014). These approaches are thus only suitable when both the genome and gene models are available for all the species in comparison. An alternative to avoid the availability and varying quality issues in genomic resources would be to generate *de novo*

transcriptome assembly for each species. Nevertheless, it comes with an additional issue of drawing significant functional information and defining comparable transcriptional units across species.



Figure 5: Annotation difference between species. The schematic here reflects how the gene models are differently resolved in other species (purple) compared to Human (green). Going from left to right we show difference in the length of 3' UTR, a missing exon or completely unannotated gene.

There is a scarcity of comprehensive assessment of methods for the usage of RNA-seq as a tool to compare species with limited reference information. With the rapid increase in RNA-seq based studies for evolutionary analysis, there is a pressing need for such assessment of different methods. Using simulations for evolution of both expression levels and sequence divergence in the primate phylogeny, we quantify the impact of varying genomic reference quality on differential expression analysis among differently diverged species.

Computational simulation

To evaluate the performance of any method, a gold standard is necessary to test hypotheses. Generally, generating a gold standard dataset is not an easy task because of the various sources of technical biases to account for, such as environmental effects, human handling, availability of resources, etc. To optimise certain methods or a partially, all the other factors are required to be stable to correctly assess the impact. Thus, one needs to devise several experimental strategies to test all possible conditions, a highly time

consuming and expensive task. Ideally, the ground truth is known and can be computationally modelled to recapitulate the properties of the data.

In the field of biology, researchers face a major question about how to get the most out of an experiment (Conesa et al. 2016). There are plenty of methods available for a wide range of applications (Hrdlickova, Toloue, and Tian 2017; Shendure and Lieberman Aiden 2012) for gene expression studies. It is crucial to systematically assess the impact of various sources of technical noise such as amplification, quality of genomic resources, sequencing layouts and computational strategy for analysis.


For instance, flux simulator (Griebel et al. 2012) can model and generate a typical RNA-seq dataset from a genomic reference and defined gene models for given error models of sequencing quality. Such simulated datasets are useful for benchmarking various mapping (Baruzzo et al. 2016) and quantification methods (Teng et al. 2016) for a range of sequencing layouts. With computational simulations performed under the empirically derived mean and dispersion of expression estimates, it is possible to determine the required sample size between conditions in different levels of biological heterogeneity to achieve optimal results (Poplawski and Binder 2017). In short, computational simulations help to gain insights into possible sources of bias and optimise the analysis strategy cost efficiently.

Results

Assessing the impact of amplification noise in RNA-seq

The impact of amplification on differential expression analyses by RNA-seq

SCIENTIFIC REPORTS



OPEN

The impact of amplification on differential expression analyses by RNA-seq

Received: 25 January 2016

Accepted: 20 April 2016

Published: 09 May 2016

Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard & Ines Hellmann

Currently, quantitative RNA-seq methods are pushed to work with increasingly small starting amounts of RNA that require amplification. However, it is unclear how much noise or bias amplification introduces and how this affects precision and accuracy of RNA quantification. To assess the effects of amplification, reads that originated from the same RNA molecule (PCR-duplicates) need to be identified. Computationally, read duplicates are defined by their mapping position, which does not distinguish PCR- from natural duplicates and hence it is unclear how to treat duplicated reads. Here, we generate and analyse RNA-seq data sets prepared using three different protocols (Smart-Seq, TruSeq and UMI-seq). We find that a large fraction of computationally identified read duplicates are not PCR duplicates and can be explained by sampling and fragmentation bias. Consequently, the computational removal of duplicates does improve neither accuracy nor precision and can actually worsen the power and the False Discovery Rate (FDR) for differential gene expression. Even when duplicates are experimentally identified by unique molecular identifiers (UMIs), power and FDR are only mildly improved. However, the pooling of samples as made possible by the early barcoding of the UMI-protocol leads to an appreciable increase in the power to detect differentially expressed genes.

High throughput RNA sequencing methods (RNA-seq) are currently replacing microarrays as the method of choice for gene expression quantification^{1–5}. For many applications RNA-seq technologies are required to become more sensitive, the goal being to detect rare transcripts in single cells. However, sensitivity, accuracy and precision of transcript quantification strongly depend on how the mRNA is converted into the cDNA that is eventually sequenced⁶. Especially when starting from low amounts of RNA, amplification is necessary to generate enough cDNA for sequencing^{7,8}. While it is known that PCR does not amplify all sequences equally well^{9–11}, PCR amplification is used in popular RNA-seq library preparation protocols such as TruSeq or Smart-Seq¹². However, it is unclear how PCR bias affects quantitative RNA-seq analyses and to what extent PCR amplification adds noise and hence reduces the precision of transcript quantification. For detecting differentially expressed genes this is even more important than accuracy because it influences the power and potentially the false discovery rate.

RNA-seq library preparation methods are designed with different goals in mind. TruSeq is a method of choice, if there is sufficient starting material, while the Smart-Seq protocol is better suited for low starting amounts^{13,14}. Furthermore, methods using UMIs and cellular barcodes have been optimized for low starting amounts and low costs, to generate RNA-seq profiles from single cells^{7,15}. To achieve these goals, the methods differ in a number of steps that will also impact the probability of read duplicates and their detection (Fig. 1). TruSeq uses heat-fragmentation of mRNA and the only amplification is the amplification of the sequencing library. Thus all PCR duplicates can be identified by their mapping positions. In contrast, in the Smart-Seq protocol full length mRNAs are reverse transcribed, pre-amplified and the amplified cDNA is then fragmented with a Tn5 transposase¹². Consequently, PCR duplicates that arise during the pre-amplification step can not be identified by their mapping positions. UMI-seq also amplifies full-length cDNA, but unique molecular identifiers (UMIs) as well as library barcodes are already introduced during reverse transcription before pre-amplification¹⁶. This early barcoding allows all samples to be pooled right after reverse transcription. The primer sequences required for the library amplification are introduced at the 3' end during reverse transcription. Thus, PCR-duplicates in UMI-seq data can always be identified via the UMI. In summary, while PCR-duplicates can be unambiguously identified in UMI-seq, for Smart-Seq and TruSeq PCR-duplicates are identified computationally as read duplicates. However,

Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians University, Großhaderner Str. 2, 82152 Martinsried, Germany. Correspondence and requests for materials should be addressed to I.H. (email: hellmann@bio.lmu.de)

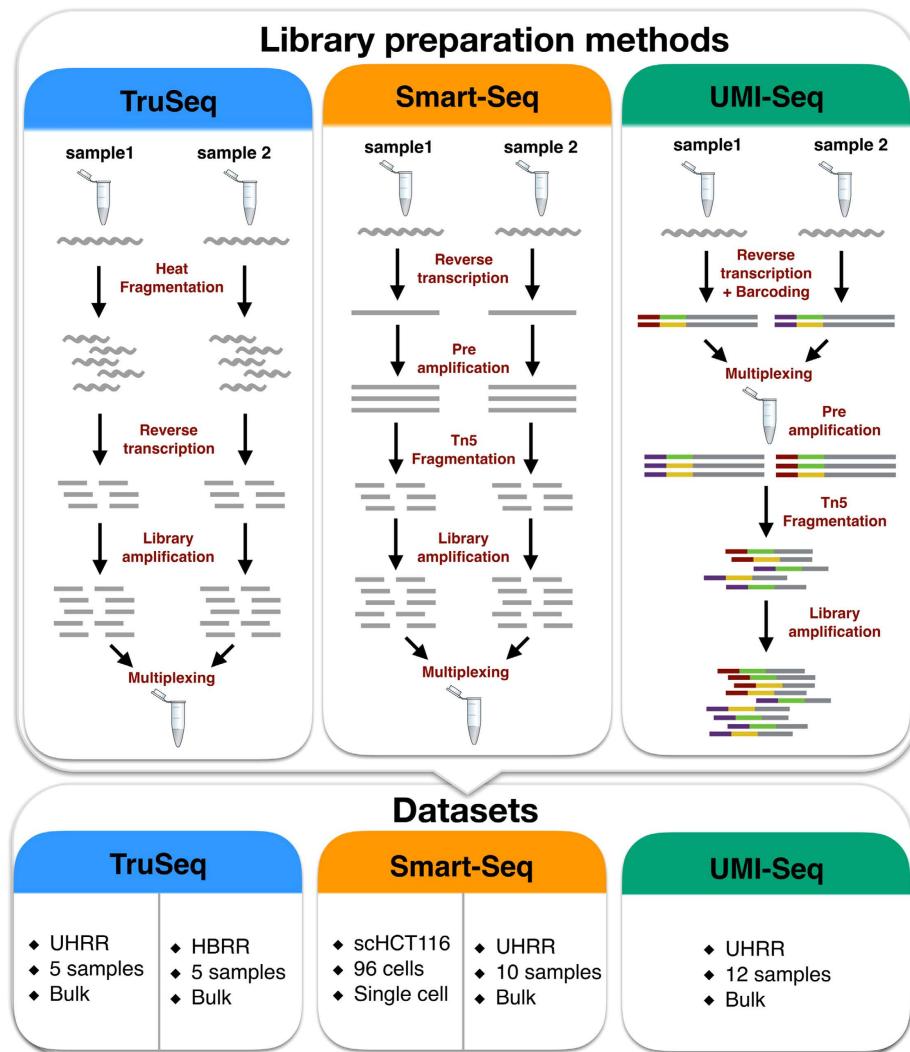


Figure 1. Schematic of library preparation protocols and datasets. The upper panel details the steps for the three sequencing library preparation methods analysed in this study. In the UMI-seq flow-chart red and purple tags represent the sample barcodes and the green and yellow tags the UMIs.

such read duplicates can also arise by sampling independent molecules. The chance that such natural duplicates, i.e. read duplicates that originated from different mRNA molecules, occur for a transcript of a given length, increases with expression levels and fragmentation bias.

That said, it is unclear whether removing read duplicates computationally improves accuracy and precision by reducing PCR bias and noise or whether it decreases accuracy and precision by removing genuine information. Here, we investigate the impact of PCR amplification on RNA-seq by analyzing datasets prepared with Smart-Seq, TruSeq and UMI-seq as well as different amounts of amplification. We investigate the source of read duplicates by analysing PCR bias and fragmentation bias, assess the accuracy using ERCCs - spike-in mRNAs of known concentrations^{17,18} - and assess precision using power simulations using PROPER¹⁹.

Results

Selection of datasets. We analyse five different datasets that represent three popular RNA-seq library preparation methods. We started with two benchmarking datasets from the literature² that sequenced five replicates of bulk mRNA using the TruSeq protocol on commercially available reference mRNAs: the Universal Human Reference RNA (UHRR; Agilent Technologies) and the Human Brain Reference RNA (HBRR, ThermoFisher Scientific). To ensure comparability, we also used UHRR aliquots to produce Smart-Seq and UMI-seq datasets in house (Table 1). However, we also wanted to include a single cell dataset, representing the most extreme and the most interesting case for low starting amounts of RNA. To this end, we chose to reanalyze the first published single cell dataset from Wu *et al.*²⁰ that sequenced the cancer cell line HCT116. The library preparation method used for the single cell data is also Smart-Seq and thus comparable to our UHRR-Smart-Seq data.

Study ID	GSE-ID	Lab	Sample size	Reads per sample (Mean \pm SD million)	Read Length	PCR cycles
scHCT116 Smart-Seq	GSE51254	Quake	96	1.8 \pm 1.1	101	21* + 12
UHRR Smart-Seq	GSE75823	Enard	10	1.5 \pm 1.1	50	10* + 12
UHRR UMI-seq	GSE75823	Enard	12	9 \pm 1	46	15* + 12
UHRR TruSeq	GSE49712	SEQC	5	125 \pm 33	101	15
HBRR TruSeq	GSE49712	SEQC	5	140 \pm 29	101	15

Table 1. Description of the datasets analysed. *preamplification PCR-cycles.

Study Name	Fraction PE-duplicates	Fraction SE-duplicates
HBRR TruSeq	0.06–0.16	0.62–0.71
scHCT116 Smart-Seq	0.013–0.59	0.064–0.94
UHRR Smart-Seq	0.081–0.18	0.36–0.47
UHRR TruSeq	0.087–0.18	0.66–0.74
UHRR UMI-seq	0.65–0.68*	

Table 2. Fraction of duplicates per sample. *Fraction of duplicates based on UMI counts.

The only drawback that we have to keep in mind for this dataset, is that it also contains true biological variation that we cannot control for, whereas the bulk datasets using the reference mRNAs should only show technical variation.

All datasets contain ERCC-spike-ins, which allows us to compare the accuracy of the quantification of RNA-levels. Furthermore, all datasets except the UHRR-UMI-seq have paired-end sequencing, which should provide more information for the computational identification of PCR duplicates.

Natural duplicates are expected to be common. The number of computationally identified paired-end read duplicates (PE-duplicates) varies between 6% and 19% for the bulk data and 1% and 59% for the single cell data. Since single-end data is commonly used for gene expression quantification, we also consider the mapping of the first read of every pair. The resulting fractions of computationally identified duplicates from single-end reads (SE-duplicates) are much higher. For the bulk data, it ranges from 36–74% and for the single cell data from 6–94% (Table 2, Fig. 2a). Surprisingly, out of the bulk datasets, the UMI-seq data show on average the highest duplicate fractions with 66% (Range: 64–68%), whereas all those duplicates are bona-fide PCR-duplicates. In the UHRR Smart-Seq data, which is the most similar dataset to the UMI-seq data, we only identified 12% PE-duplicates computationally (Fig. 2a). Although these numbers are not strictly comparable due to some differences in the library preparation (e.g. 5 more PCR-cycles for the UMI-data see Table 1 and a stronger 3' bias (Supplementary Figure S1)), it nevertheless strongly indicates that many PCR-duplicates in Smart-Seq libraries occur during pre-amplification and thus cannot be detected by computational means.

Generally, the fraction of read duplicates is expected to depend on library complexity, fragmentation method and sequencing depth. Sequencing depth is the factor that gives us the most straight-forward predictions and in the case of SE-duplicates they are by in large independent of other parameters such as the fragment size distribution. As expected, we observe a positive correlation between the number of reads that were sequenced and the fraction of SE-duplicates (Fig. 2b,c). In order to test to what extent simple sampling can explain the number of SE-duplicates, we calculate the expected fraction of SE-duplicates, given the observed number of reads per gene and the gene lengths (see Methods, Fig. 2b,c). Note that in the case of Smart-Seq this approach will only evaluate the effect of the library PCR, but be oblivious to PCR duplicates that arose during pre-amplification. We find that for TruSeq and Smart-Seq the majority of SE-duplicates are expected under this simple model of random sampling (Fig. 2b,c). For the TruSeq data our simple model underestimates the fraction of duplicates on average by 10% (8.1–13.6%), for the single cell Smart-Seq data by 19% (0.3–67%) and for the bulk Smart-Seq data by 16.6% (11.5–22.3%). Thus, irrespective of the library preparation protocol a large fraction of computationally identified SE-duplicates could easily be natural duplicates (Fig. 2b,c).

In contrast to this simple sampling expectation for SE-duplicates, fragments produced during PCR-amplification after adapter ligation, will necessarily produce fragments with the same 5' and 3' end and consequently will have identical mapping for both ends. If the sampling was shallow enough so that we would not expect to draw the same 5' end twice by chance, the 3' end position should also be identical and no reads with only one matching 5' end are expected. If same 5' ends are more frequent due to biased fragmentation, we expect a higher ratio of SE- to PE-duplicates. Thus, the relationship between PE- and SE-duplicates contains information about the relative amounts of duplicates produced by fragmentation as compared to amplification. More specifically, we expect that the fragmentation component of the PE- vs. SE-duplicates should be captured by a quadratic fit with an intercept of zero (Fig. 3).

The only dataset for which the quadratic term is not significant is the UHRR-TruSeq dataset. This could be seen as an indication of a higher proportion of PCR-duplicates, but it is more likely due to the low sample size of only 5 replicates. More importantly, the quadratic term is significant and positive for the HBRR TruSeq, the UHRR Smart-Seq and the scHCT116 datasets, supporting the notion that at least for those datasets library PCR

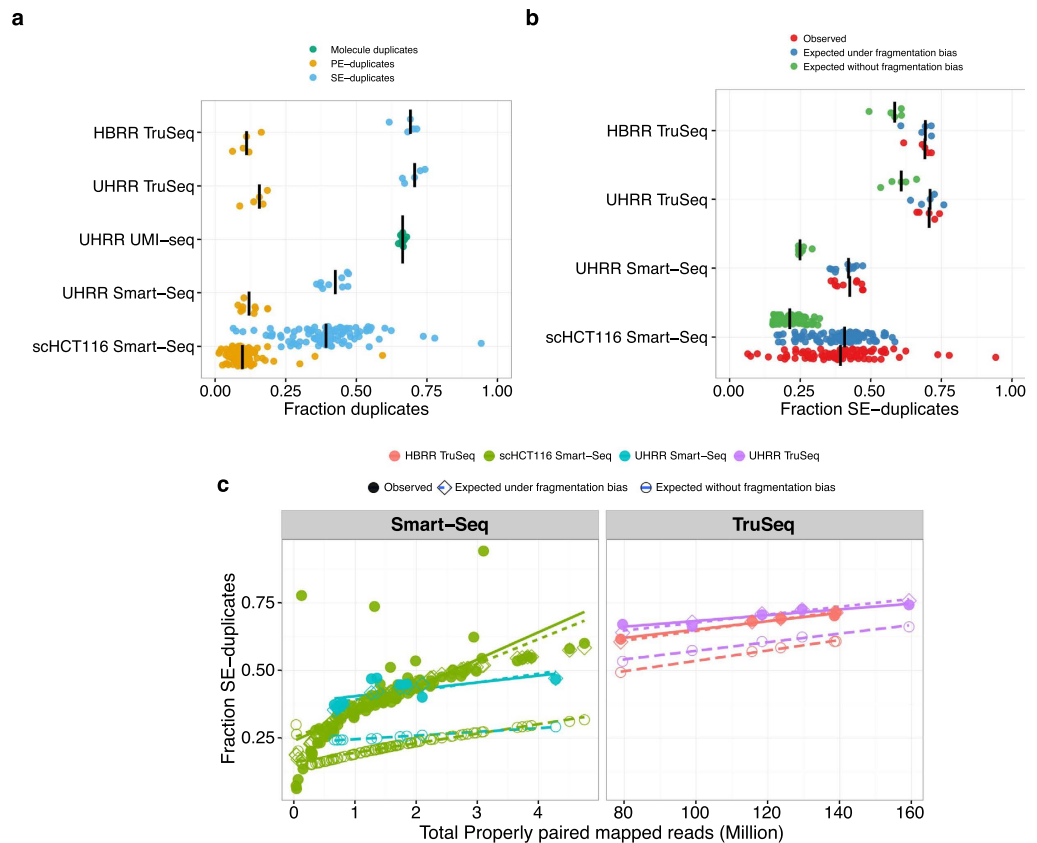


Figure 2. The Fraction of SE-duplicates increases with the total number of reads. In panel (a), we plot the fraction of computationally identified SE-duplicates (blue) and PE-duplicates (yellow) per sample. For the UMI-seq data, we identify duplicates only based on the experimental evidence provided by the UMIs. The black line marks the median for each dataset. If the correlation between sequencing depth and duplicates is due to sampling and fragmentation, we can quantify this impact. In (b), we plot the observed SE-duplicate fractions (red) and expected fractions (sampling—green, sampling + fragmentation—blue). (c) The left panel shows the two Smart-Seq datasets (UHRR- blue, scHCT116- green) and the right panel the TruSeq data (HBRR- red, UHRR- purple). Filled circles represent the observed fraction of SE-duplicates. Open symbols represent simulated data: Open diamonds mark the expected fractions of SE-duplicates under a simple sampling model and open circles are the expectations for a sampling model with fragmentation bias. The lines are the log-linear fits between sampling depth and SE-duplicates per dataset.

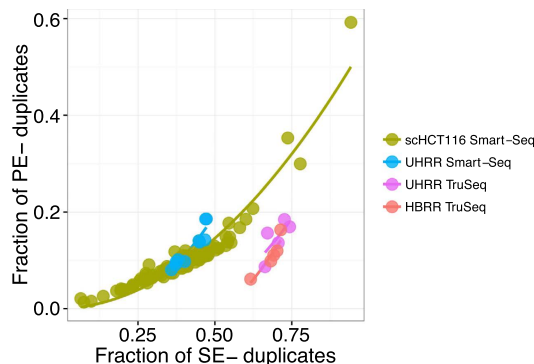


Figure 3. The relation between SE- and PE-duplicates. The relation between SE- and PE-duplicates is expected to follow a quadratic function, if the majority of duplicates are natural, i.e. due to fragmentation and sampling. Here, we show a quadratic fit for the different datasets (UHRR-TruSeq—purple, HBRR-TruSeq—red, UHRR-Smart-Seq—blue, scHCT116—green).

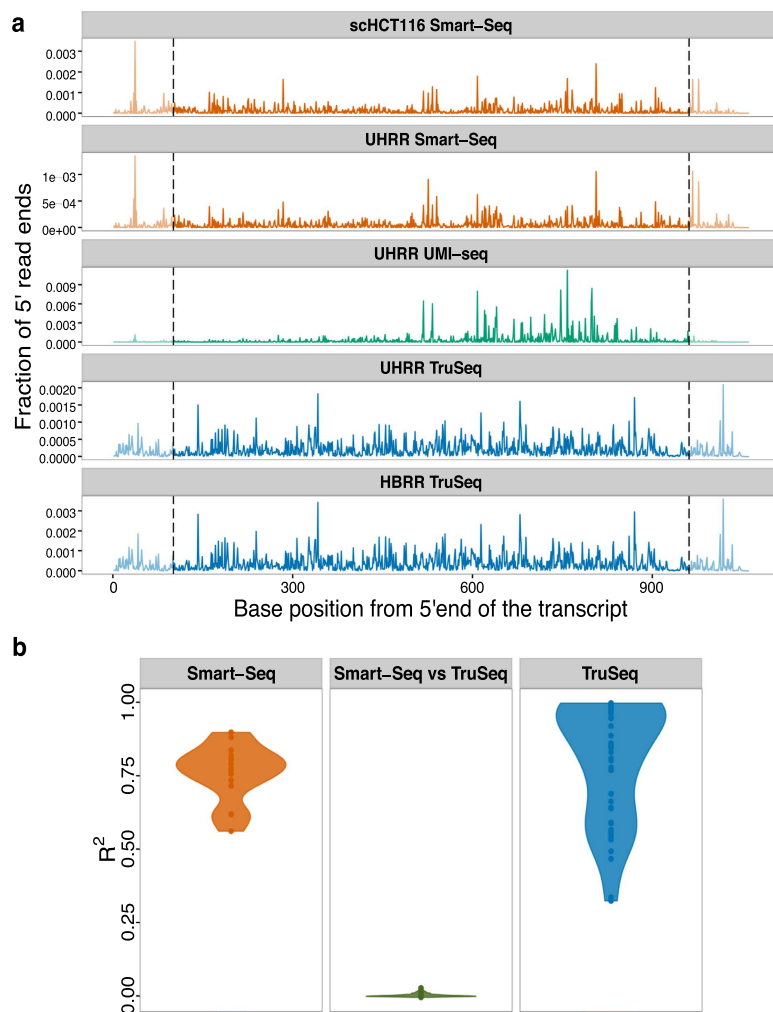


Figure 4. The fragmentation patterns of the ERCCs are highly reproducible for different samples prepared with the same RNA-seq library method. (a) Here, we plot the fraction of 5' read ends per position of ERCC-00002. Because the TruSeq libraries (blue) had read lengths of 100 bases, we do not consider the ends (grey dashed lines) for the calculation of the pair-wise R^2 values. Also, note that UMI-seq creates a stronger 3' bias. (b) Violin plot of the adjusted R^2 of a linear model of 5' read ends from different samples. The reproducibility of fragmentation is highest between Smart-Seq samples (orange), a little lower between the TruSeq samples and there is no correlation between samples from one Smart-Seq and one TruSeq sample (middle, green).

amplification is not the dominant source of duplicates. This is also consistent with our finding that most observed SE-duplicates are simply due to sampling (Supplementary Table S1 and Fig. 3).

Fragmentation is biased. The model above assumes that fragmentation does occur randomly. However, some sites are more likely to break than others and this might increase the fraction of SE-duplicates. To evaluate the impact and nature of fragmentation bias, we analysed ERCC spike-ins because they are exactly the same in all datasets. First, we test whether the variance in the frequency of 5' end mapping positions of ERCCs in one sample can explain a significant part of this variance in other samples prepared with the same method. On average, we find R^2 's of 0.77 and 0.85 for the Smart-Seq and TruSeq protocols, respectively. Note, that this high R^2 holds for samples that were prepared in different labs: for example the R^2 between the Smart-Seq samples prepared in our lab and the single cell data from the Quake lab ranges between 0.56–0.90. In contrast, if the R^2 is calculated for the comparison between one TruSeq and one Smart-Seq library, it drops to 0.0012 (Fig. 4a,b). Because the UMI-seq method specifically enriches for reads close to the 3' end of the transcript, we cannot compare fragmentation across the entire length of the transcript. However, if we limit ourselves to the 600 most 3' basepairs, we still find that the fragmentation pattern of the UMI-seq data shows a higher concordance with the two other datasets prepared also using the Smart-Seq protocol (mean $R^2 = 0.08$) than with the TruSeq data (mean $R^2 = 0.002$; Supplementary Figure S2). All in all, this is strong evidence that fragmentation reproducibly prefers the same sites given a library preparation protocol and thus read sampling is not random.

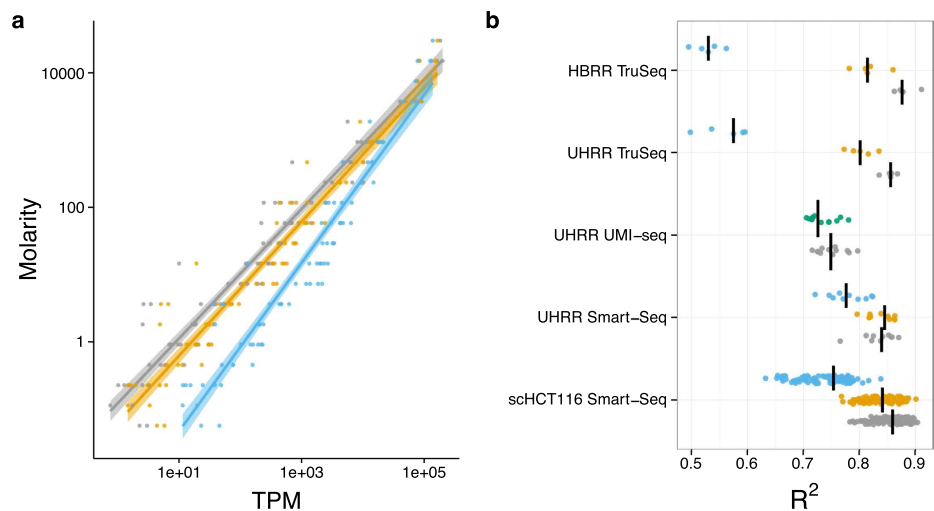


Figure 5. Removing duplicates does not improve the accuracy of expression quantification as measured using the ERCC spike-ins. Expression levels as quantified in transcripts per million reads (TPM) are a good predictor of the concentrations of the ERCC spike-ins. The log-linear fit of TPM vs. Molarity for one exemplary sample of the UHRR-TruSeq dataset is shown in (a). The most accurate prediction of ERCC molarity is the TPM estimator using all reads (grey). Removing duplicates as PE (yellow) makes the fit a little worse and removing SE-duplicates (blue) much worse. The adjusted R^2 for all samples are summarized in (b), the median for each dataset is marked as black line. The R^2 of the TPM estimate from the removal of PCR-duplicates using UMIs (green) is surprisingly similar to keeping PCR-duplicates (grey).

To identify potential causes for these non-random fragmentation patterns, we correlated the GC-content of the 15 bases around a given position with the number of 5' read ends. This explained very little of the fragmentation patterns in the TruSeq-data (median $R^2 = 0.0064$, 59% of the pair-wise comparisons significant with $p < 0.05$), and none in the Smart-Seq data (median $R^2 = 0.00002$, 18% significant with $p < 0.05$, Supplementary Figure S3a and Supplementary Table S2). Next, we built a binding motif for the Transposase²¹ from our UHRR-Smart-Seq data and, unsurprisingly, found that the motif has a very low information content (Supplementary Figure S3b) and accordingly a weak effect on the 5' read end count (median $R^2 = 0.0019$, 48% & 58% significant with $p < 0.05$ for scHCT116 & UHRR Smart-Seq, Supplementary Figure S3a and Supplementary Table S2).

Although we could not identify the cause for the fragmentation bias in the sequence patterns around the fragmentation site, we can still quantify the maximal impact of fragmentation bias on the number of SE-duplicates, simply by adjusting the effective length of the transcripts. For the TruSeq data, we estimate that a fragmentation bias that reduces the effective length by ~ 2 -fold gives a reasonably good fit, leaving on average 1% (0.1–3.0%) of the SE-duplicates unexplained. For the UHRR-Smart-Seq data, a ~ 38.5 -fold reduction in the effective length is needed and leaves only 3% (0.6–5.1%) of the duplicates unexplained. For the single cell data, the fragmentation bias that gives overall the best fit is a ~ 8 -fold reduction, however the fit is worse since the fraction of unexplained duplicates is still at $\sim 7\%$ and varies between 0.3% and 61% (Fig. 2b,c). In summary, we find that fragmentation bias contributes considerably to computationally identified read duplicates and is stronger for Smart-Seq, i.e. for enzymatic fragmentation, than for TruSeq, i.e. heat fragmentation.

Removal of duplicates does not improve the accuracy of quantification. To evaluate the impact of PCR duplicates on the accuracy of transcript quantification, we use again the ERCC spike-in mRNAs. Although, the absolute amounts of ERCC-spike ins might vary due to handling, the relative abundances of these 92 reference mRNAs can serve as a standard for quantification. Ideally, the known concentrations of the ERCCs should explain the complete variance in read counts and any deviations are a sign of measurement errors. We calculate the R^2 values of a log-linear fit of transcripts per million (TPM) versus ERCC concentration to quantify how well TPM estimates molecular concentrations and compare the fit among the different duplicate treatments. In no instance does removing read duplicates improve the fit, but in most cases the fit gets significantly worse (t-test, $p < 2 \times 10^{-3}$) except for the computational PE-duplicate removal of the UHRR-Smart-Seq and the duplicate removal using UMIs (Fig. 5). These results also hold when we use a more complex linear model including ERCC-length and GC-content (Supplementary Figure S4).

Removal of duplicates does not improve power. Most of the time we are not interested in absolute quantification, but are content to find relative differences, i.e. differentially expressed (DE) genes between groups of samples. The extra noise from the PCR-amplification has the potential to create false positives as well as to obscure truly DE genes. In order to assess the impact of duplicates on the power and the false discovery rate (FDR) to detect DE genes, we simulated data based on the estimated gene expression distributions of the five datasets. For comparability, we first equalized the sampling depth by reducing the number of mapped reads to 3

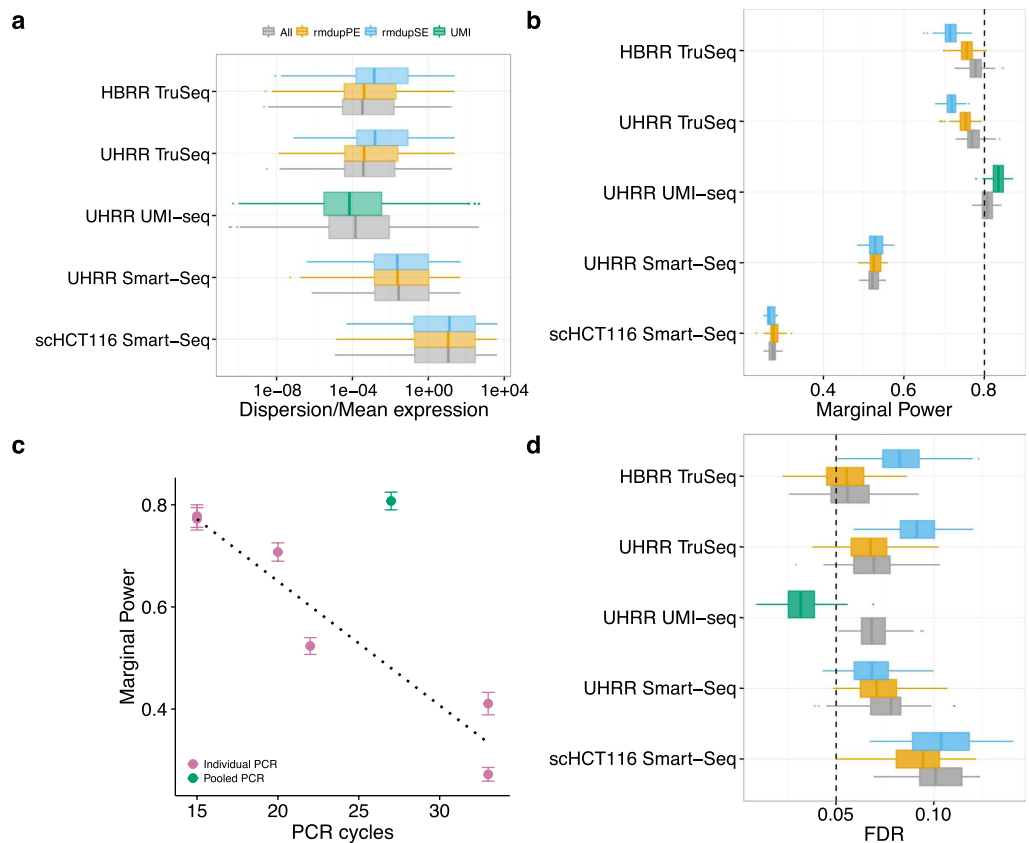


Figure 6. Duplicate removal has little influence on the power and FDR to detect DE-genes in comparison to the library preparation method. We estimated the distributions of mean expression and dispersion across genes for each dataset using DESeq2 after downsampling the datasets to 3 or 1 million reads. The distributions are estimated for the data including all reads (grey), removing PE-duplicates (yellow), removing SE-duplicates (blue) and for the UHRR-UMI-seq dataset removing duplicates using UMIs (green). We summarize distributions of dispersion/mean in (a). The estimated mean and dispersion distributions served as input for our power simulations using PROPER¹⁹. We did 100 simulations per dataset, whereas each dataset had two groups of six replicates (45 for scHT116) with 5% of the genes being differentially expressed between groups. In panel (b), we report the marginal power to detect a log₂-fold change of 0.5 and in panel (d) the corresponding FDR, whereas the nominal FDR was set to $\alpha = 0.05$ (dashed line). In panel (c), we plot our estimates of the marginal power against the number of PCR-cycles for each dataset. Error bars are standard deviation to the mean marginal power over 100 simulations. We find a surprisingly simple linear decline in power with the number of PCR-cycles, if we only consider datasets where PCR amplification was done separately for each sample of the dataset (violet). To confirm this simple fit we added two other datasets: (1) Bulk Smart-Seq dataset of mouse brain bulk RNA amplified using 20 PCR-cycles and (2) Single cell Smart-Seq dataset of 96 mouse embryonic stem cells that were amplified using 33 cycles. The only outlier is the UMI-seq dataset for which samples were pooled prior to amplification (green).

million and 1 million for bulk and single cell data, respectively. Next, we estimated gene-wise base mean expression and dispersion using DESeq2²².

There are no big differences in the distributions of mean baseline expression and dispersion estimates from the different duplicate treatments for the two Smart-Seq datasets, whereas there is a shift towards lower means and higher dispersions, when removing SE-duplicates for the TruSeq datasets. Dispersions shift only to lower values if we exclude duplicates based on identification by UMIs (Fig. 6a, Supplementary Figure S5). The empirical mean and dispersion distributions are then used to simulate two groups with six replicates for bulk-RNA-seq datasets and 45 replicates for the single cell dataset. In all cases we simulate that 5% of the genes are differentially expressed with log₂-fold changes drawn from a normal distribution with $N(0, 1.5)$ ¹⁹. We analysed 100 simulations per data-set using DESeq2 and calculate FDR and power for detecting DE-genes with a log₂-fold change of at least 0.5.

Except for the UHRR-UMI-seq dataset, the nominal FDR that we set to $\alpha = 5\%$ is exceeded: the means vary between 5.4% and 10.1%, whereas the HBRR TruSeq has the lowest and the scHCT116 Smart-Seq data the highest FDR (Fig. 6d). Computational removal of SE-duplicates increases the FDR by ~2% in the HBRR-TruSeq and the UHRR-TruSeq, has no significant impact on the scHCT116 dataset and, surprisingly, improves the FDR by

1% in the UHRR-Smart-Seq data (Fig. 6d). The computational removal of PE-duplicates harbors less potential for harm, in that it leaves the FDR unchanged for both TruSeq datasets and even slightly improves the FDR for the Smart-Seq datasets. Again, the only substantial improvement is achieved by duplicate removal using UMIs, which reduces the FDR from 7% to 3%. (t-test, $p < 1 \times 10^{-15}$).

The differences in the power are more striking. As for the FDR, the major differences are not between duplicate treatments, but between the datasets. For the TruSeq and the UHRR-UMI-seq datasets, the average power to detect a log₂-fold change of 0.5 is ~80% (Fig. 6b). For those datasets the changes in power due to duplicate removal are only marginal and for the computational removal using PE-duplicates it actually decreases the power for the TruSeq datasets by 2%, while for the UMI-seq data duplicate removal increases power by 2%. The power for the UHRR-Smart-Seq and the scHCT116 Smart-Seq datasets is much lower with 52% and 27%, respectively, and duplicate removal increases the power by only 1%.

The large differences in power between the datasets are unlikely to be ameliorated by increasing the number of replicates per group. In addition to the 6 and 45 replicates for which the results are reported above, we also conducted simulations for 12 and 90 replicates for bulk and the single cell data, respectively. This doubling in replicate number increases the power for the UHRR-Smart-Seq dataset only from 52 to 63% and for the single cell dataset from 27 to 34% (Supplementary Figure S6, Supplementary Table 3).

Discussion

RNA-seq has become a standard method for gene expression quantification and in most cases the sequencing library preparation involves amplification steps. Ideally, we would like to count the number of RNA molecules in the sample and thus would want to keep only one read per molecule. A common strategy applied for amplification correction in SNP-calling and ChIP-Seq protocols^{23,24} is to simply remove reads based on their 5' ends, so called read duplicates. Here, we show that this strategy is not suitable for RNA-seq data, because the majority of such SE-duplicates is likely due to sampling. For highly transcribed genes, it is simply unavoidable that multiple reads have the same 5' end, also if they originated from different RNA-molecules. We find that only ~10% (TruSeq) and ~20% (Smart-Seq) of the read duplicates cannot be explained by a simple sampling model with random fragmentation. This fraction decreases even more, if we factor in that the fragmentation of mRNA or cDNA during library preparation is clearly non-random, as evidenced by a strong correlation between the 5' read positions of the ERCC-spike-ins across samples. Because local sequence content has little or no detectable effect on fragmentation, we cannot predict fragmentation, but we can quantify the observed effect. For example, we find that a fragmentation bias that halves the number of break points can fit the observed proportion of duplicates for TruSeq libraries well. For the Smart-Seq datasets, fragmentation biases would have to be much higher to explain the observed numbers of read duplicates. Furthermore, the fit between model estimates and the observed duplicate fractions is worse than for the TruSeq data and the model estimates for fragmentation bias are also inconsistent between the datasets (38.5 for the UHRR and 8 for the scHCT116).

Since computational methods cannot distinguish between fragmentation and PCR duplicates, the removal of read duplicates could introduce a bias rather than removing it. Using the ERCC-spike-ins, we can indeed show that removing duplicates computationally does not improve a fit to the known concentrations, but rather makes it worse, especially if only single-end reads are available (Fig. 5). This is in line with our observation that most single end duplicates are due to sampling and fragmentation. Hence, removing duplicates is similar to a saturation effect known for microarrays^{25–27}.

Moreover, the Smart-Seq protocol, which was designed for small starting amounts, involves PCR amplification before the final fragmentation of the sequencing library. Thus in the case of Smart-Seq, computational methods cannot identify PCR duplicates that occur during the pre-amplification step. When we use unique molecular identifiers (UMIs), we find that 66% of the reads are PCR duplicates and only 34% originate from independent mRNA molecules. In contrast, when using paired-end mapping for a comparable Smart-Seq library, we identify 13% as duplicates and 87% as unique. This might in part be due to the fact that in UMI-Seq we sequence mainly 3' ends of transcripts, thus decreasing the complexity of the library, which in turn increases the potential for PCR duplicates for a given sequencing depth (Fig. 4a, Supplementary Figure S1). However, it is unlikely that library complexity can explain the 53% difference in duplicate occurrence. This difference is more likely to be due to PCR-duplicates that are generated during pre-amplification and thus remain undetectable by computational means.

All in all, computational methods are limited when it comes to removing PCR-duplicates, but how much noise or bias do PCR duplicates introduce? In other words, we want to know how PCR-duplicates impact the power and the false discovery rate for the detection of differentially expressed genes. Both, power and FDR, are determined by the gene-wise mean expression and dispersion. Based on simulated differential expression using the empirically determined mean and dispersion distributions, we find that computational removal of duplicates has either a negligible or a negative impact on FDR and power, and we therefore recommend not to remove read duplicates. In contrast, if PCR duplicates are removed using UMIs, both FDR and power improve. Even though the effects in the bulk data analysed here are relatively small: FDR is improved by 4% and the power by 2%, UMIs will become more important when using smaller amounts of starting material as it is the case for single-cell RNA-seq^{6,28}.

The major differences in power are between the datasets with the TruSeq and the UMI-seq data achieving a power of around 80%, the UHRR-Smart-Seq 52% and the single cell Smart-Seq data (scHCT116) only 27%. Note that this apparently bad performance of the single cell Smart-Seq data is at least in part due to an unfair comparison. While all the other datasets were produced using commercially available mRNA and thus represent true technical replicates, the single cell data necessarily represent biological replicates and thus are expected to have a larger inherent variance and thus lower power.

However, also the UHRR Smart-Seq bulk data achieves with 52% a much lower power than the other bulk datasets. One possible explanation for the differences in power is the total number of PCR-cycles involved in

the library preparation. With every PCR-cycle the power to detect a log₂-fold change of 0.5 appears to drop by 2.4% (Fig. 6c). The only exception is the UMI-seq dataset, that gives a power of 81%, even if duplicates are not removed, which is comparable to the power reached with TruSeq data despite the UMI-seq method having 12 more PCR-cycles. Technically UMI-seq is most similar to the Smart-Seq method. The biggest difference between the two methods is that all UMI-seq libraries are pooled before PCR-amplification, suggesting that the PCR-noise is due to the different PCR-reactions and not due to amplification efficiency per-se.

We conclude that computational removal of duplicates is not recommendable for differential expression analysis and if sufficient starting material is available so that only few PCR-cycles are necessary, the loss in power due to PCR duplicates is negligible. However, if more amplification is needed, power would be improved if all samples are pooled early on, and for really low amounts as for single cell data also the gain in power that is achieved by removing PCR-duplicates using UMIs will become important.

Methods

Datasets. We used six datasets representing the TruSeq, Smart-Seq and UMI-seq protocols and varying amounts of starting material from bulk RNA or single cell RNA. All analysed datasets contain the ERCCs spike-in RNAs. This is a set of 92 artificial poly-adenylated RNAs designed to match the characteristics of naturally occurring RNAs with respect to their length (273–2022 bp), their GC-content (31–53%) and concentrations of the ERCCs (0.01–30,000 attomol/ μ l). The recommended ERCC spike-in amounts result in 5–10⁷ ERCC RNA molecules in the cDNA synthesis reaction.

To reduce biological variation, we used the well-characterized Universal Human Reference RNA (UHRR; Agilent Technologies) for the two datasets produced for this study. We downloaded UHRR- and HBRR-TruSeq data from SEQC/MAQC-III². Finally, we also analyse the single cell data published in Wu *et al.*²⁰, for which the colorectal cancer cell-line HCT116 was used (Table 1). The input mostly being commercially distributed human samples, we expect all biological samples analysed in this study to have similarly high quality and complexity. All data that were generated for this project were submitted to GEO under accession GSE75823.

RNA-seq library preparation and sequencing. For the Smart-Seq libraries, 250 ng of Universal Human Reference RNA (UHRR; Agilent Technologies) and ERCC spike-in control mix I (Life Technologies) were used and cDNA was synthesized as described in the Smart-Seq2 protocol from Picelli *et al.*¹³. However, because we used more mRNA to begin with, we reduced the number of pre-amplification PCR cycles to 9 cycles instead of the 18–21 recommended in Picelli *et al.*¹³. 1 ng of pre-amplified cDNA was then used as input for Tn5 transposon tagmentation by the Nextera XT Kit (Illumina), followed by 12 PCR cycles of library amplification. For sequencing, equal amounts of all libraries were pooled.

For the UMI-seq libraries, we started with 10 ng of UHRR-RNA to synthesise cDNA as described in Soumillon *et al.*¹⁶. This protocol is very similar to the Smart-Seq protocol, however the first strand cDNA is decorated with sample-specific barcodes and unique molecular identifiers. The barcoded cDNA from all samples was then pooled, purified and unincorporated primers digested with Exonuclease I (NEB). Pre-amplification was performed by single-primer PCR for 15 cycles. 1 ng of full-length cDNA was then used as input for the Nextera XT library preparation with the modification of adding a custom i5 primer to enrich for barcoded 3' ends.

Library pools were sequenced on an Illumina HiSeq1500. The Smart-Seq libraries were sequenced using 50 cycles of paired-end sequencing on a High-Output flow-cell. The UMI-seq libraries were sequenced on a rapid flow-cell with paired-end layout, where the first read contains the sequences of the sample barcode and the UMI sequence using 17 cycles. The second read contains the actual cDNA fragment with 46 cycles.

Data Processing. For Smart-Seq and TruSeq libraries, the sequenced reads were mapped to the human genome (hg19) and the splice site information from the ensembl annotation (GRCh37.75) using STAR(version:2.4.0.1)²⁹ with the default parameters, reporting only the best hit per read. The genome index was created with `-sjdbOverhang 'readlength-1'`. Because the ERCCs are transcript sequences no splice-aware mapping is necessary and therefore we used NextGenMap for the ERCCs³⁰. Except for three parameters, (1) the maximum fragment size which was set to 10kb, (2) the minimum identity set to 90% and (3) reporting only the best hit per read, we also used the default parameters for NextGenMap. Note that we also included hg19 and did not map to ERCC sequences only. The mapped reads were assigned to genes [Ensembl database annotation version GRCh37.75] using FeatureCount from the bioconductor package Rsubread³¹ (see Supplementary text).

For UMI-seq data, cDNA reads were mapped to the transcriptome as recommended in Soumillon *et al.*¹⁶ using the Ensembl annotation [version GRCh37.75] and NextGenMap³⁰ (Supplementary text). If either the sample barcode or the UMI had at least one base with sequence quality ≤ 10 or contained 'N's the read was discarded. Next, we generated count tables for reads or UMIs per gene. Finally, mitochondrial and ambiguously assigned reads were removed from all libraries.

Duplicate detection and removal. We defined single-end (SE) read duplicates as reads that map to the same 5' position, have the same strand and the same CIGAR value. Because we cannot determine the exact mapping position for 5' soft clipped reads, we discard them. To flag paired-end duplicates (PE), we used the same requirements as for the SE-duplicates, those requirements had just to be fulfilled for both reads of a pair.

Model for the fraction of sampling and fragmentation duplicates. We obtain an expectation for the number of reads if duplicates are identified via their 5' position and only one read per 5' end position is kept. The only input parameters are the observed number of reads per gene (r_G) and the effective length of the gene ($L_{eG} = L - 2 \times \text{read-length}$). Then the expected number of unique reads can be estimated as

$$E[r_{G_{RMDUP}}] = s \sum_{k \in 1 \dots r_G} r_G P(X = k) / k \quad (1)$$

whereas $P(X = k)$ can be calculated using a positive Poisson distribution with $\lambda_G = r_G / L_{eG}$ and s is a scaling factor $s = 1 / \sum_{k \in 1 \dots r_G} P(X = k)$.

In order to estimate the level of fragmentation bias, we simply modified the effective length L_{eG} by a factor $f \times L_{eG}$.

Fragmentation pattern analysis. To compare fragmentation sites across libraries, we counted 5' read starts per position for the ERCCs across all datasets using samtools and in house perl scripts. To avoid edge effects in later analyses, we excluded the first and last 100 bases of each ERCC, whereas 100 bases is the maximum read length of datasets analysed here.

We generated a Position Weight Matrix (PWM) for the transposase (Tn5) motif by simply stacking up the 30 bases of the putative Transposase binding sites from all UHRR-Smart-Seq reads. Those 30 bases are identified as 6 bases upstream of the 5' read end and the 24 downstream²¹. The resulting PWM was then used to calculate motif scores across the ERCCs using the Bioconductor package PWMEnrich³².

Power evaluation for differential expression. For power analysis, we estimated the mean baseline expression and dispersion for all datasets after downsampling them to 3 and 1 million reads for bulk and single cell data, respectively. This was done for all three duplicate treatments (keep all, remove SE and remove PE) using DESeq2²² with standard parameters. Furthermore, genes with very low dispersions (< 0.001) were removed. We chose the sample sizes 3, 6 and 12 per condition for the bulk data and 30, 45 and 90 for the single cell dataset, because they seemed to be a good representation of the current literature. For the simulations, we use an in-house adaptation of the Bioconductor-package PROPER¹⁹. As suggested in Wu *et al.*¹⁹, we set the fraction of differentially expressed genes between groups to 0.05 and the log₂-fold change for the DE-genes was drawn from a normal distribution with $N(0, 1.5)$. We generated 100 simulations per original input data-set and analysed them using DESeq2. Next, we calculated the power to detect a log₂-fold change of at least 0.5 and the according FDR using $\alpha = 0.05$.

References

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. 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).
3. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
4. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
5. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
6. Ziegenhain, C. *et al.* Comparative analysis of single-cell RNA-sequencing methods. *bioRxiv* doi: 10.1101/035758 (2016).
7. 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).
8. Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* **42**, 8845–8860 (2014).
9. Kozarewa, I. *et al.* Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (G + C)-biased genomes. *Nat. Methods* **6**, 291–295 (2009).
10. Mamanova, L. *et al.* FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat. Methods* **7**, 130–132 (2010).
11. Lahens, N. F. *et al.* IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.* **15**, R86 (2014).
12. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
13. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
14. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
15. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
16. Soumillon, M. *et al.* Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv* doi: 10.1101/003236 (2014).
17. Baker, S. C. *et al.* The external RNA controls consortium: a progress report. *Nat. Methods* **2**, 731–734 (2005).
18. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
19. Wu, H., Wang, C. & Wu, Z. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* **31**, 233–241 (2015).
20. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
21. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* **11**, R119 (2010).
22. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
23. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
24. Chen, Y. *et al.* Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat. Methods* **9**, 609–614 (2012).
25. Siegmund, K. H., Steiner, U. E. & Richert, C. ChipCheck - a program predicting total hybridization equilibria for DNA binding to small oligonucleotide microarrays. *J. Chem. Inf. Comput. Sci.* **43**, 2153–2162 (2003).
26. Dodd, L. E., Korn, E. L., McShane, L. M., Chandramouli, G. V. R. & Chuang, E. Y. Correcting log ratios for signal saturation in cDNA microarrays. *Bioinformatics* **20**, 2685–2693 (2004).
27. Hsiao, L. L. *et al.* Correcting for signal saturation errors in the analysis of microarray data. *Biotechniques* **32**, 330–2, 334, 336 (2002).
28. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
29. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

30. Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790–2791 (2013).
31. Liao, Y., Smyth, G. K. & Shi, W. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
32. Stojnic, R. & Diez, D. PWMEnrich: PWM enrichment analysis. R package version 4.6.0. Cambridge Systems Biology Institute, University of Cambridge, UK. URL <https://www.bioconductor.org/packages/release/bioc/html/PWMEnrich.html> (2015).

Acknowledgements

We thank Khalis Afnan and Sabrina Weser for help with the RNA-seq library preparation. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the SFB1243 (Subprojects A14/A15) as well as a travel grant to C.Z. by the Boehringer Ingelheim Fonds.

Author Contributions

S.P. and C.Z. conceived the study. C.Z. prepared RNA-seq libraries. S.P., I.H. and B.V. analyzed the data. I.H., S.P. and W.E. wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

Accession codes: RNA-seq data generated for this study is submitted to GEO under the accession code: GSE75823.

Supplementary information accompanies this paper at <http://www.nature.com/srep>

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

How to cite this article: Parekh, S. *et al.* The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 25533; doi: 10.1038/srep25533 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

The impact of amplification on differential expression analyses by RNA-seq

Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, Ines Hellmann*

Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians University, Großhaderner Str. 2, 82152 Martinsried, Germany.

* hellmann@bio.lmu.de

Supplementary figures

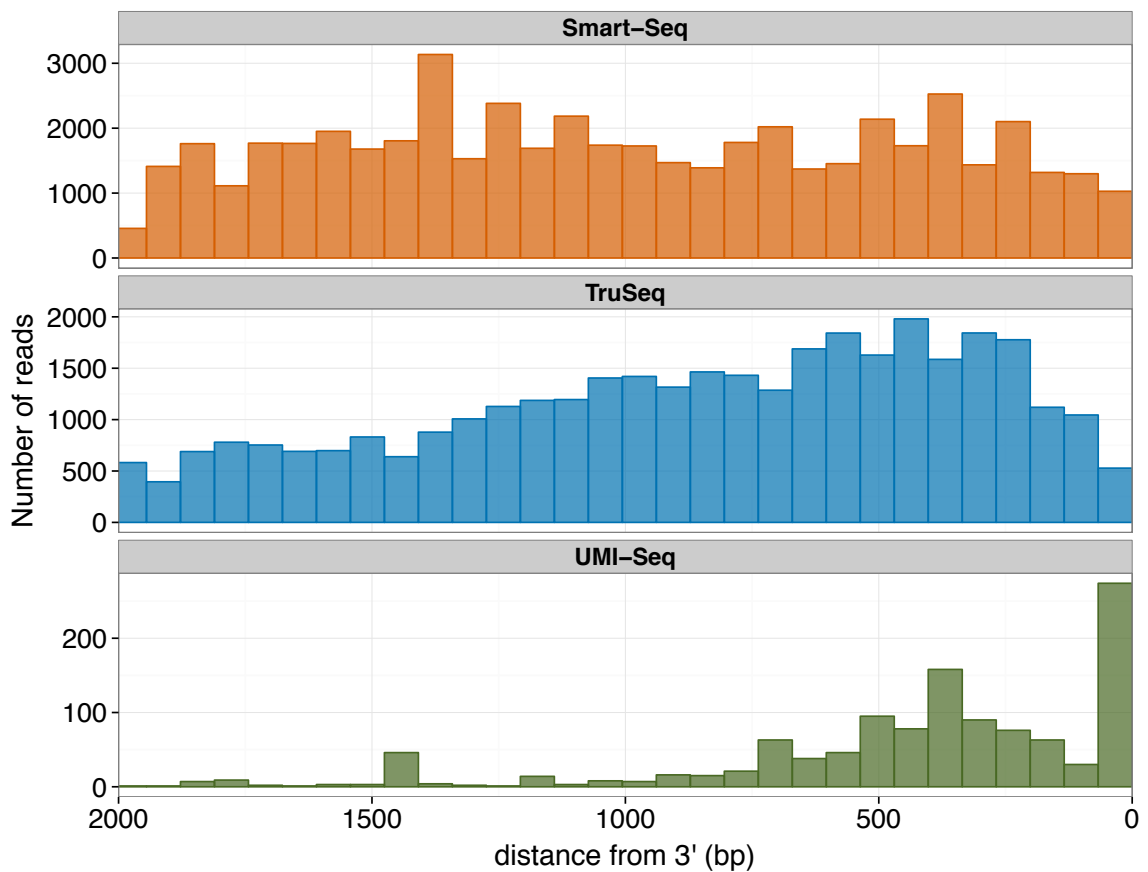


Figure S1: 3' bias in fragmentation site is prominent in UMI-seq. The histogram showing distance of the fragmentation site from 3' end of the gene measured from ERCC spike-ins of length $\sim 2kb$. Colors represent library preparation methods, 'blue' - Smart-Seq, 'orange' - TruSeq, 'green' - UMI-seq.

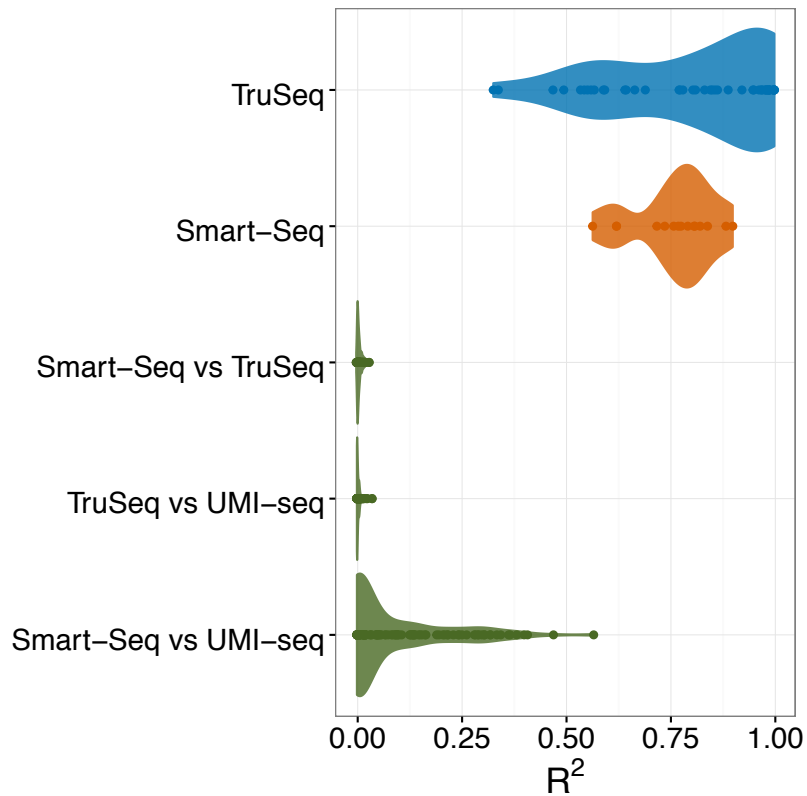


Figure S2: The fragmentation patterns of the most 3' 600bp of ERCCs are relatively reproducible between Smart-Seq and UMI-seq. Violin plots of the adjusted R^2 from a linear model between fraction of 5' read ends from different samples. The adjusted R^2 are calculated considering full length for Smart-Seq and TruSeq methods whereas for comparison to UMI-seq the most 3' 600bp are considered. The reproducibility of fragmentation is highest within Smart-Seq (orange) and TruSeq samples (blue). Fragmentation reproducibility between Smart-Seq and UMI-seq samples (green) is higher than compared to TruSeq (green), as both methods use transposase tagmentation.

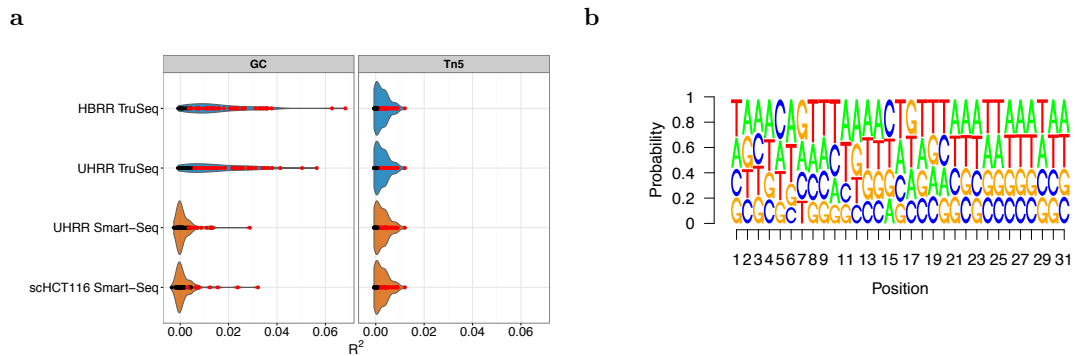


Figure S3: Fragmentation does not appear to have a cutting site preference. Colors of the violin plots represent library preparation methods, 'blue - Smart-Seq, 'orange' - TruSeq and dots are colored by the significance of the fit where 'red' - $p\text{-value} \leq 0.05$ and 'black' - $p\text{-value} > 0.05$. **a)** The left panel shows violin plots of the adjusted R^2 of linear model fit between background corrected GC content of 15bases window and fraction of 5' read ends of the middle base in the window for each ERCC spike-in and the right panel shows the adjusted R^2 of linear model fit between Tn5 motif score calculated for ERCC spike-in RNAs. **b)** Sequence logo of the Tn5 motif derived from UHRR Smart-Seq dataset.

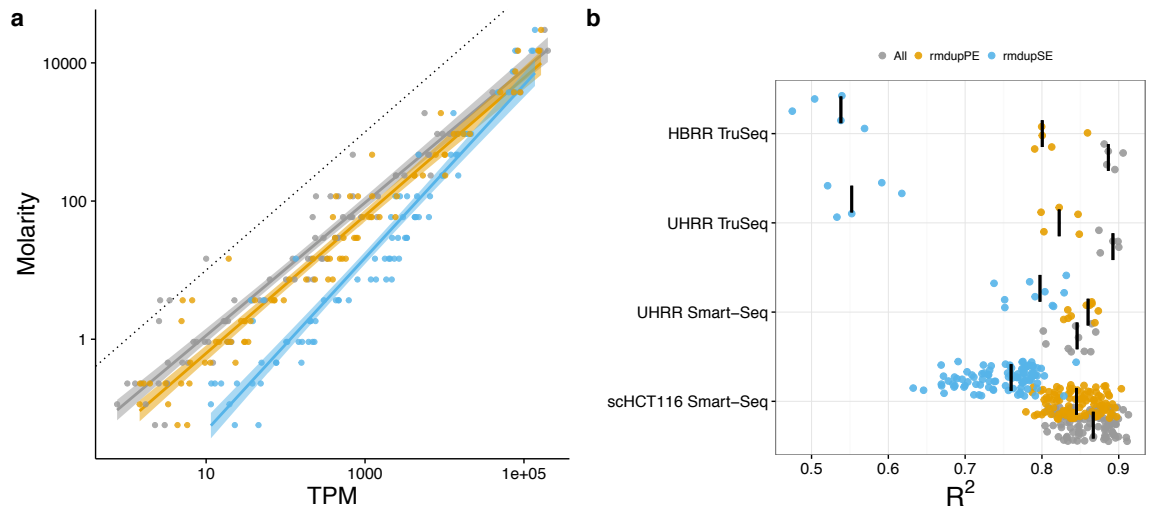


Figure S4: Removing duplicates does not improve the accuracy of expression quantification as measured using the ERCC spike-ins. Expression levels as quantified in transcripts per million reads (TPM) are considered to be good measure of ERCC spike ins. However, other factors like capture and sequencing efficiency can not be explained by TPM. One exemplary sample of the UHRR-TruSeq dataset as shown in Figure 5 of the main text is shown in **a**). The dashed grey line shows the bisecting line. We calculated the log-linear fit of counts per million (CPM) vs. Molarity also controlling for GC content and length of the transcript. The adjusted R^2 for all samples are summarized in **b**), the median for each dataset is marked as black line. The colors represent different duplicates treatment. All reads (grey), removing PE-duplicates (yellow) and removing SE-duplicates (blue).

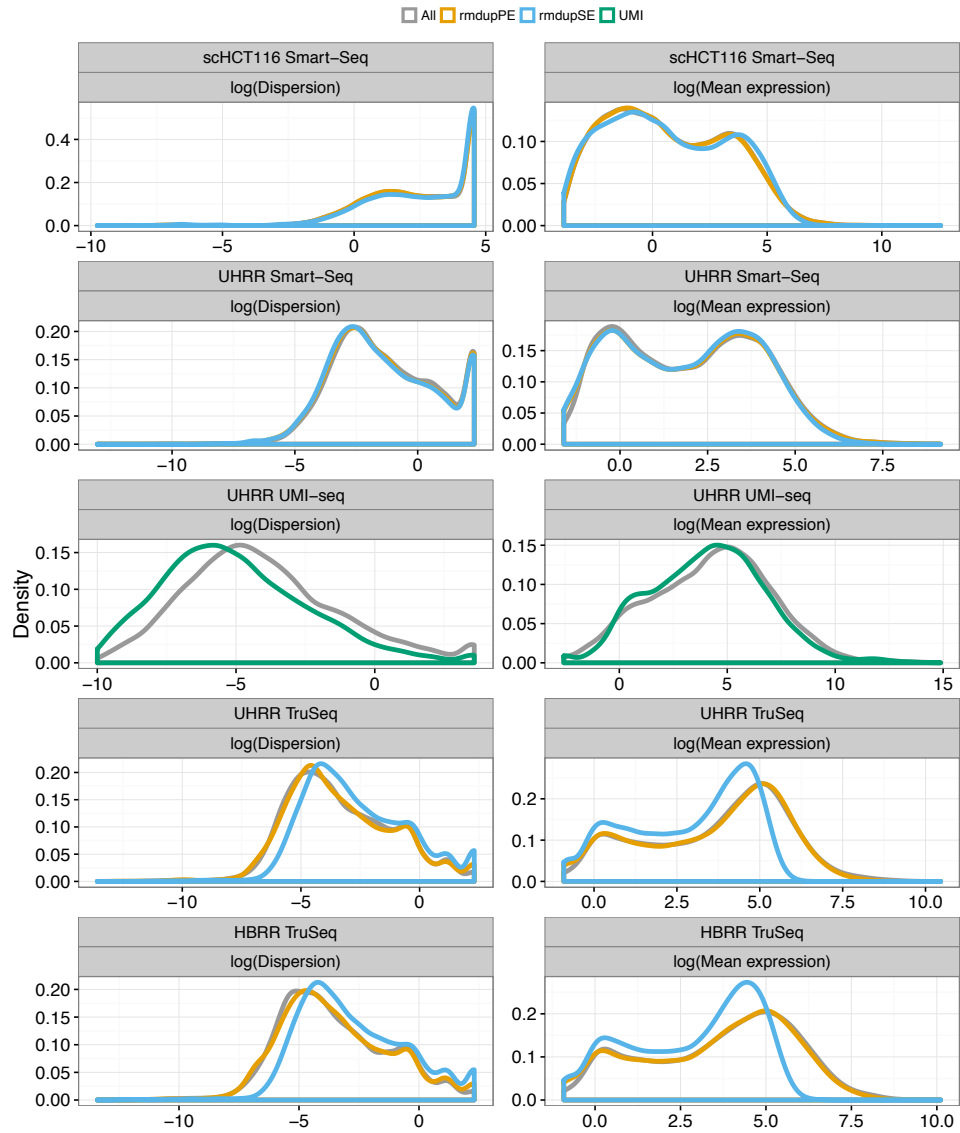


Figure S5: Empirical mean and dispersion distributions are used to estimate power to detect differential expression. The left panel shows density plot of $\log(\text{dispersion})$ and the right panel the $\log(\text{mean baseline expression})$ measured by DESeq2 for each study. Different duplicates treatments are represented by colors, All reads- grey, removing PE-duplicates- orange, removing SE-duplicates- blue and removing duplicate molecules in UMI-seq as green.

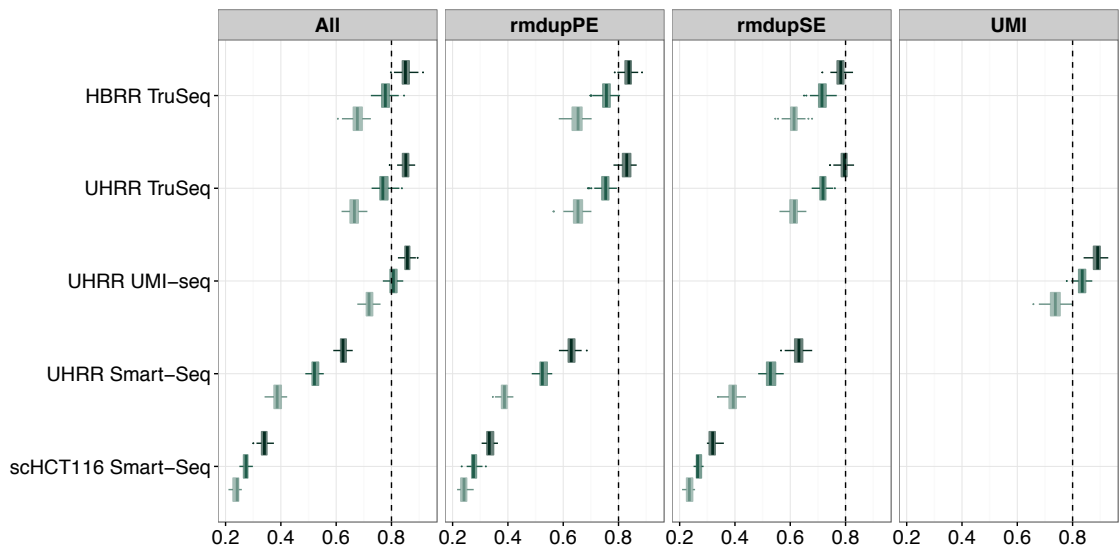


Figure S6: Power to detect differential expression increases with increased sample size. The box-plot shows marginal power to detect 0.5 log₂foldchange at 5% nominal FDR for different sample sizes. Colors gradient from light to dark represent sample sizes 3,6 and 12 for the bulk and 30,45 and 90 for the single cell datasets.

Supplementary text

Detailed commands used for mapping are given below.

STAR genome generate

```
STAR --runThreadN 10 --runMode genomeGenerate --genomeDir hg19STARindex --genomeFastaFiles hg19.fa --sjdbGTFfile GRCh37.75.gtf --sjdbOverhang 'readLen-1'
```

STAR mapping

```
STAR --readFilesIn R1.fastq R2.fastq --runThreadN 10 --outFileNamePrefix samplename --outFilterMultimapNmax 1 --outSAMunmapped Within --outSAMtype BAM SortedByCoordinate --sjdbGTFfile GRCh37.75.gtf --genomeDir hg19STARindex --sjdbOverhang 'readLen-1' --outFilterType BySJout --outSJfilterReads Unique
```

NextGenMap mapping

For ERCC spike-ins

```
ngm.4.12 -1 R1.fastq -2 R2.fastq -t 10 -i 0.9 -X 10000 -r ERCCs.fa -o samplename.sam
```

For UMI-seq data

```
ngm.4.12 -q R1.fastq -t 10 -i 0.9 -r GRCh37.75.fa -o samplename.sam
```

Supplementary tables

Table S1: Summary of squared terms from quadratic fit between PE-dup and SE-dup ($PE\text{-dup} \sim SE\text{-dup} + (SE\text{-dup})^2 + 0$)

Study name	Beta ²	Std. Error	t value	Pr(> t)
scHCT116 Smart-Seq	0.542	0.0302	17.94	0.0000
UHRR Smart-Seq	1.168	0.246	4.739	0.001
UHRR TruSeq	0.840	0.619	1.356	0.268
HBRR TruSeq	1.134	0.338	3.350	0.044

Table S2: Median R² and percentage of significant ERCCs for the lm fit between GC content/Tn5 motif score and 5' read ends

Study name	GC		Tn5	
	R ²	%Significant*	R ²	%Significant*
scHCT116 Smart-Seq	-0.00027	16%	0.00112	49%
UHRR Smart-Seq	0.00020	19%	0.00174	59%
UHRR TruSeq	0.00614	57%	0.00077	43%
HBRR TruSeq	0.00657	61%	0.00077	43%

*Percentage of ERCCs with p-value ≤ 0.05

Table S3: Summary of power analysis

Study name	Sample size	Mean FDR	Marginal power	Avg # of TD	Avg # of FD	FDC	DupType	PCRcycles	Amount(ug)
HBRR TruSeq	3	0.06	0.68	239.63	16.28	0.07	All	15	1.00
HBRR TruSeq	3	0.06	0.65	232.52	16.35	0.07	rndupPE	15	1.00
HBRR TruSeq	3	0.07	0.61	266.98	20.45	0.08	rndupSE	15	1.00
HBRR TruSeq	6	0.06	0.78	277.37	19.16	0.07	All	15	1.00
HBRR TruSeq	6	0.05	0.76	273.61	17.75	0.06	rndupPE	15	1.00
HBRR TruSeq	6	0.08	0.72	315.48	31.46	0.10	rndupSE	15	1.00
HBRR TruSeq	12	0.06	0.85	307.49	21.32	0.07	All	15	1.00
HBRR TruSeq	12	0.05	0.84	298.30	19.26	0.06	rndupPE	15	1.00
HBRR TruSeq	12	0.07	0.78	352.17	30.74	0.09	rndupSE	15	1.00
scHCT116 Smart-Seq	30	0.14	0.24	194.30	33.80	0.17	All	33	0.00
scHCT116 Smart-Seq	30	0.14	0.24	208.35	34.00	0.16	rndupPE	33	0.00
scHCT116 Smart-Seq	30	0.15	0.23	211.20	37.70	0.18	rndupSE	33	0.00
scHCT116 Smart-Seq	45	0.10	0.27	230.45	26.60	0.12	All	33	0.00
scHCT116 Smart-Seq	45	0.09	0.28	246.70	25.35	0.10	rndupPE	33	0.00
scHCT116 Smart-Seq	45	0.10	0.27	251.00	29.35	0.12	rndupSE	33	0.00
scHCT116 Smart-Seq	90	0.06	0.34	293.92	21.13	0.07	All	33	0.00
scHCT116 Smart-Seq	90	0.07	0.33	307.00	22.35	0.07	rndupPE	33	0.00
scHCT116 Smart-Seq	90	0.07	0.32	308.55	22.75	0.07	rndupSE	33	0.00
UHRR UMI-seq	3	0.06	0.72	447.41	33.19	0.07	All	27	0.01
UHRR UMI-seq	3	0.03	0.74	238.36	7.00	0.03	UMI	27	0.01
UHRR UMI-seq	6	0.07	0.81	507.54	43.54	0.09	All	27	0.01
UHRR UMI-seq	6	0.03	0.83	271.73	10.30	0.04	UMI	27	0.01
UHRR UMI-seq	12	0.06	0.86	553.42	43.01	0.08	All	27	0.01
UHRR UMI-seq	12	0.04	0.89	301.07	13.42	0.04	UMI	27	0.01
UHRR Smart-Seq	3	0.06	0.39	288.66	18.89	0.07	All	22	0.25
UHRR Smart-Seq	3	0.06	0.39	282.26	17.25	0.06	rndupPE	22	0.25
UHRR Smart-Seq	3	0.05	0.39	283.54	15.46	0.05	rndupSE	22	0.25
UHRR Smart-Seq	6	0.08	0.52	404.17	34.57	0.09	All	22	0.25
UHRR Smart-Seq	6	0.07	0.53	399.62	32.43	0.08	rndupPE	22	0.25
UHRR Smart-Seq	6	0.07	0.53	398.36	30.53	0.08	rndupSE	22	0.25
UHRR Smart-Seq	12	0.06	0.63	489.58	35.81	0.07	All	22	0.25
UHRR Smart-Seq	12	0.06	0.63	483.90	34.61	0.07	rndupPE	22	0.25
UHRR Smart-Seq	12	0.06	0.63	481.09	32.36	0.07	rndupSE	22	0.25
UHRR TruSeq	3	0.08	0.67	274.02	25.72	0.09	All	15	1.00
UHRR TruSeq	3	0.08	0.65	269.81	25.53	0.09	rndupPE	15	1.00
UHRR TruSeq	3	0.08	0.61	316.45	30.10	0.10	rndupSE	15	1.00
UHRR TruSeq	6	0.07	0.77	319.40	26.78	0.08	All	15	1.00
UHRR TruSeq	6	0.07	0.75	314.12	25.36	0.08	rndupPE	15	1.00
UHRR TruSeq	6	0.09	0.72	375.37	41.36	0.11	rndupSE	15	1.00
UHRR TruSeq	12	0.06	0.85	350.17	24.90	0.07	All	15	1.00
UHRR TruSeq	12	0.05	0.83	345.31	22.83	0.07	rndupPE	15	1.00
UHRR TruSeq	12	0.08	0.79	412.77	39.44	0.10	rndupSE	15	1.00

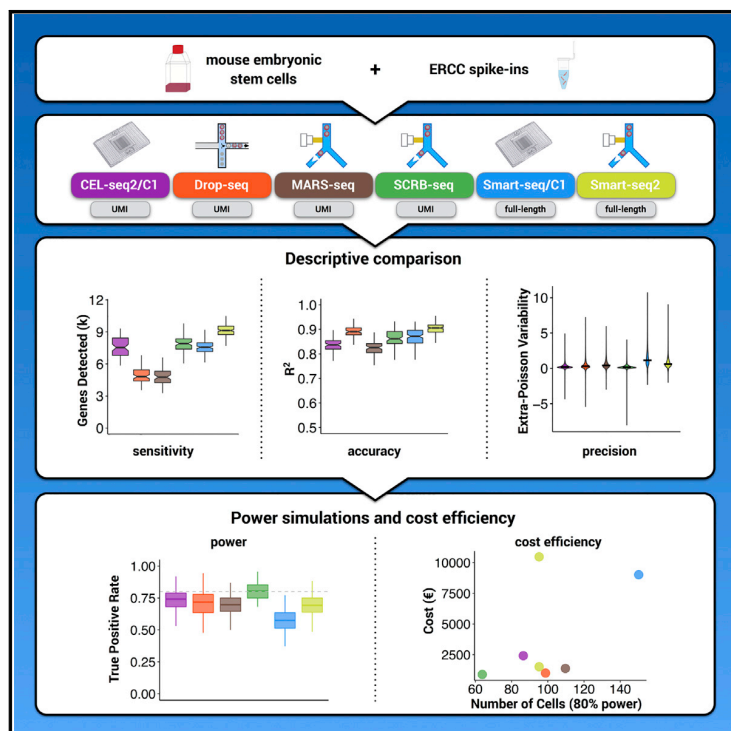
Identifying and addressing issues in single-cell RNA-seq analysis

Comparative Analysis of Single-Cell RNA Sequencing Methods

Molecular Cell

Comparative Analysis of Single-Cell RNA Sequencing Methods

Graphical Abstract



Authors

Christoph Ziegenhain, Beate Vieth, Swati Parekh, ..., Holger Heyn, Ines Hellmann, Wolfgang Enard

Correspondence

enard@bio.lmu.de

In Brief

Ziegenhain et al. generated data from mouse ESCs to systematically evaluate six prominent scRNA-seq methods. They used power simulations to compare cost efficiencies, allowing for informed choice among existing protocols and providing a framework for future comparisons.

Highlights

- The study represents the most comprehensive comparison of scRNA-seq protocols
- Power simulations quantify the effect of sensitivity and precision on cost efficiency
- The study offers an informed choice among six prominent scRNA-seq methods
- The study provides a framework for benchmarking future protocol improvements



Comparative Analysis of Single-Cell RNA Sequencing Methods

Christoph Ziegenhain,¹ Beate Vieth,¹ Swati Parekh,¹ Björn Reinius,^{2,3} Amy Guillaumet-Adkins,^{4,5} Martha Smets,⁶ Heinrich Leonhardt,⁶ Holger Heyn,^{4,5} Ines Hellmann,¹ and Wolfgang Enard^{1,7,*}

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

²Ludwig Institute for Cancer Research, Box 240, 171 77 Stockholm, Sweden

³Department of Cell and Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden

⁴CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain

⁵Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain

⁶Department of Biology II and Center for Integrated Protein Science Munich (CIPSM), Ludwig-Maximilians University, Großhaderner Straße 2, 82152 Martinsried, Germany

⁷Lead Contact

*Correspondence: enard@bio.lmu.de

<http://dx.doi.org/10.1016/j.molcel.2017.01.023>

SUMMARY

Single-cell RNA sequencing (scRNA-seq) offers new possibilities to address biological and medical questions. However, systematic comparisons of the performance of diverse scRNA-seq protocols are lacking. We generated data from 583 mouse embryonic stem cells to evaluate six prominent scRNA-seq methods: CEL-seq2, Drop-seq, MARS-seq, SCRB-seq, Smart-seq, and Smart-seq2. While Smart-seq2 detected the most genes per cell and across cells, CEL-seq2, Drop-seq, MARS-seq, and SCRB-seq quantified mRNA levels with less amplification noise due to the use of unique molecular identifiers (UMIs). Power simulations at different sequencing depths showed that Drop-seq is more cost-efficient for transcriptome quantification of large numbers of cells, while MARS-seq, SCRB-seq, and Smart-seq2 are more efficient when analyzing fewer cells. Our quantitative comparison offers the basis for an informed choice among six prominent scRNA-seq methods, and it provides a framework for benchmarking further improvements of scRNA-seq protocols.

INTRODUCTION

Genome-wide quantification of mRNA transcripts is highly informative for characterizing cellular states and molecular circuitries (ENCODE Project Consortium, 2012). Ideally, such data are collected with high spatial resolution, and single-cell RNA sequencing (scRNA-seq) now allows for transcriptome-wide analyses of individual cells, revealing exciting biological and medical insights (Kolodziejczyk et al., 2015a; Wagner et al., 2016). scRNA-seq requires the isolation and lysis of single cells, the conversion of their RNA into cDNA, and the amplification of cDNA to generate high-throughput sequencing libraries. As the

amount of starting material is so small, this process results in substantial technical variation (Kolodziejczyk et al., 2015a; Wagner et al., 2016).

One type of technical variable is the sensitivity of a scRNA-seq method (i.e., the probability to capture and convert a particular mRNA transcript present in a single cell into a cDNA molecule present in the library). Another variable of interest is the accuracy (i.e., how well the read quantification corresponds to the actual concentration of mRNAs), and a third type is the precision with which this amplification occurs (i.e., the technical variation of the quantification). The combination of sensitivity, precision, and number of cells analyzed determines the power to detect relative differences in expression levels. Finally, the monetary cost to reach a desired level of power is of high practical relevance. To make a well-informed choice among available scRNA-seq methods, it is important to quantify these parameters comparably. Some strengths and weaknesses of different methods are already known. For example, it has previously been shown that scRNA-seq conducted in the small volumes available in the automated microfluidic platform from Fluidigm (C1 platform) outperforms CEL-seq2, Smart-seq, or other commercially available kits in microliter volumes (Hashimshony et al., 2016; Wu et al., 2014). Furthermore, the Smart-seq protocol has been optimized for sensitivity, more even full-length coverage, accuracy, and cost (Picelli et al., 2013), and this improved Smart-seq2 protocol (Picelli et al., 2014b) has also become widely used (Gokce et al., 2016; Reinius et al., 2016; Tirosh et al., 2016).

Other protocols have sacrificed full-length coverage in order to sequence part of the primer used for cDNA generation. This enables early barcoding of libraries (i.e., the incorporation of cell-specific barcodes), allowing for multiplexing the cDNA amplification and thereby increasing the throughput of scRNA-seq library generation by one to three orders of magnitude (Hashimshony et al., 2012; Jaitin et al., 2014; Klein et al., 2015; Macosko et al., 2015; Soumillon et al., 2014). Additionally, this approach allows the incorporation of unique molecular identifiers (UMIs), random nucleotide sequences that tag individual

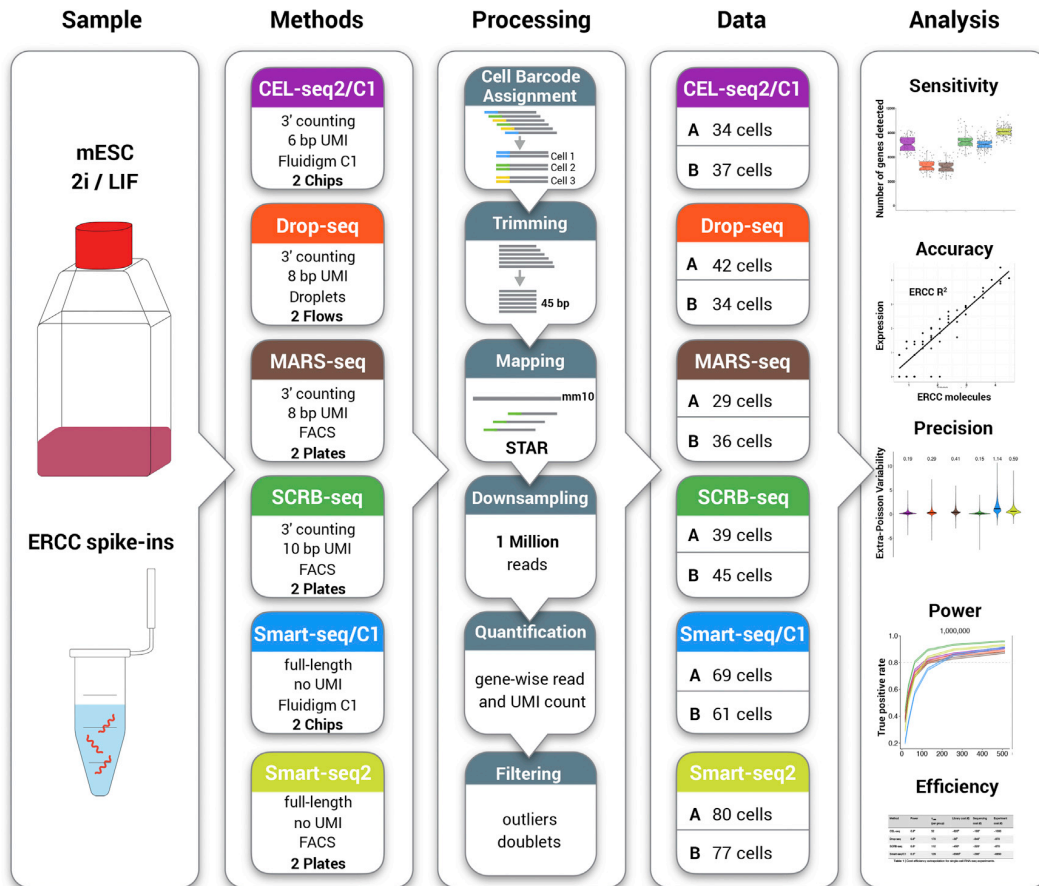


Figure 1. Schematic Overview of the Experimental and Computational Workflow

Mouse embryonic stem cells (mESCs) cultured in 2i/LIF and ERCC spike-in RNAs were used to generate single-cell RNA-seq data with six different library preparation methods (CEL-seq2/C1, Drop-seq, MARS-seq, SCRIB-seq, Smart-seq/C1, and Smart-seq2). The methods differ in the usage of unique molecular identifier (UMI) sequences, which allow the discrimination between reads derived from original mRNA molecules and duplicates generated during cDNA amplification. Data processing was identical across methods, and the given cell numbers per method and replicate were used to compare sensitivity, accuracy, precision, power, and cost efficiency. The six scRNA-seq methods are denoted by color throughout the figures of this study as follows: purple, CEL-seq2/C1; orange, Drop-seq; brown, MARS-seq; green, SCRIB-seq; blue, Smart-seq; and yellow, Smart-seq2. See also Figures S1 and S2.

mRNA molecules and, hence, allow for the distinction between original molecules and amplification duplicates that derive from the cDNA or library amplification (Kivioja et al., 2011). Utilization of UMI information improves quantification of mRNA molecules (Grün et al., 2014; Islam et al., 2014), and it has been implemented in several scRNA-seq protocols, such as STRT (Islam et al., 2014), CEL-seq (Grün et al., 2014; Hashimshony et al., 2016), CEL-seq2 (Hashimshony et al., 2016), Drop-seq (Macosko et al., 2015), inDrop (Klein et al., 2015), MARS-seq (Jaitin et al., 2014), and SCRIB-seq (Soumillon et al., 2014).

However, a thorough and systematic comparison of relevant parameters across scRNA-seq methods is still lacking. To address this issue, we generated 583 scRNA-seq libraries from mouse embryonic stem cells (mESCs), using six different methods in two replicates, and we compared their sensitivity, accuracy, precision, power, and efficiency (Figure 1).

RESULTS

Generation of scRNA-Seq Libraries

Variation in gene expression as observed among single cells is caused by biological and technical variation (Kolodziejczyk et al., 2015a; Wagner et al., 2016). We used mESCs cultured under two inhibitor/leukemia inhibitory factor (2i/LIF) conditions to obtain a relatively homogeneous cell population (Grün et al., 2014; Kolodziejczyk et al., 2015b), so that biological variation was similar among experiments and, hence, we mainly compared technical variation. In addition, we spiked in 92 poly-adenylated synthetic RNA transcripts of known concentration designed by the External RNA Control Consortium (ERCCs) (Jiang et al., 2011). For all six tested scRNA-seq methods (Figure 2), we generated libraries in two independent replicates.

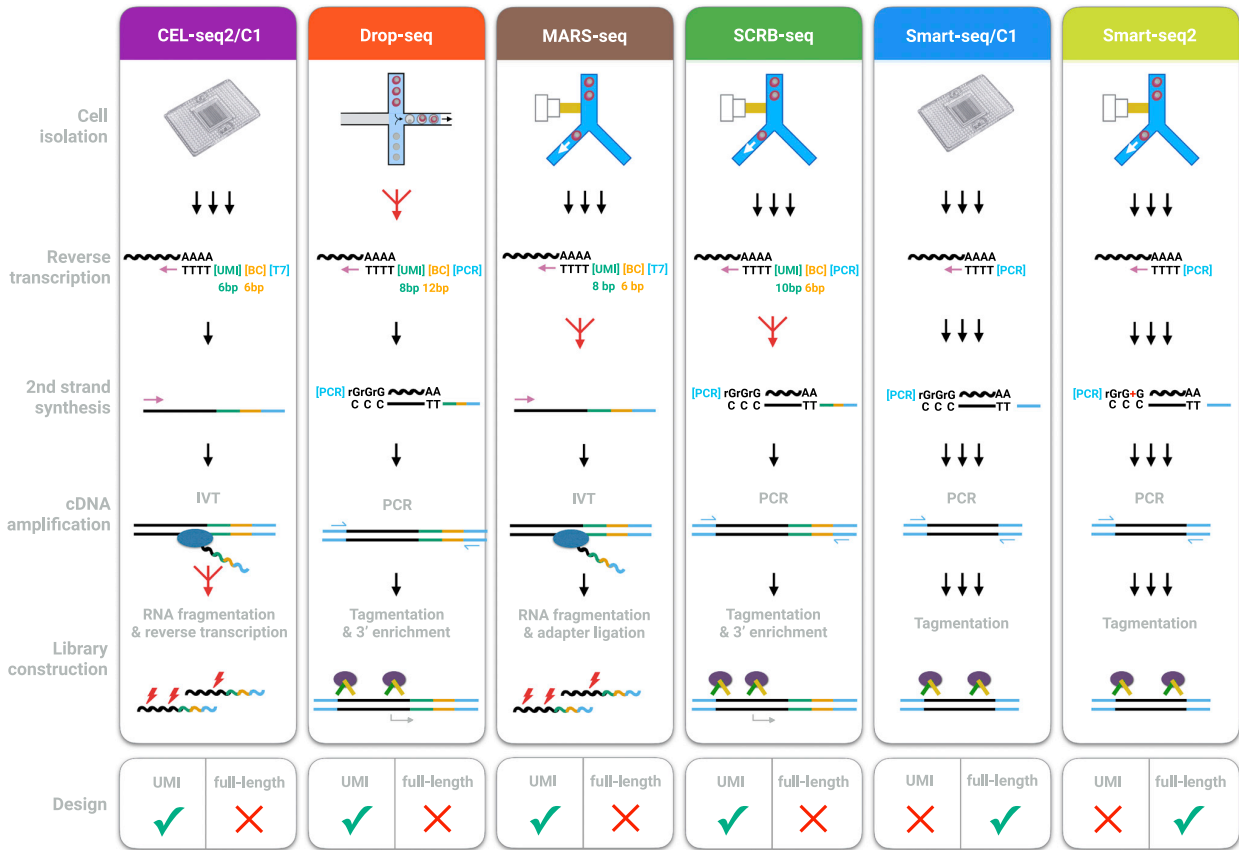


Figure 2. Schematic Overview of Library Preparation Steps

For details, see the text. See also Table S1.

For each replicate of the Smart-seq protocol, we performed one run on the C1 platform from Fluidigm (Smart-seq/C1) using microfluidic chips that automatically capture up to 96 cells (Wu et al., 2014). We imaged captured cells, added lysis buffer together with the ERCCs, and we used the commercially available Smart-seq kit (Clontech) to generate full-length double-stranded cDNA that we converted into 96 sequencing libraries by tagmentation (Nextera, Illumina).

For each replicate of the Smart-seq2 protocol, we sorted mESCs by fluorescence activated cell sorting (FACS) into 96-well PCR plates containing lysis buffer and the ERCCs. We generated cDNA as described (Picelli et al., 2013, 2014b), and we used an in-house-produced Tn5 transposase (Picelli et al., 2014a) to generate 96 libraries by tagmentation. While Smart-seq/C1 and Smart-seq2 are very similar protocols that generate full-length libraries, they differ in how cells are isolated, their reaction volume, and in that the Smart-seq2 chemistry has been systematically optimized (Picelli et al., 2013, 2014b). The main disadvantage of both Smart-seq protocols is that the generation of full-length cDNA libraries precludes an early barcoding step and the incorporation of UMIs.

For each replicate of the SCR-seq protocol (Soumillon et al., 2014), we also sorted mESCs by FACS into 96-well PCR plates

containing lysis buffer and the ERCCs. Similar to the Smart-seq protocols, cDNA was generated by oligo-dT priming, template switching, and PCR amplification of full-length cDNA. However, the oligo-dT primers contained well-specific (i.e., cell-specific) barcodes and UMIs. Hence, cDNA from one plate could be pooled and then converted into sequencing libraries, using a modified tagmentation approach that enriches for the 3' ends. SCR-seq is optimized for small volumes and few handling steps.

The fourth method evaluated was Drop-seq, a recently developed microdroplet-based approach (Macosko et al., 2015). Here a flow of beads suspended in lysis buffer and a flow of a single-cell suspension were brought together in a microfluidic chip that generated nanoliter-sized emulsion droplets. On each bead, oligo-dT primers carrying a UMI and a unique, bead-specific barcode were covalently bound. Cells were lysed within these droplets, their mRNA bound to the oligo-dT-carrying beads, and, after breaking the droplets, cDNA and library generation was performed for all cells in parallel in one single tube. The ratio of beads to cells (20:1) ensured that the vast majority of beads had either no cell or one cell in its droplet. Hence, similar to SCR-seq, each cDNA molecule was labeled with a bead-specific (i.e., cell-specific) barcode and a UMI. We confirmed that

the Drop-seq protocol worked well in our setup by mixing mouse and human T cells, as recommended by [Macosko et al. \(2015\)](#) ([Figure S1A](#)). The main advantage of the protocol is that a high number of scRNA-seq libraries can be generated at low cost. One disadvantage of Drop-seq is that the simultaneous inclusion of ERCC spike-ins is quite expensive, as their addition would generate cDNA from ERCCs also in beads that have zero cells and thus would double the sequencing costs. As a proxy for the missing ERCC data, we used a published dataset ([Macosko et al., 2015](#)), where ERCC spike-ins were sequenced using the Drop-seq method without single-cell transcriptomes.

As a fifth method we chose CEL-seq2 ([Hashimshony et al., 2016](#)), an improved version of the original CEL-seq ([Hashimshony et al., 2012](#)) protocol, as implemented for microfluidic chips on Fluidigm's C1 ([Hashimshony et al., 2016](#)). As for Smart-seq/C1, this allowed us to capture 96 cells in two independent replicates and to include ERCCs in the cell lysis step. Similar to Drop-seq and SCR-seq, cDNA was tagged with barcodes and UMIs; but, in contrast to the four PCR-based methods described above, CEL-seq2 relies on linear amplification by *in vitro* transcription after the initial reverse transcription. The amplified, bar-coded RNAs were harvested from the chip, pooled, fragmented, and reverse transcribed to obtain sequencing libraries.

MARS-seq, the sixth method evaluated, is a high-throughput implementation of the original CEL-seq method ([Jaitin et al., 2014](#)). In this protocol, cells were sorted by FACS in 384-well plates containing lysis buffer and the ERCCs. As in CEL-seq and CEL-seq2, amplified RNA with barcodes and UMIs were generated by *in vitro* transcription, but libraries were prepared on a liquid-handling platform. An overview of the methods and their workflows is provided in [Figure 2](#) and in [Table S1](#).

Processing of scRNA-Seq Data

For each method, we generated at least 48 libraries per replicate and sequenced between 241 and 866 million reads ([Figure 1](#); [Figure S1B](#)). All data were processed identically, with cDNA reads clipped to 45 bp and mapped using Spliced Transcripts Alignment to a Reference (STAR) ([Dobin et al., 2013](#)) and UMIs quantified using the Drop-seq pipeline ([Macosko et al., 2015](#)). To adjust for differences in sequencing depths, we selected all libraries with at least one million reads, and we downsampled them to one million reads each. This resulted in 96, 79, 73, 93, 162, and 187 libraries for CEL-seq2/C1, Drop-seq, MARS-seq, SCR-seq, Smart-seq/C1, and Smart-seq2, respectively.

To exclude doublets (libraries generated from two or more cells) in the Smart-seq/C1 data, we analyzed microscope images and identified 16 reaction chambers with multiple cells. For the four UMI methods, we calculated the number of UMIs per library, and we found that libraries that have more than twice the mean total UMI count can be readily identified ([Figure S1C](#)). It is unclear whether these libraries were generated from two separate cells (doublets) or, for example, from one large cell before mitosis. However, for the purpose of this method comparison, we removed these three to nine libraries. To filter out low-quality libraries, we used a method that exploits the fact that transcript detection and abundance in low-quality libraries correlate poorly with high-quality libraries as well as with other low-quality libraries ([Petropoulos et al., 2016](#)). Therefore, we determined

the maximum Spearman correlation coefficient for each cell in all-to-all comparisons that allowed us to identify low-quality libraries as outliers of the distributions of correlation coefficients by visual inspection ([Figure S1D](#)). This filtering led to the removal of 21, 0, 4, 0, 16, and 30 cells for CEL-seq2/C1, Drop-seq, MARS-seq, SCR-seq, Smart-seq/C1, and Smart-seq2, respectively.

In summary, we processed and filtered our data so that we ended up with a total of 583 high-quality scRNA-seq libraries that could be used for a fair comparison of the sensitivity, accuracy, precision, power, and efficiency of the methods.

Single-Cell Libraries Are Sequenced to a Reasonable Level of Saturation at One Million Reads

For all six methods, >50% of the reads could be unambiguously mapped to the mouse genome ([Figure 3A](#)), which is comparable to previous results ([Jaitin et al., 2014](#); [Wu et al., 2014](#)). Overall, between 48% (Smart-seq2) and 30% (Smart-seq/C1) of all reads were exonic and, thus, were used to quantify gene expression levels. However, the UMI data showed that only 14%, 5%, 7%, and 15% of the exonic reads were derived from independent mRNA molecules for CEL-seq2/C1, Drop-seq, MARS-seq, and SCR-seq, respectively ([Figure 3A](#)). To quantify the relationship between the number of detected genes or mRNA molecules and the number of reads in more detail, we downsampled reads to varying depths, and we estimated to what extent libraries were sequenced to saturation ([Figure S2](#)). The number of unique mRNA molecules plateaued at 56,760 UMIs per library for CEL-seq2/C1 and 26,210 UMIs per library for MARS-seq, was still marginally increasing at 17,210 UMIs per library for Drop-seq, and was considerably increasing at 49,980 UMIs per library for SCR-seq ([Figure S2C](#)). Notably, CEL-seq2/C1 and MARS-seq showed a steeper slope at low sequencing depths than both Drop-seq and SCR-seq, potentially due to a less biased amplification by *in vitro* transcription. Hence, among the UMI methods, CEL-seq2/C1 and SCR-seq libraries had the highest complexity of mRNA molecules, and this complexity was sequenced to a reasonable level of saturation with one million reads.

To investigate saturation also for non-UMI-based methods, we applied a similar approach at the gene level by counting the number of genes detected by at least one read. By fitting an asymptote to the downsampled data, we estimated that ~90% (Drop-seq and SCR-seq) to 100% (CEL-seq2/C1, MARS-seq, Smart-Seq/C1, and Smart-seq2) of all genes present in a library were detected at one million reads ([Figure 3B](#); [Figure S2A](#)). In particular, the deep sequencing of Smart-seq2 libraries showed clearly that the number of detected genes did not change when increasing the sequencing depth from one million to five million reads per cell ([Figure S2B](#)).

All in all, these analyses show that scRNA-seq libraries were sequenced to a reasonable level of saturation at one million reads, a cutoff that also has been suggested previously for scRNA-seq datasets ([Wu et al., 2014](#)). While it can be more efficient to invest in more cells at lower coverage (see our power analyses below), one million reads per cell is a reasonable sequencing depth for our purpose of comparing scRNA-seq methods.

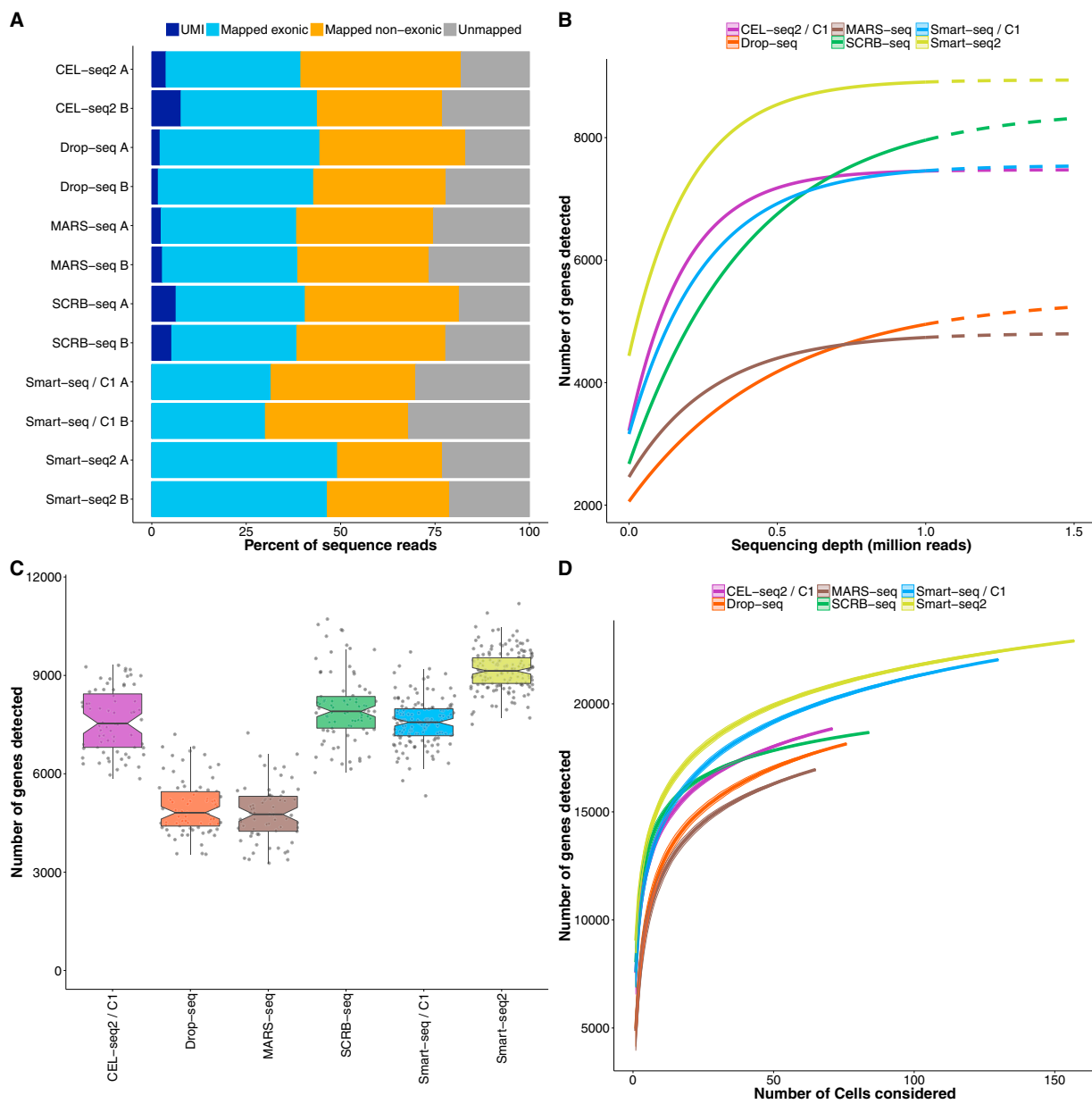


Figure 3. Sensitivity of scRNA-Seq Methods

(A) Percentage of reads (downsampled to one million per cell) that cannot be mapped to the mouse genome (gray) are mapped to regions outside exons (orange) or inside exons (blue). For UMI methods, dark blue denotes the exonic reads with unique UMIs.

(B) Median number of genes detected per cell (counts ≥ 1) when downsampling total read counts to the indicated depths. Dashed lines above one million reads represent extrapolated asymptotic fits.

(C) Number of genes detected (counts ≥ 1) per cell. Each dot represents a cell and each box represents the median and first and third quartiles per replicate and method.

(D) Cumulative number of genes detected as more cells are added. The order of cells considered was drawn randomly 100 times to display mean \pm SD (shaded area). See also [Figures S3 and S4](#).

Smart-Seq2 Has the Highest Sensitivity

Taking the number of detected genes per cell as a measure of sensitivity, we found that Drop-seq and MARS-seq had the lowest

sensitivity, with a median of 4,811 and 4,763 genes detected per cell, respectively, while CEL-seq2/C1, SCRB-seq, and Smart-seq/C1 detected a median of 7,536, 7,906, and 7,572 genes per

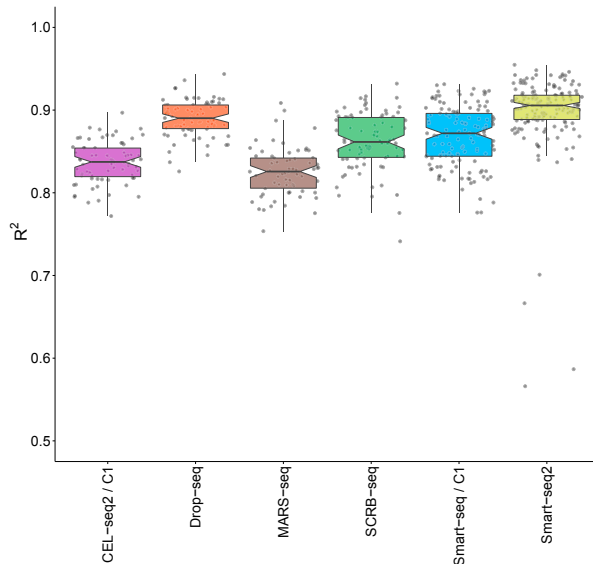


Figure 4. Accuracy of scRNA-Seq Methods

ERCC expression values (counts per million reads for Smart-seq/C1 and Smart-seq2 and UMIs per million reads for all others) were correlated to their annotated molarity. Shown are the distributions of correlation coefficients (adjusted R^2 of linear regression model) across methods. Each dot represents a cell/bead and each box represents the median and first and third quartiles. See also Figure S5.

cell (Figure 3C). Smart-seq2 detected the highest number of genes per cell with a median of 9,138. To compare the total number of genes detected across many cells, we pooled the sequence data of 65 cells per method, and we detected $\sim 19,000$ genes for CEL-Seq2/C1, $\sim 17,000$ for MARS-seq, $\sim 18,000$ for Drop-seq and SCRIB-seq, $\sim 20,000$ for Smart-seq/C1, and $\sim 21,000$ for Smart-seq2 (Figure 3D). While the majority of genes ($\sim 13,000$) were detected by all methods, ~ 400 genes were specific to each of the 3' counting methods, and $\sim 1,000$ genes were specific to each of the two full-length methods (Figure S3A). This higher sensitivity of both full-length methods also was apparent when plotting the genes detected in all available cells, as the 3' counting methods leveled off below 20,000 genes while the two full-length methods leveled off above 20,000 genes (Figure 3D). Such a difference could be caused by genes that have 3' ends that are difficult to map. However, we found that genes specific to Smart-seq2 and Smart-seq/C1 map as well to 3' ends as genes with similar length distribution that are not specifically detected by full-length methods (Figure S3B). Hence, it seems that full-length methods turn a slightly higher fraction of transcripts into sequenceable molecules than 3' counting methods and are more sensitive in this respect. Importantly, method-specific genes are detected in very few cells (87% of genes occur in one or two cells) with very low counts (mean counts < 0.2 , Figure S3C). This suggests that they are unlikely to remain method specific at higher expression levels and that their impact on conclusions drawn from scRNA-seq data is rather limited (Lun et al., 2016).

Next, we investigated how reads are distributed along the mRNA transcripts for all genes. As expected, the 3' counting

methods showed a strong bias of reads mapped to the 3' end (Figure S3D). However, it is worth mentioning that a considerable fraction of reads also covered other segments of the transcripts, probably due to internal oligo-dT priming (Nam et al., 2002). Smart-seq2 showed a more even coverage than Smart-seq, confirming previous findings (Picelli et al., 2013). A general difference in expression values between 3' counting and full-length methods also was reflected in their strong separation by the first principal component, explaining 37% of the total variance, and when taking into account that one needs to normalize for gene length for the full-length methods (Figure S4E).

As an absolute measure of sensitivity, we compared the probability of detecting the 92 spiked-in ERCCs, for which the number of molecules available for library construction is known (Figures S4A and S4B). We determined the detection probability of each ERCC RNA as the proportion of cells with at least one read or UMI count for the particular ERCC molecule (Marinov et al., 2014). For Drop-seq, we used the previously published ERCC-only dataset (Macosko et al., 2015), and for the other five methods, 2%–5% of the one million reads per cell mapped to ERCCs that were sequenced to complete saturation at that level (Figure S5B). A 50% detection probability was reached at $\sim 7, 11, 14, 16, 17,$ and 28 ERCC molecules for Smart-seq2, Smart-seq/C1, CEL-seq2/C1, SCRIB-seq, Drop-seq, and MARS-seq, respectively (Figure S4C). Notably, the sensitivity estimated from the number of detected genes does not fully agree with the comparison based on ERCCs. While Smart-seq2 was the most sensitive method in both cases, Drop-seq performed better and SCRIB-seq and MARS-seq performed worse when using ERCCs. The separate generation and sequencing of the Drop-seq ERCC libraries could be a possible explanation for their higher sensitivity. However, it remains unclear why SCRIB-seq and MARS-seq had a substantially lower sensitivity when using ERCCs. It has been noted before that ERCCs can be problematic for modeling endogenous mRNAs (Risso et al., 2014), potentially due to their shorter length, shorter poly-A tail, and their missing 5' cap (Grün and van Oudenaarden, 2015; Stegle et al., 2015). While ERCCs are still useful to gauge the absolute range of sensitivities, the thousands of endogenous mRNAs are likely to be a more reliable estimate for comparing sensitivities as we used the same cell type for all methods.

In summary, we find that Smart-seq2 is the most sensitive method, as it detects the highest number of genes per cell and the most genes in total across cells and has the most even coverage across transcripts. Smart-seq/C1 is slightly less sensitive per cell and detects almost the same number of genes across cells with slightly less even coverage. Among the 3' counting methods, CEL-seq2/C1 and SCRIB-seq detect about as many genes per cell as Smart-seq/C1, whereas Drop-seq and MARS-seq detect considerably fewer genes.

Accuracy of scRNA-Seq Methods

To measure the accuracy of transcript level quantifications, we compared the observed expression values (counts per million or UMIs per million) with the known concentrations of the 92 ERCC transcripts (Figure S5A). For each cell, we calculated the coefficient of determination (R^2) for a linear model fit (Figure 4). Methods differed significantly in their accuracy (Kruskal-Wallis

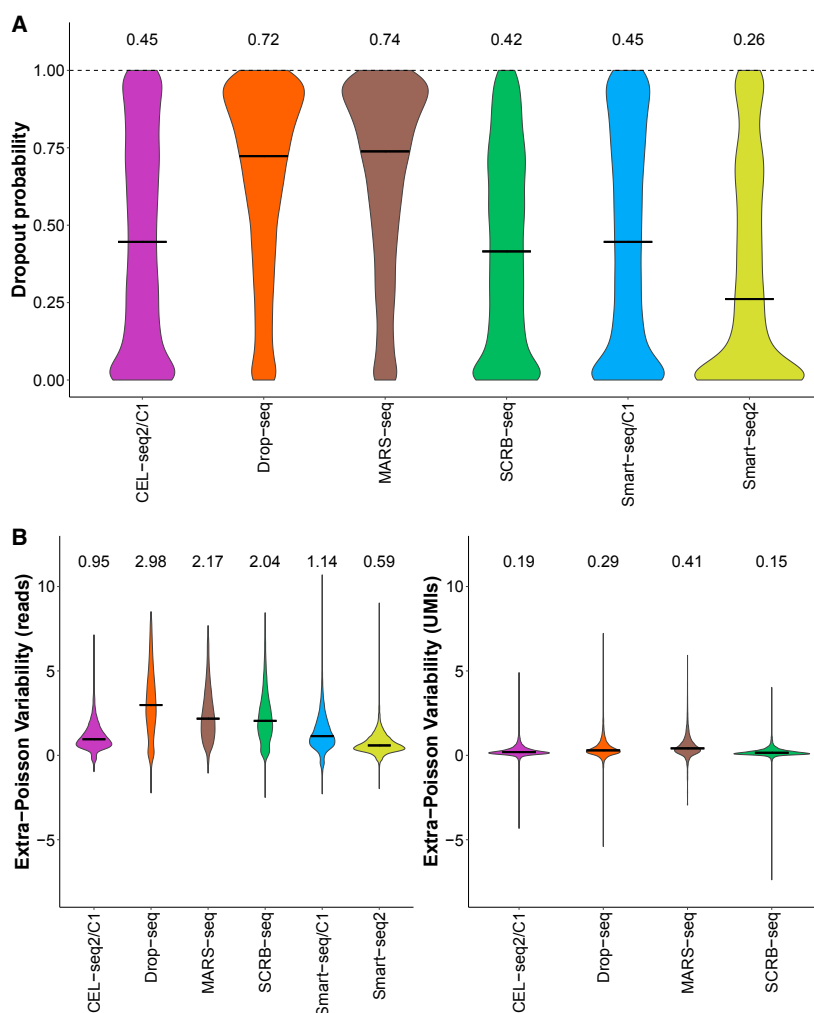


Figure 5. Precision of scRNA-Seq Methods

We compared precision among methods using the 13,361 genes detected in at least 25% of all cells by any method in a subsample of 65 cells per method.

(A) Distributions of dropout rates across the 13,361 genes are shown as violin plots, and medians are shown as bars and numbers.

(B) Extra Poisson variability across the 13,361 genes was calculated by subtracting the expected amount of variation due to Poisson sampling (square root of mean divided by mean) from the CV (SD divided by mean). Distributions are shown as violin plots and medians are shown as bars and numbers. For 349, 336, 474, 165, 201, and 146 genes for CEL-seq2/C1, Drop-seq, MARS-seq, SCR-seq, Smart-seq/C1, and Smart-seq2, respectively, no extra Poisson variability could be calculated. See also [Figures S6](#) and [S7](#).

test, $p < 2.2e-16$), but all methods had a fairly high R^2 ranging between 0.83 (MARS-seq) and 0.91 (Smart-seq2). This suggests that, for all methods, transcript concentrations across this broad range can be predicted fairly well from expression values. As expected, accuracy was worse for narrower and especially for lower concentration ranges ([Figure S5C](#)). It is worth emphasizing that the accuracy assessed here refers to absolute expression levels across genes within cells. This accuracy can be important, for example, to identify marker genes with a high absolute mRNA expression level. However, the small differences in accuracy seen here will rarely be a decisive factor when choosing among the six protocols.

Precision of Amplified Genes Is Strongly Increased by UMIs

While a high accuracy is necessary to compare absolute expression levels, one of the most common experimental aims is to compare relative expression levels to identify differentially expressed genes or different cell types. Hence, the precision (i.e.,

the reproducibility of the expression-level estimate) is a major factor when choosing a method. As we used the same cell type under the same culture conditions for all methods, the amount of biological variation should be the same in the cells analyzed by each of the six methods. Hence, we can assume that differences in the total variation among methods are due to differences in their technical variation. Technical variation is substantial in scRNA-seq data primarily because a substantial fraction of mRNAs is lost during cDNA generation and small amounts of cDNA get amplified. Therefore, both the dropout probability and the amplification noise need to be considered when quantifying variation.

Indeed, a mixture model including a dropout probability and a negative binomial distribution, modeling the overdispersion in the count data, have been shown to represent scRNA-seq data better than the negative binomial alone ([Finak et al., 2015](#); [Kharchenko et al., 2014](#)).

To compare precision without penalizing more sensitive methods, we selected a common set of 13,361 genes that were detected in 25% of the cells by at least one method ([Figure S6A](#)). We then analyzed these genes in a subsample of 65 cells per method to avoid a bias due to unequal numbers of cells. We estimated the dropout probability as the fraction of cells with zero counts ([Figure 5A](#); [Figure S6B](#)). As expected from the number of detected genes per cell ([Figure 3C](#)), MARS-seq had the highest median dropout probability (74%) and Smart-seq2 had the lowest (26%) ([Figure 5A](#)). To estimate the amplification noise of detected genes, we calculated the coefficient of variation (CV, SD divided by the mean, including zeros), and we subtracted the expected amount of variation due to Poisson sampling (i.e., the square root of the mean divided by the mean). This was possible

for 96.5% (MARS-seq) to 98.9% (Smart-seq2) of all the 13,361 genes. This extra Poisson variability includes biological variation (assumed to be the same across methods in our data) and technical variation, and the latter includes noise introduced by amplification (Brennecke et al., 2013; Grün et al., 2014; Stegle et al., 2015). That amplification noise can be a major factor is seen by the strong increase of extra Poisson variability when ignoring UMIs and considering read counts only (Figure 5B, left; Figure S7A). This is expected, as UMIs should remove amplification noise, which has been described previously for CEL-seq (Grün et al., 2014). For SCR-seq and Drop-seq, which are PCR-based methods, UMIs removed even more extra Poisson variability than for CEL-seq2/C1 and MARS-seq (Figure 5B), which is in line with the notion that amplification by PCR is more noisy than amplification by *in vitro* transcription. Of note, Smart-seq2 had the lowest amplification noise when just considering reads (Figure 5B, left), potentially because its higher sensitivity requires less amplification and, hence, leads to less noise.

In summary, Smart-seq2 detects the common set of 13,361 genes in more cells than the UMI methods, but it has, as expected, more amplification noise than the UMI-based methods. How the different combinations of dropout rate and amplification noise affect the power of the methods is not evident, neither from this analysis nor from the total coefficient of variation that ignores the strong mean variance and mean dropout dependencies of scRNA-seq data (Figure S7B).

Power Is Determined by a Combination of Dropout Rates and Amplification Noise and Is Highest for SCR-seq

To estimate the combined impact of sensitivity and precision on the power to detect differential gene expression, we simulated scRNA-seq data given the observed dropout rates and variance for the 13,361 genes. As these depend strongly on the expression level of a gene, it is important to retain the mean variance and mean dropout relationships. To this end, we estimated the mean, the variance (i.e., the dispersion parameter of the negative binomial distribution), and the dropout rate for each gene and method. We then fitted a cubic smoothing spline to the resulting pairs of mean and dispersion estimates to predict the dispersion of a gene given its mean (Figure S8A). Furthermore, we applied a local polynomial regression model to account for the dropout probability given a gene's mean expression (Figure S8B). When simulating data according to these fits, we recovered distributions of dropout rates and variance closely matching the observed data (Figures S8C and S8D). To compare the power for differential gene expression among the methods, we simulated read counts for two groups of n cells and added log-fold changes to 5% of the 13,361 genes in one group. To mimic a biologically realistic scenario, these log-fold changes were drawn from observed differences between microglial subpopulations from a previously published dataset (Zeisel et al., 2015). Simulated datasets were tested for differential expression using limma (Ritchie et al., 2015), and the true positive rate (TPR) and the false discovery rate (FDR) were calculated. Of note, this does include undetected genes, i.e., the 2.5% (SCR-seq) to 6.8% (MARS-seq) of the 13,361 genes that had fewer than two measurements in a particular method (Figure S6B) and for which we could not estimate the variance. In our simulations, these

genes could be drawn as differentially expressed, and in our TPR they were then counted as false negatives for the particular method. Hence, our power simulation framework considers the full range of dropout rates and is not biased against more sensitive methods.

First, we analyzed how the number of cells affects TPR and FDR by running 100 simulations each for a range of 16 to 512 cells per group (Figure 6A). FDRs were similar in all methods ranging from 3.9% to 8.7% (Figure S9A). TPRs differed considerably among methods and SCR-seq performed best, reaching a median TPR of 80% with 64 cells. CEL-seq2/C1, Drop-seq, MARS-seq, and Smart-seq2 performed slightly worse, reaching 80% power with 86, 99, 110, and 95 cells per group, respectively, while Smart-seq/C1 needed 150 cells to reach 80% power (Figure 6A). When disregarding UMIs, Smart-seq2 performed best (Figure 6B), as expected from its low dropout rate and its low amplification noise when considering reads only (Figure 5B). Furthermore, power dropped especially for Drop-seq and SCR-seq (Figure 6B), as expected from the strong increase in amplification noise of these two methods when considering reads only (Figure 5B). When we stratified our analysis (considering UMIs) across five bins of expression levels, the ranking of methods was recapitulated and showed that the lowest expression bin strongly limited the TPR in all methods (Figure S9B). This ranking also was recapitulated when we analyzed a set of 19 genes previously reported to contain cell-cycle variation in the 2i/LIF culture condition (Kolodziejczyk et al., 2015b). The variance of these cell-cycle genes was clearly higher than the variance of 19 pluripotency and housekeeping (ribosomal) genes in all methods. The p value of that difference was lowest for SCR-seq, the most powerful method, and highest for Smart-seq/C1, the least powerful method (Figure S10D).

Notably, this power analysis, as well as the sensitivity, accuracy, and precision parameters analyzed above, includes the variation that is generated in the two technical replicates (batches) per method that we performed (Figure 1). These estimates were very similar among our technical replicates, and, hence, our method comparison is valid with respect to batch variations (Figures S10B–S10D). In addition, as batch effects are known to be highly relevant for interpreting scRNA-seq data (Hicks et al., 2015), we gauged the magnitude of batch effects with respect to identifying differentially expressed genes. To this end, we used limma to identify differentially expressed genes between batches (FDR < 1%), using 25 randomly selected cells per batch and method. All methods had significantly more genes differentially expressed between batches than expected from permutations (zero to four genes), with a median of 119 (Drop-seq) to ~1,135 (CEL-seq2/C1) differentially expressed genes (Figure S10A). Notably, genes were affected at random across methods, as there was no significant overlap among them (extended hypergeometric test [Kalinka, 2013], $p > 0.84$). Hence, this analysis once more emphasizes that batches are important to consider in the design of scRNA-seq experiments (Hicks et al., 2015). While a quantitative comparison of the magnitude of batch effects among methods would require substantially more technical replicates per method, the methods differ in their flexibility to incorporate batch effect into the experimental design, which is an important aspect to consider as discussed below.

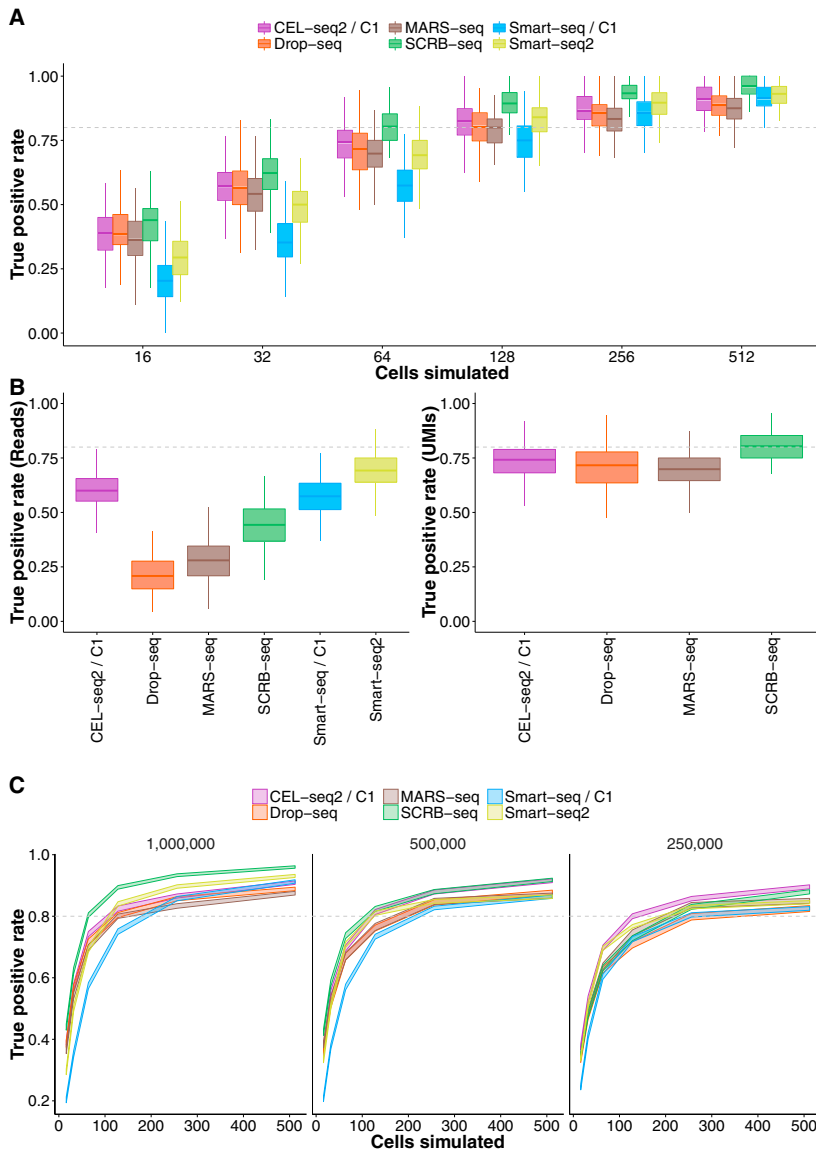


Figure 6. Power of scRNA-Seq Methods

Using the empirical mean/dispersion and mean/dropout relationships (Figures S8A and S8B), we simulated data for two groups of n cells each for which 5% of the 13,361 genes were differentially expressed, with log-fold changes drawn from observed differences between microglial subpopulations from a previously published dataset (Zeisel et al., 2015). The simulated data were then tested for differential expression using limma (Ritchie et al., 2015), from which the average true positive rate (TPR) and the average false discovery rate (FDR) were calculated (Figure S9A).

(A) TPR for one million reads per cell for sample sizes $n = 16$, $n = 32$, $n = 64$, $n = 128$, $n = 256$, and $n = 512$ per group. Boxplots represent the median and first and third quartiles of 100 simulations.

(B) TPR for one million reads per cell for $n = 64$ per group with and without using UMI information. Boxplots represent the median and first and third quartiles of 100 simulations.

(C) TPRs as in (A) using mean/dispersion and mean/dropout estimates from one million (as in A), 0.5 million, and 0.25 million reads. Line areas indicate the median power with SE from 100 simulations. See also Figures S8–S10 and Table 1.

including the scientific questions addressed, the experimental design, or the sample availability. However, the monetary cost is certainly an important one, and we used the results of our simulations to compare the costs among the methods for a given level of power.

Cost Efficiency Is Similarly High for Drop-Seq, MARS-Seq, SCR-Seq, and Smart-Seq2

Given the number of cells needed to reach 80% power as simulated above for three sequencing depths (Figure 6C), we calculated the minimal costs to generate and sequence these libraries.

For example, at a sequencing depth of one million reads, SCR-seq requires 64 cells per group to reach 80% power. Generating 128 SCR-seq libraries costs ~260\$ and generating 128 million reads costs ~640\$. Note that the necessary paired-end reads for CEL-seq2/C1, SCR-seq, MARS-seq, and Drop-seq can be generated using a 50-cycle sequencing kit, and, hence, we assume that sequencing costs are the same for all methods.

Calculating minimal costs this way, Drop-seq (690\$) is the most cost-effective method when sequencing 254 cells at a depth of 250,000 reads, and SCR-seq (810\$), MARS-seq (820\$), and Smart-seq2 (1,090\$) are slightly more expensive at the same performance (Table 1). For Smart-seq2 it should be stressed that the use of in-house-produced Tn5 transposase (Picelli et al., 2014a) is required to keep the cost at this level, as

As a next step, we analyzed how the performance of the six methods depends on sequencing depth. To this end, we performed power simulations as above, but we estimated the mean dispersion and mean dropout relationships from data downsampled to 500,000 or 250,000 reads per cell. Overall, the decrease in power was moderate (Figure 6C; Table 1) and followed the drop in sensitivity at different sequencing depths (Figure 3B). While Smart-seq2 and CEL-seq2/C1 needed just 1.3-fold more cells at 0.25 million reads than at one million reads to reach 80% power, SCR-seq and Drop-seq required 2.6-fold more cells (Table 1). In summary, SCR-seq is the most powerful method at one million reads and half a million reads, but CEL-seq2/C1 is the most powerful method at a sequencing depth of 250,000 reads. The optimal balance between the number of cells and their sequencing depth depends on many factors,

Table 1. Cost Efficiency Extrapolation for Single-Cell RNA-Seq Experiments

Method	TPR ^a	FDR ^a (%)	Cell per Group ^b	Library Cost (\$)	Minimal Cost ^c (\$)
CEL-seq2/C1	0.8	~6.1	86/100/110	~9	~2,420/2,310/2,250
Drop-seq	0.8	~8.4	99/135/254	~0.1	~1,010/700/690
MARS-seq	0.8	~7.3	110/135/160	~1.3	~1,380/1,030/820
SCRB-seq	0.8	~6.1	64/90/166	~2	~900/810/1,080
Smart-seq/C1	0.8	~4.9	150/172/215	~25	~9,010/9,440/11,290
Smart-seq2 (commercial)	0.8	~5.2	95/105/128	~30	~10,470/11,040/13,160
Smart-seq2 (in-house Tn5)	0.8	~5.2	95/105/128	~3	~1,520/1,160/1,090

See also [Figure 6](#).

^aTrue positive rate and false discovery rate are based on simulations ([Figure 6](#); [Figure S9](#)).

^bSequencing depth of one, 0.5, and 0.25 million reads.

^cAssuming \$5 per one million reads.

was done in our experiments. When instead using the Tn5 transposase of the commercial Nextera kit as described ([Picelli et al., 2014b](#)), the costs for Smart-seq2 are 10-fold higher. Even if one reduces the amount of Nextera transposase to a quarter, as done in the Smart-seq/C1 protocol, the Smart-seq2 protocol is still four times more expensive than the early barcoding methods. CEL-seq2/C1 is fairly expensive due to the microfluidic chips that make up 69% of the library costs, and Smart-seq/C1 is almost 13-fold less efficient than Drop-seq due to its high library costs that arise from the microfluidic chips, the commercial Smart-seq kit, and the costs for commercial Nextera XT kits.

Of note, these calculations are the minimal costs of the experiment and several factors are not considered, such as labor costs, costs to set up the methods, costs to isolate cells of interest, or costs due to practical constraints in generating a fixed number of scRNA-seq libraries with a fixed number of reads. In many experimental settings, independent biological and/or technical replicates are needed when investigating particular factors, such as genotypes or developmental time points, and Smart-seq/C1, CEL-seq2/C1, and Drop-seq are less flexible in distributing scRNA-seq libraries across replicates than the other three methods that use PCR plates. Furthermore, the costs are increased by unequal sampling from the included cells as well as from sequencing reads from cells that are excluded. In our case, between 6% (SCRB-seq) and 32% (Drop-seq) of the reads came from cell barcodes that were not included. While it is difficult to exactly calculate and compare these costs among methods, it is clear that they will increase the costs for Drop-seq relatively more than for the other methods. In summary, we find that Drop-seq, SCRБ-seq, and MARS-seq are the most cost-effective methods, closely followed by Smart-seq2, if using an in-house-produced transposase.

DISCUSSION

Here we have provided an in-depth comparison of six prominent scRNA-seq protocols. To this end, we generated data for all six compared methods from the same cells, cultured under the same condition in the same laboratory. While there would be many more datasets and methods for a comparison of the sensitivity and accuracy of the ERCCs ([Svensson et al., 2016](#)), our approach provides a more controlled and comprehensive com-

parison across thousands of endogenous genes. This is important, as can be seen by the different sensitivity estimates that we obtained for Drop-seq, MARS-seq, and SCRБ-seq using the ERCCs. In our comparison, we clearly find that Smart-seq2 is the most sensitive method, closely followed by SCRБ-seq, Smart-seq/C1, and CEL-seq2/C1, while Drop-seq and MARS-seq detect nearly 50% fewer genes per cell ([Figures 3B and 3C](#)). In addition, Smart-seq2 shows the most even read coverage across transcripts ([Figure S3D](#)), making it the most appropriate method for the detection of alternative splice forms and for analyses of allele-specific expression using SNPs ([Deng et al., 2014](#); [Reinius et al., 2016](#)). Hence, Smart-seq2 is certainly the most suitable method when an annotation of single-cell transcriptomes is the focus. Furthermore, we find that Smart-seq2 is also the most accurate method (i.e., it has the highest correlation of known ERCC spike-in concentrations and read counts per million), which is probably related to its higher sensitivity. Hence, differences in expression values across transcripts within the same cell predict differences in the actual concentrations of these transcripts well. All methods do this rather well, at least for higher expression levels, and we think that the small differences among methods will rarely be a decisive factor. Importantly, the accuracy of estimating transcript concentrations across cells (relevant, e.g., for comparing the total RNA content of cells) depends on different factors and cannot be compared well among the tested methods as it would require known concentration differences of transcripts across cells. However, it is likely that methods that can use UMIs and ERCCs (CEL-seq2/C1, MARS-seq, and SCRБ-seq) would have a strong advantage in this respect.

How well relative expression levels of the same genes can be compared across cells depends on two factors. First, how often (i.e., in how many cells and from how many molecules) it is measured. Second, with how much technical variation (i.e., with how much noise, e.g., from amplification) it is measured. For the first factor (dropout probability), we find Smart-seq2 to be the best method ([Figure 5A](#)), as expected from its high gene detection sensitivity. For the second factor (extra Poisson variability), we find the four UMI methods to perform better ([Figure 5B](#)), as expected from their ability to eliminate variation introduced by amplification. To assess the combined effect of these two factors, we performed simulations for differential gene

expression scenarios (Figure 6). This allowed us to translate the sensitivity and precision parameters into the practically relevant power to detect differentially expressed genes. Of note, our power estimates include the variation that is caused by the two different replicates per method that constitutes an important part of the variation. Our simulations show that, at a sequencing depth of one million reads, SCRIB-seq has the highest power, probably due to a good balance of high sensitivity and low amplification noise. Furthermore, amplification noise and power strongly depend on the use of UMIs, especially for the PCR-based methods (Figures 5B and 6B; Figure S7). Notably, this is due to the large amount of amplification needed for scRNA-seq libraries, as the effect of UMIs on power for bulk RNA-seq libraries is negligible (Parekh et al., 2016).

Perhaps practically most important, our power simulations also allow us to compare the efficiency of the methods by calculating the costs to generate the data for a given level of power. Using minimal cost calculations, we find that Drop-seq is the most cost-effective method, closely followed by SCRIB-seq, MARS-seq, and Smart-seq2. However, Drop-seq costs are likely to be more underestimated, due to lower flexibility in generating a specified number of libraries and the higher fraction of reads that come from bad cells. Hence, all four UMI methods are in practice probably similarly cost-effective. In contrast, for Smart-seq2 to be similarly cost-effective it is absolutely necessary to use in-house-produced transposase or to drastically reduce volumes of commercial transposase kits (Lamble et al., 2013; Mora-Castilla et al., 2016).

Given comparable efficiencies of Drop-seq, MARS-seq, SCRIB-seq, and Smart-seq2, additional factors will play a role when choosing a suitable method for a particular question. Due to its low library costs, Drop-seq is probably preferable when analyzing large numbers of cells at low coverage (e.g., to find rare cell types). On the other hand, Drop-seq in its current setup requires a relatively large amount of cells (>6,500 for 1 min of flow). Hence, if few and/or unstable cells are isolated by FACS, the SCRIB-seq, MARS-seq, or Smart-seq2 protocols are probably preferable. Additional advantages of these methods over Drop-seq include that technical variation can be estimated from ERCCs for each cell, which can be helpful to estimate biological variation (Kim et al., 2015; Vallejos et al., 2016), and that the exact same setup can be used to generate bulk RNA-seq libraries. While SCRIB-seq is slightly more cost-effective than MARS-seq and has the advantage that one does not need to produce the transposase in-house, Smart-seq2 is preferable when transcriptome annotation, identification of sequence variants, or the quantification of different splice forms is of interest. Furthermore, the presence of batch effects shows that experiments need to be designed in a way that does not confound batches with biological factors (Hicks et al., 2015). Practically, plate-based methods might currently accommodate complex experimental designs with various biological factors more easily than microfluidic chips.

We find that Drop-seq, MARS-seq, SCRIB-seq, and Smart-seq2 (using in-house transposase) are 2- to 13-fold more cost efficient than CEL-seq2/C1, Smart-seq/C1, and Smart-seq2 (using commercial transposase). Hence, the latter methods

would need to increase in their power and/or decrease in their costs to be competitive. The efficiency of the Fluidigm C1 platform can be further increased by microfluidic chips with a higher throughput, as available in the high-throughput (HT) mRNA-seq integrated fluidic circuit (IFC) chip. While CEL-seq2/C1 has been found to be more sensitive than the plate-based version of CEL-seq2 (Hashimshony et al., 2016), the latter might be more efficient when considering its lower costs. Our finding that Smart-seq2 is the most sensitive protocol also hints toward further possible improvements of SCRIB-seq and Drop-seq. As these methods also rely on template switching and PCR amplification, the improvements found in the systematic optimization of Smart-seq2 (Picelli et al., 2013) also could improve the sensitivity of SCRIB-seq and Drop-seq. Furthermore, the costs of SCRIB-seq libraries per cell can be halved when switching to a 384-well format (Soumillon et al., 2014). Similarly, improvements made for CEL-seq2 (Hashimshony et al., 2016) could be incorporated into the MARS-seq protocol. Hence, it is clear that scRNA-seq protocols will become even more efficient in the future. The results of our comparative analyses of six currently prominent scRNA-seq methods may facilitate such developments, and they provide a framework for method evaluation in the future.

In summary, we systematically compared six prominent scRNA-seq methods and found that Drop-seq is preferable when quantifying transcriptomes of large numbers of cells with low sequencing depth, SCRIB-seq and MARS-seq is preferable when quantifying transcriptomes of fewer cells, and Smart-seq2 is preferable when annotating and/or quantifying transcriptomes of fewer cells as long one can use in-house-produced transposase. Our analysis allows an informed choice among the tested methods, and it provides a framework for benchmarking future improvements in scRNA-seq methodologies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
 - Published data
 - Single cell RNA-seq library preparations
 - DNA sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
 - Basic data processing and sequence alignment
 - Power Simulations
 - ERCC capture efficiency
 - Cost efficiency calculation
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes ten figures and one table and can be found with this article online at <http://dx.doi.org/10.1016/j.molcel.2017.01.023>.

AUTHOR CONTRIBUTIONS

C.Z. and W.E. conceived the experiments. C.Z. prepared scRNA-seq libraries and analyzed the data. B.V. implemented the power simulation framework and estimated the ERCC capture efficiencies. S.P. helped in data processing and power simulations. B.R. prepared the Smart-seq2 scRNA-seq libraries. A.G.-A. and H.H. established and performed the MARS-seq library preps. M.S. performed the cell culture of mESCs. W.E. and H.L. supervised the experimental work and I.H. provided guidance in data analysis. C.Z., I.H., B.R., and W.E. wrote the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank Rickard Sandberg for facilitating the Smart-seq2 sequencing. We thank Christopher Mulholland for assistance with FACS, Dominik Alterauge for help establishing the Drop-seq method, and Stefan Krebs and Helmut Blum from the LAFUGA platform for sequencing. We are grateful to Magali Soumillon and Tarjei Mikkelsen for providing the SCRB-seq protocol. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through LMUexcellent and the SFB1243 (Subproject A01/A14/A15) as well as a travel grant to C.Z. by the Boehringer Ingelheim Fonds.

Received: August 8, 2016
 Revised: December 1, 2016
 Accepted: January 17, 2017
 Published: February 9, 2017

REFERENCES

- Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* *10*, 1093–1095.
- Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* *343*, 193–196.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* *29*, 15–21.
- ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* *489*, 57–74.
- Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Pric, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* *16*, 278.
- Frazee, A.C., Jaffe, A.E., Langmead, B., and Leek, J.T. (2015). Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* *31*, 2778–2784.
- Gokce, O., Stanley, G.M., Treutlein, B., Neff, N.F., Camp, J.G., Malenka, R.C., Rothwell, P.E., Fuccillo, M.V., Südhof, T.C., and Quake, S.R. (2016). Cellular taxonomy of the mouse striatum as revealed by single-cell RNA-seq. *Cell Rep.* *16*, 1126–1137.
- Grün, D., and van Oudenaarden, A. (2015). Design and analysis of single-cell sequencing experiments. *Cell* *163*, 799–810.
- Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat. Methods* *11*, 637–640.
- Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* *2*, 666–673.
- Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., et al. (2016). CEL-seq2: sensitive highly-multiplexed single-cell RNA-seq. *Genome Biol.* *17*, 77.
- Hicks, S.C., Teng, M., and Irizarry, R.A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv*. <http://dx.doi.org/10.1101/025528>.
- Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* *11*, 163–166.
- Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* *343*, 776–779.
- Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* *21*, 1543–1551.
- Kalinka, A.T. (2013). The probability of drawing intersections: extending the hypergeometric distribution. *arXiv*, arXiv:1305.0717. <https://arxiv.org/abs/1305.0717>.
- Kharchenko, P.V., Silberstein, L., and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* *11*, 740–742.
- Kim, J.K., Kolodziejczyk, A.A., Ilicic, T., Teichmann, S.A., and Marioni, J.C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* *6*, 8687.
- Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* *9*, 72–74.
- Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* *161*, 1187–1201.
- Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015a). The technology and biology of single-cell RNA sequencing. *Mol. Cell* *58*, 610–620.
- Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C.H., Ilicic, T., Henriksson, J., Natarajan, K.N., Tuck, A.C., Gao, X., Bühler, M., Liu, P., et al. (2015b). Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* *17*, 471–485.
- Lamble, S., Batty, E., Attar, M., Buck, D., Bowden, R., Lunter, G., Crook, D., El-Fahmawi, B., and Piazza, P. (2013). Improved workflows for high throughput library preparation using the transposome-based Nextera system. *BMC Biotechnol.* *13*, 104.
- Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* *15*, R29.
- Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* *69*, 915–926.
- Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* *41*, e108.
- Lun, A.T.L., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* *17*, 75.
- Macosko, E.Z., Basu, A., Satija, R., Nemeshe, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* *161*, 1202–1214.
- Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* *24*, 496–510.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* *17*, 10–12.
- Mora-Castilla, S., To, C., Vaezeslami, S., Morey, R., Srinivasan, S., Dumdie, J.N., Cook-Andersen, H., Jenkins, J., and Laurent, L.C. (2016). Miniaturization technologies for efficient single-cell library preparation for next-generation sequencing. *J. Lab. Autom.* *21*, 557–567.

- Nam, D.K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., Chen, J., Rowley, J.D., and Wang, S.M. (2002). Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc. Natl. Acad. Sci. USA* *99*, 6152–6156.
- Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., and Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* *6*, 25533.
- Petropoulos, S., Edsgård, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* *165*, 1012–1026.
- Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* *10*, 1096–1098.
- Picelli, S., Björklund, A.K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014a). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* *24*, 2033–2040.
- Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014b). Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* *9*, 171–181.
- Reinius, B., Mold, J.E., Ramsköld, D., Deng, Q., Johnsson, P., Michaëlsson, J., Frisén, J., and Sandberg, R. (2016). Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.* *48*, 1430–1435.
- Renaud, G., Stenzel, U., Maricic, T., Wiebe, V., and Kelso, J. (2015). deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics* *31*, 770–772.
- Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* *32*, 896–902.
- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* *43*, e47.
- Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A., and Mikkelsen, T.S. (2014). Characterization of directed differentiation by high-throughput single-cell RNA-seq. *bioRxiv*. <http://dx.doi.org/10.1101/003236>.
- Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* *16*, 133–145.
- Svensson, V., Natarajan, K.N., Ly, L.-H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., and Teichmann, S.A. (2016). Power analysis of single cell RNA-sequencing experiments. *bioRxiv*. <http://dx.doi.org/10.1101/073692>.
- Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* *352*, 189–196.
- Vallejos, C.A., Richardson, S., and Marioni, J.C. (2016). Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol.* *17*, 70.
- Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* *34*, 1145–1160.
- Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., and Quake, S.R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* *11*, 41–46.
- Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betscholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* *347*, 1138–1142.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Esgro recombinant mouse LIF	Millipore	ESG1107
CHIR99021	Axon Med Chem	1386
PD0325901	Axon Med Chem	1408
2-Mercaptoethanol	Sigma-Aldrich	M3148
FBS	Sigma-Aldrich	F7524
Penicillin/Streptomycin	Sigma-Aldrich	P4333
MEM non-essential amino acids	Sigma-Aldrich	M7145
L-glutamine	Sigma-Aldrich	G7513
Dulbecco's modified Eagle's medium	Sigma-Aldrich	D6429
Perfluorooctanol	Sigma-Aldrich	370533
Maxima H- Reverse Transcriptase	Thermo Fisher Scientific	EP0753
SuperScript II	Life Technologies	18064071
Exonuclease I	New England Biolabs	M0293L
RNAProtect Cell Reagent	QIAGEN	76526
RNase inhibitor	Promega	N2515
RNase inhibitor	Lucigen	30281-2-LU
Phusion HF buffer	New England Biolabs	B0518S
Proteinase K	Ambion	AM2546
KAPA HiFi HotStart polymerase	KAPA Biosystems	KAPBKK2602
Phusion HF PCR Master Mix	Thermo Fisher Scientific	F531L
dNTPs	New England Biolabs	N0447L
Triton X-100	Sigma-Aldrich	T8787
SDS	Sigma-Aldrich	L3771
Tn5 transposase	Picelli et al., 2014a	N/A
Critical Commercial Assays		
C1 Single-Cell System	Fluidigm	N/A
C1 IFC for Open App (10-17 μ m)	Fluidigm	100-8134
C1 IFC for mRNA-seq (10-17 μ m)	Fluidigm	100-6041
Nextera XT DNA Sample Preparation Kit	Illumina	FC-131-1096
SMARTer Ultra Low RNA Kit for Fluidigm C1	Clontech	634833
MinElute Gel Extraction Kit	QIAGEN	28606
Deposited Data		
single-cell RNA-seq data	This paper	GEO: GSE75790
Drop-seq ERCC data	Macosko et al., 2015	GEO: GSE66694
Experimental Models: Cell Lines		
J1 mouse embryonic stem cells	Li et al., 1992	N/A
Sequence-Based Reagents		
Nextera XT Index Kit	Illumina	FC-121-1012
SCRB-seq P5 primer, AATGATACGGCGACCACCG AGATCTACACTCTTTCCCTACACGACGCTCTTC CG*A*T*C*T, * PTO bond	IDT	N/A
SCRB-seq oligo-dT primer, Biotin-ACACTCTTTCCCT ACACGACGCTCTTCGATCT[BC6][N10][T30]VN	IDT	"TruGrade Ultramer"

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SCRB-seq template-switch oligo, iCiGiCAGACTCTTTCC CTACACGACGCrGrGrG	Eurogentech	N/A
Drop-seq P5 primer, AATGATACGGCGACCACCGAGA TCTACACGCCT GTCCGCGAAGCAGTGGTATCAACG CAGAGT*A*C, * PTO bond	IDT	N/A
Drop-seq oligo-dT primer beads, Bead-Linker- TTTTTTTAAGCAGTGGTATCAAC GCAGAGTAC[BC12][N8][T30]	Chemgenes	MACOSKO-2011-10
Drop-seq template-switch oligo, AAGCAGTGGTATCA ACGCAGAGTGAATrGrGrG	IDT	N/A
CEL-seq2 oligo-dT primer, GCCGGTAATACGACTCACTATA GGGAGTTCTACAGTCCGACGATC[N6][BC6][T25]	Sigma-Aldrich	N/A
ERCC RNA Spike-In Mix	Ambion	4456740
Software and Algorithms		
STAR	Dobin et al., 2013	https://github.com/alexdobin/STAR
Drop-seq tools	Macosko et al., 2015	http://mccarrolllab.com/dropseq/
featureCounts	Liao et al., 2013	https://bioconductor.org/packages/release/bioc/html/Rsubread.html
R	N/A	www.r-project.org
Other		
Drop-seq PDMS device	Nanoshift	Drop-seq
2% E-Gel Agarose EX Gels	Life Technologies	G402002

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the corresponding author Wolfgang Enard (enard@biologie.uni-muenchen.de).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

J1 mouse embryonic stem cells ([Li et al., 1992](#)) were maintained on gelatin-coated dishes in Dulbecco's modified Eagle's medium supplemented with 16% fetal bovine serum (FBS, Sigma-Aldrich), 0.1 mM β -mercaptoethanol (Sigma-Aldrich), 2 mM L-glutamine, 1x MEM non-essential amino acids, 100 U/ml penicillin, 100 μ g/ml streptomycin (Sigma-Aldrich), 1000 U/ml recombinant mouse LIF (Millipore) and 2i (1 μ M PD032591 and 3 μ M CHIR99021 (Axon Medchem, Netherlands). J1 embryonic stem cells were obtained from E. Li and T. Chen and mycoplasma free determined by a PCR-based test. Cell line authentication was not recently performed.

METHOD DETAILS**Published data**

Drop-seq ERCC ([Macosko et al., 2015](#)) data were obtained under accession GEO: GSE66694. Raw fastq files were extracted using the SRA toolkit (2.3.5). We trimmed cDNA reads to the same length and processed raw reads in the same way as data sequenced for this study.

Single cell RNA-seq library preparations**CEL-seq2/C1**

CEL-seq2/C1 libraries were generated as previously described ([Hashimshony et al., 2016](#)). Briefly, cells (200,000/ml), ERCC spike-ins, reagents and barcoded oligo-dT primers (Sigma-Aldrich) were loaded on a 10-17 μ m C1 Open-App microfluidic IFC (Fluidigm). Cell lysis, reverse transcription, second strand synthesis and in-vitro transcription were performed on-chip. Subsequently, harvested aRNA was pooled from 48 capture sites. After fragmentation and clean-up, 5 μ l of aRNA was used to construct final libraries by reverse transcription (SuperScript II, Thermo Fisher) and library PCR (Phusion HF, Thermo Fisher).

Drop-seq

Drop-seq experiments were performed as published (Macosko et al., 2015) and successful establishment of the method in our lab was confirmed by a species-mixing experiment (Figure S1A). For this work, J1 mES cells (100/μl) and barcode-beads (120/μl, Chem-genes) were co-flown in Drop-seq PDMS devices (Nanoshift) at rates of 4000 μl/hr. Collected emulsions were broken by addition of perfluorooctanol (Sigma-Aldrich) and mRNA on beads was reverse transcribed (Maxima RT, Thermo Fisher). Unused primers were degraded by addition of Exonuclease I (New England Biolabs). Washed beads were counted and aliquoted for pre-amplification (2000 beads / reaction). Nextera XT libraries were constructed from 1 ng of pre-amplified cDNA with a custom P5 primer (IDT).

MARS-seq

To construct single cell libraries from polyA-tailed RNA, we applied massively parallel single-cell RNA sequencing (MARS-Seq) (Jaitin et al., 2014). Briefly, single cells were FACS-sorted into 384-well plates, containing lysis buffer and reverse-transcription (RT) primers. The RT primers contained the single cell barcodes and unique molecular identifiers (UMIs) for subsequent de-multiplexing and correction for amplification biases, respectively. Spike-in transcripts (ERCC, Ambion) were added, polyA-containing RNA was converted into cDNA as previously described and then pooled using an automated pipeline (liquid handling robotics). Subsequently, samples were linearly amplified by in vitro transcription, fragmented, and 3' ends were converted into sequencing libraries. The libraries consisted of 48 single cell pools.

SCRB-seq

RNA was stabilized by resuspending cells in RNeasy Protect Cell Reagent (QIAGEN) and RNase inhibitors (Promega). Prior to FACS sorting, cells were diluted in PBS (Invitrogen). Single cells were sorted into 5 μl lysis buffer consisting of a 1/500 dilution of Phusion HF buffer (New England Biolabs) and ERCC spike-ins (Ambion), spun down and frozen at -80°C. Plates were thawed and libraries prepared as described previously (Soumillon et al., 2014). Briefly, RNA was desiccated after protein digestion by Proteinase K (Ambion). RNA was reverse transcribed using barcoded oligo-dT primers (IDT) and products pooled and concentrated. Unincorporated barcode primers were digested using Exonuclease I (New England Biolabs). Pre-amplification of cDNA pools were done with the KAPA HiFi HotStart polymerase (KAPA Biosystems). Nextera XT libraries were constructed from 1 ng of pre-amplified cDNA with a custom P5 primer (IDT).

Smart-seq/C1

Smart-seq/C1 libraries were prepared on the Fluidigm C1 system using the SMARTer Ultra Low RNA Kit (Clontech) according to the manufacturer's protocol. Cells were loaded on a 10-17 μm RNA-seq microfluidic IFC at a concentration of 200,000/ml. Capture site occupancy was surveyed using the Operetta (Perkin Elmer) automated imaging platform.

Smart-seq2

mESCs were sorted into 96-well PCR plates containing 2 μl lysis buffer (1.9 μl 0.2% Triton X-100; 0.1 μl RNase inhibitor (Lucigen)) and spike-in RNAs (Ambion), spun down and frozen at -80°C. To generate Smart-seq2 libraries, priming buffer mix containing dNTPs and oligo-dT primers was added to the cell lysate and denatured at 72°C. cDNA synthesis and pre-amplification of cDNA was performed as described previously (Picelli et al., 2014b, 2013). Sequencing libraries were constructed from 2.5 ng of pre-amplified cDNA using an in-house generated Tn5 transposase (Picelli et al., 2014a). Briefly, 5 μl cDNA was incubated with 15 μl tagmentation mix (1 μl of Tn5; 2 μl 10x TAPS MgCl₂ Tagmentation buffer; 5 μl 40% PEG8000; 7 μl water) for 8 min at 55°C. Tn5 was inactivated and released from the DNA by the addition of 5 μl 0.2% SDS and 5 min incubation at room temperature. Sequencing library amplification was performed using 5 μl Nextera XT Index primers (Illumina) that had been first diluted 1:5 in water and 15 μl PCR mix (1 μl KAPA HiFi DNA polymerase (KAPA Biosystems); 10 μl 5x KAPA HiFi buffer; 1.5 μl 10mM dNTPs; 2.5 μl water) in 10 PCR cycles. Barcoded libraries were purified and pooled at equimolar ratios.

DNA sequencing

For SCR-seq and Drop-seq, final library pools were size-selected on 2% E-Gel Agarose EX Gels (Invitrogen) by excising a range of 300-800 bp and extracting DNA using the MinElute Kit (QIAGEN) according to the manufacturer's protocol.

Smart-seq/C1, CEL-seq2/C1, Drop-seq and SCR-seq library pools were sequenced on an Illumina HiSeq1500. Smart-seq2 pools were sequenced on Illumina HiSeq2500 (Replicate A) and HiSeq2000 (Replicate B) platforms. MARS-seq library pools were sequenced on an Illumina HiSeq2500 using the Rapid mode. Smart-seq/C1 and Smart-seq2 libraries were sequenced 45 cycles single-end, whereas CEL-seq2/C1, Drop-seq and SCR-seq libraries were sequenced paired-end with 15-20 cycles to decode cell barcodes and UMI from read 1 and 45 cycles into the cDNA fragment. MARS-seq libraries were paired-end sequenced with 52 cycles on read 1 into the cDNA and 15 bases for read 2 to obtain cell barcodes and UMIs. Similar sequencing qualities were confirmed by FastQC v0.10.1 (Figure S1B).

QUANTIFICATION AND STATISTICAL ANALYSIS**Basic data processing and sequence alignment**

Smart-seq/C1/Smart-seq2 libraries (i5 and i7) and CELseq2/C1/Drop-seq/SCR-seq pools (i7) were demultiplexed from the Illumina barcode reads using deML (Renaud et al., 2015). MARS-seq library pools were demultiplexed with the standard Illumina pipeline. All reads were trimmed to the same length of 45 bp by cutadapt (Martin, 2011) (v1.8.3) and mapped to the mouse genome (mm10)

including mitochondrial genome sequences and unassigned scaffolds concatenated with the ERCC spike-in reference. Alignments were calculated using STAR 2.4.0 (Dobin et al., 2013) using all default parameters.

For libraries containing UMIs, cell- and gene-wise count/UMI tables were generated using the published Drop-seq pipeline (v1.0) (Macosko et al., 2015). We discarded the last 2 bases of the Drop-seq cell and molecular barcodes to account for bead synthesis errors. For Smart-seq/C1 and Smart-seq2, features were assigned and counted using the Rsubread package (v1.20.2) (Liao et al., 2013).

Power Simulations

We developed a framework in R for statistical power evaluation of differential gene expression in single cells. For each method, we estimated the mean expression, dispersion and dropout probability per gene from the same number of cells per method. In the read count simulations, we followed the framework proposed in Polyester (Frazee et al., 2015), i.e., we retained the observed mean-variance dependency by applying a cubic smoothing spline fit to capture the heteroscedasticity observed. Furthermore, we included a local polynomial regression fit for the mean-dropout relationship. In each iteration, we simulated count measurements for the 13,361 genes for sample sizes of 2^4 , 2^5 , 2^6 , 2^7 , 2^8 and 2^9 cells per group. The read count for a gene i in a cell j is modeled as a product of a binomial and negative binomial distribution:

$$X_{ij} \sim B(p = 1 - p_0) * NB(\mu, \theta).$$

The mean expression magnitude μ was randomly drawn from the empirical distribution. 5 percent of the genes were defined as differentially expressed with an effect size drawn from the observed fold changes between microglial subpopulations in Zeisel et al. (Zeisel et al., 2015). The dispersion θ and dropout probability p_0 were predicted by above mentioned fits.

For each method and sample size, 100 RNA-seq experiments were simulated and tested for differential expression using limma (Ritchie et al., 2015) in combination with voom (Law et al., 2014) (v3.26.7). The power simulation framework was implemented in R (v3.3.0).

ERCC capture efficiency

To estimate the single molecule capture efficiency, we assume that the success or failure of detecting an ERCC is a binomial process, as described before (Marinov et al., 2014). Detections are independent from each other and are thus regarded as independent Bernoulli trials. We recorded the number of cells with nonzero and zero read or UMI counts for each ERCC per method and applied a maximum likelihood estimation to fit the probability of successful detection. The fit line was shaded with the 95% Wilson score confidence interval.

Cost efficiency calculation

We based our cost efficiency extrapolation on the power simulations starting from empirical data at different sequencing depths (250,000 reads, 500,000 reads, 1,000,000 reads; Figure 6C). We determined the number of cells required per method and depth for adequate power (80%) by an asymptotic fit to the median powers. For the calculation of sequencing cost, we assumed 5€ per million raw reads, independent of method. Although UMI-based methods need paired-end sequencing, we assumed a 50 cycle sequencing kit is sufficient for all methods. We used prices in Euro as a basis and consider an exchange course of 1:1 for the given prices in USD.

DATA AND SOFTWARE AVAILABILITY

The accession number for the raw and analyzed scRNA-seq data reported in this paper is GEO: GSE75790.

Molecular Cell, Volume 65

Supplemental Information

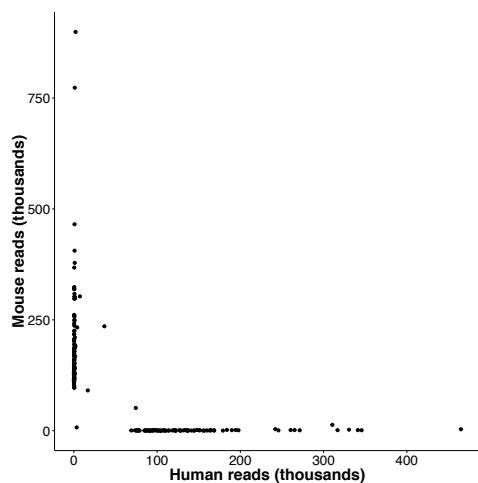
Comparative Analysis of Single-Cell RNA Sequencing Methods

Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard

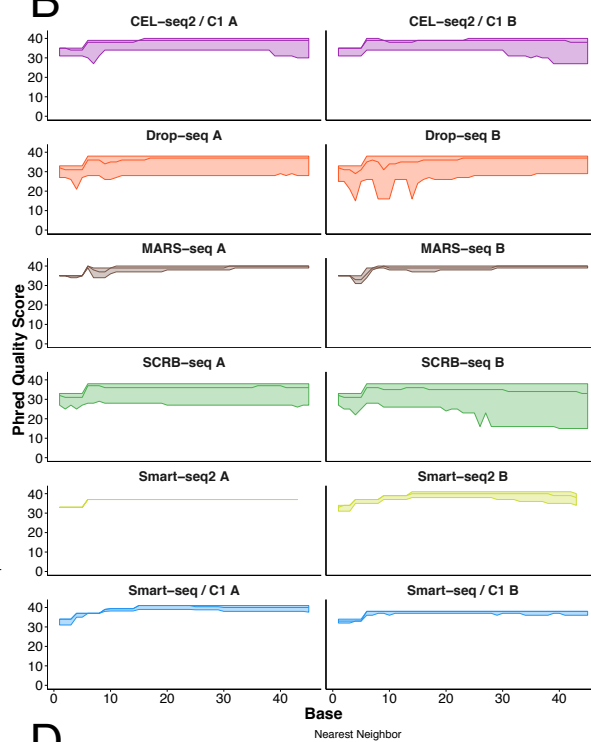
Supplementary Figures

Figure S1

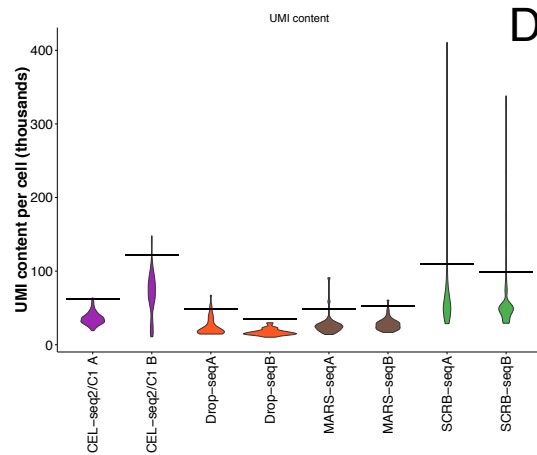
A



B



C



D

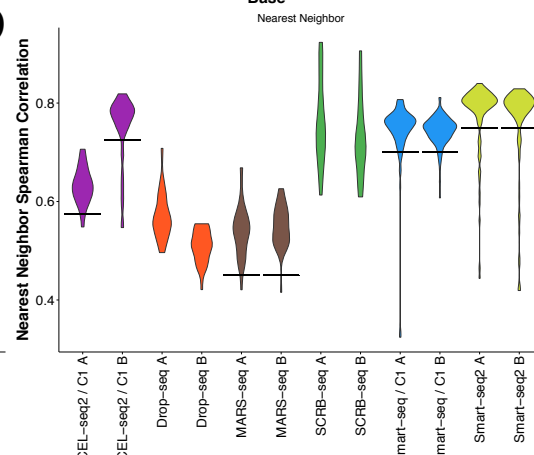


Figure S2

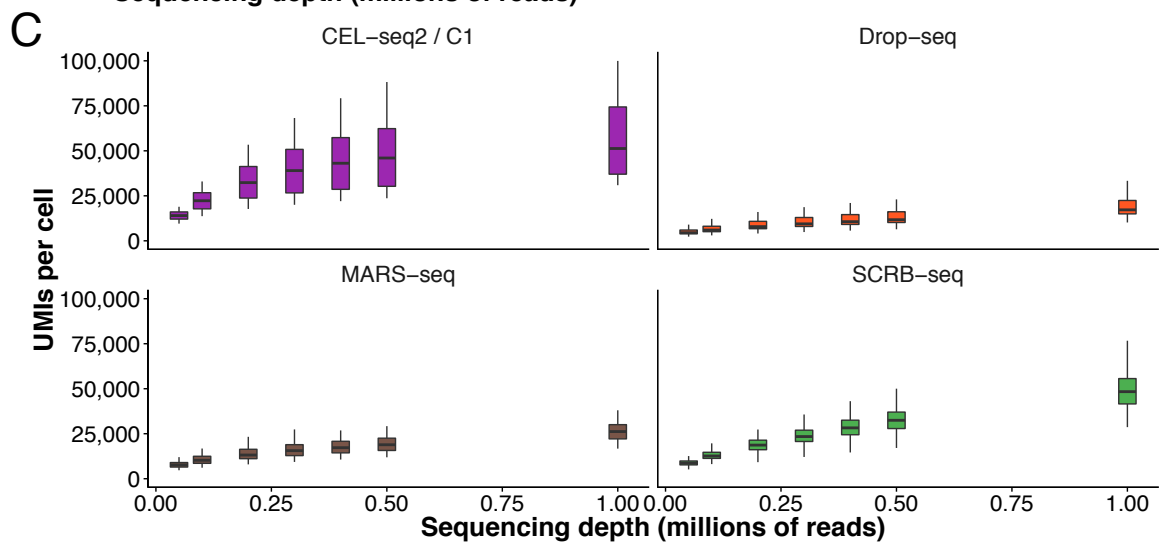
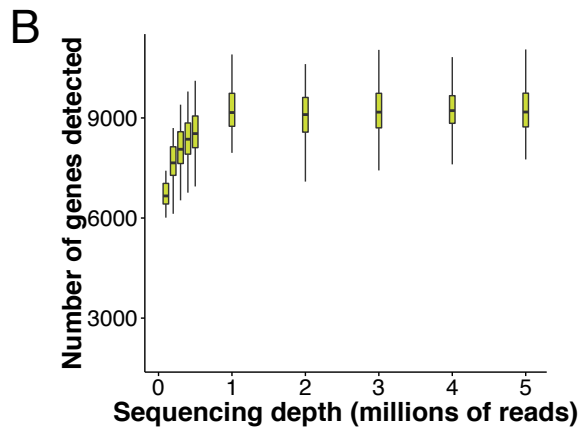
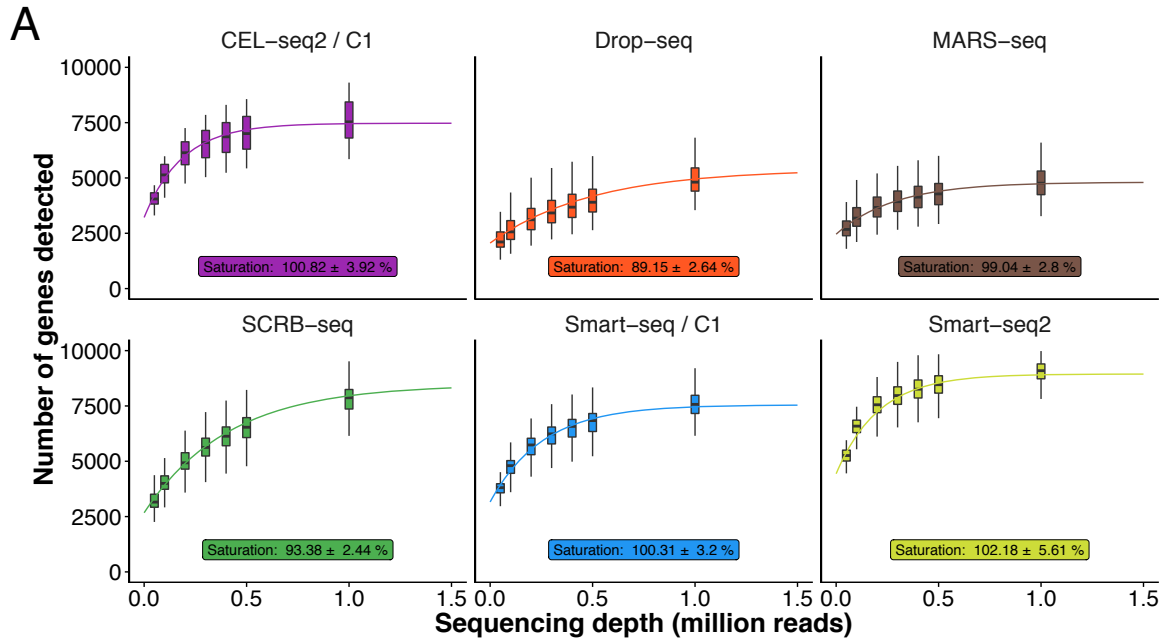


Figure S3

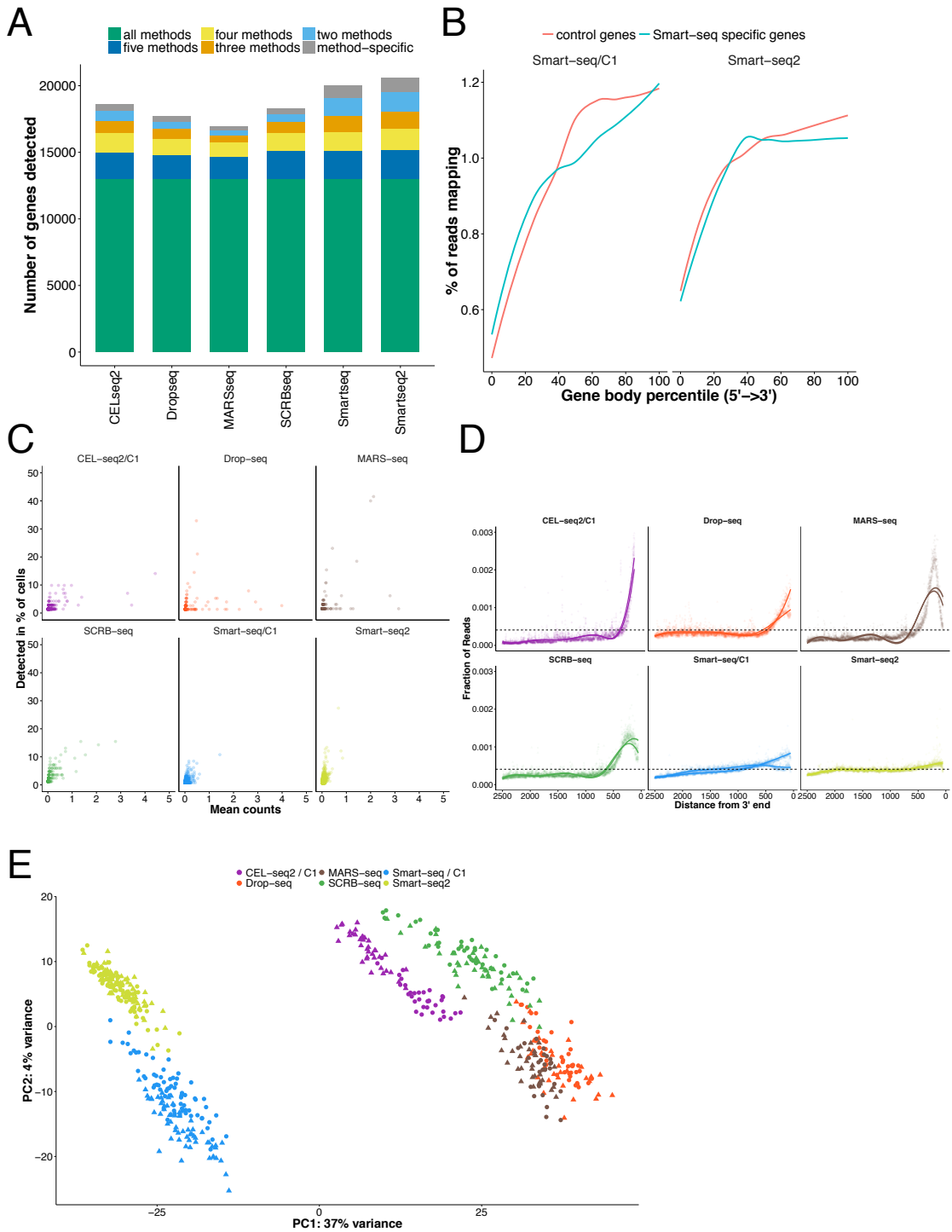
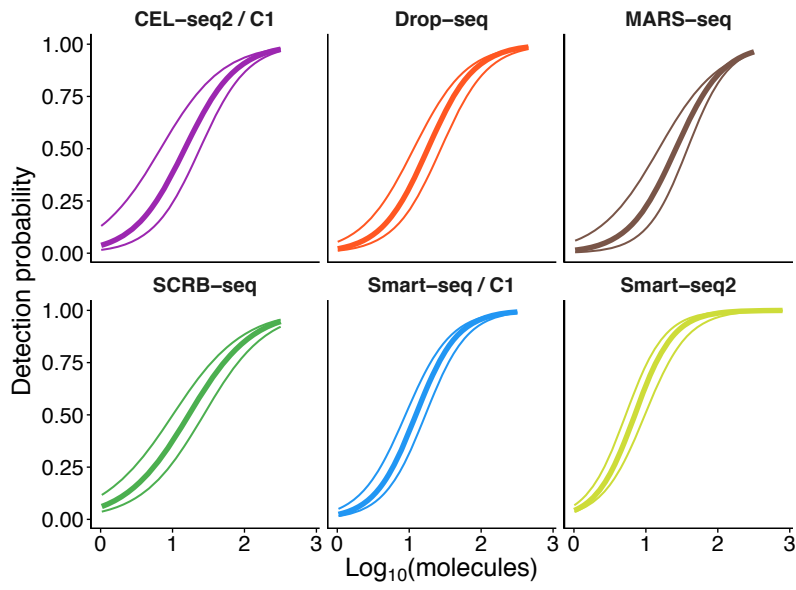
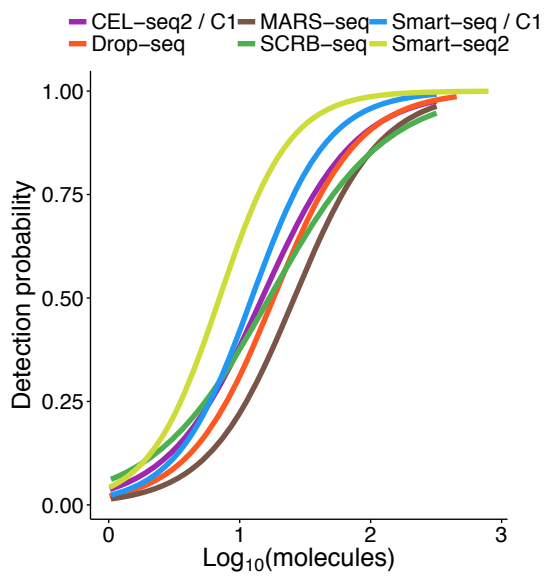


Figure S4

A



B



C

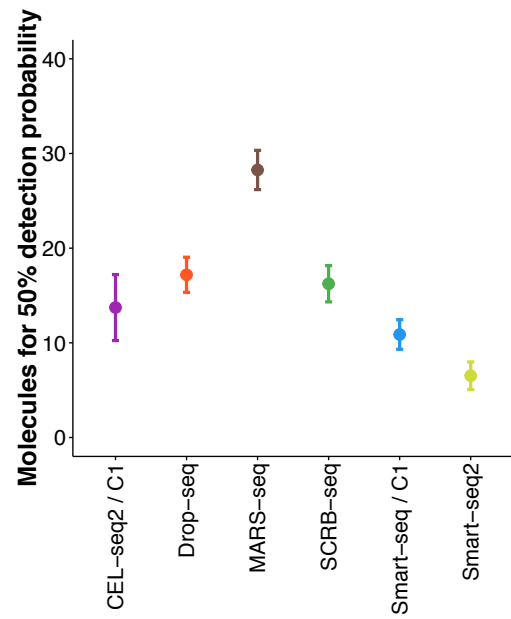


Figure S5

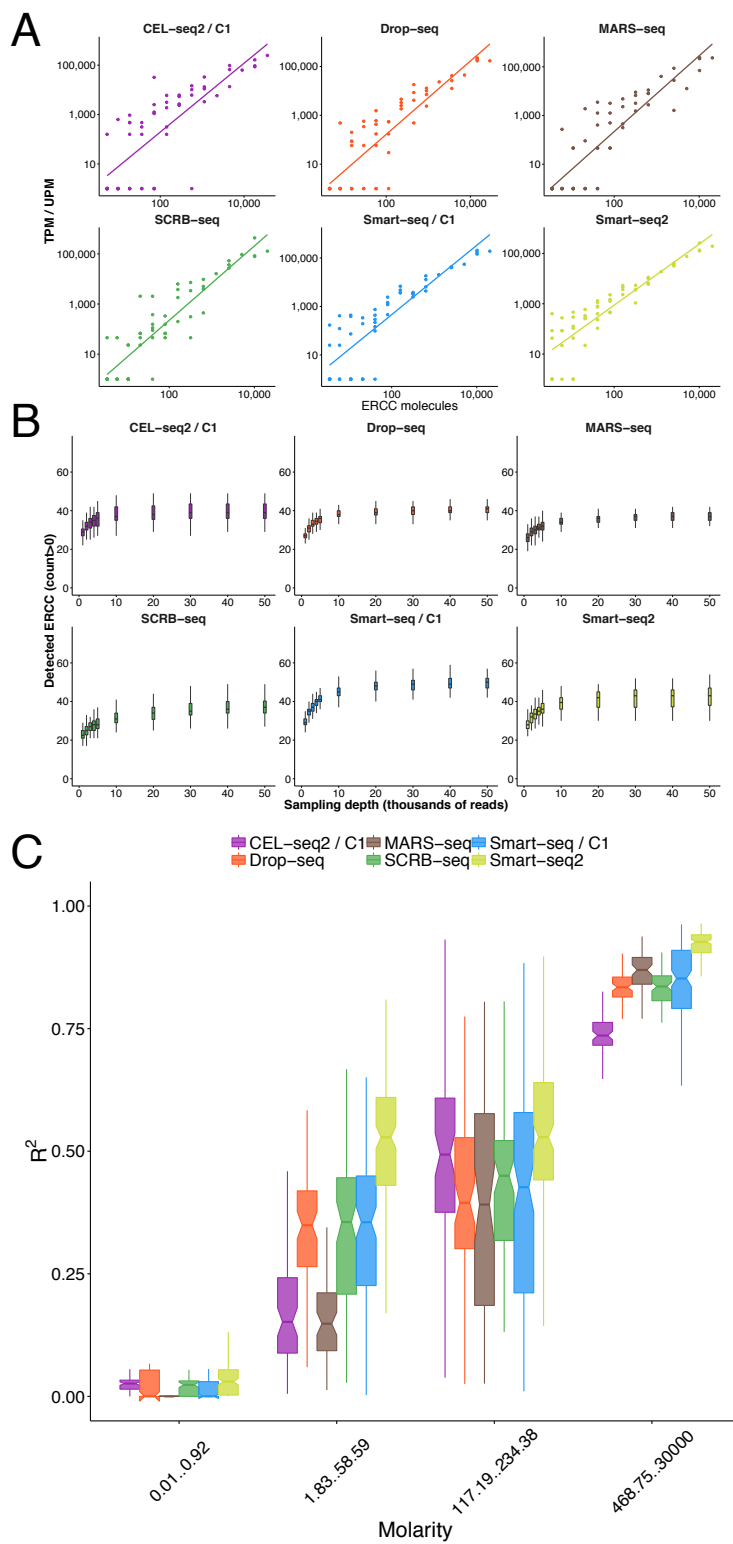


Figure S6

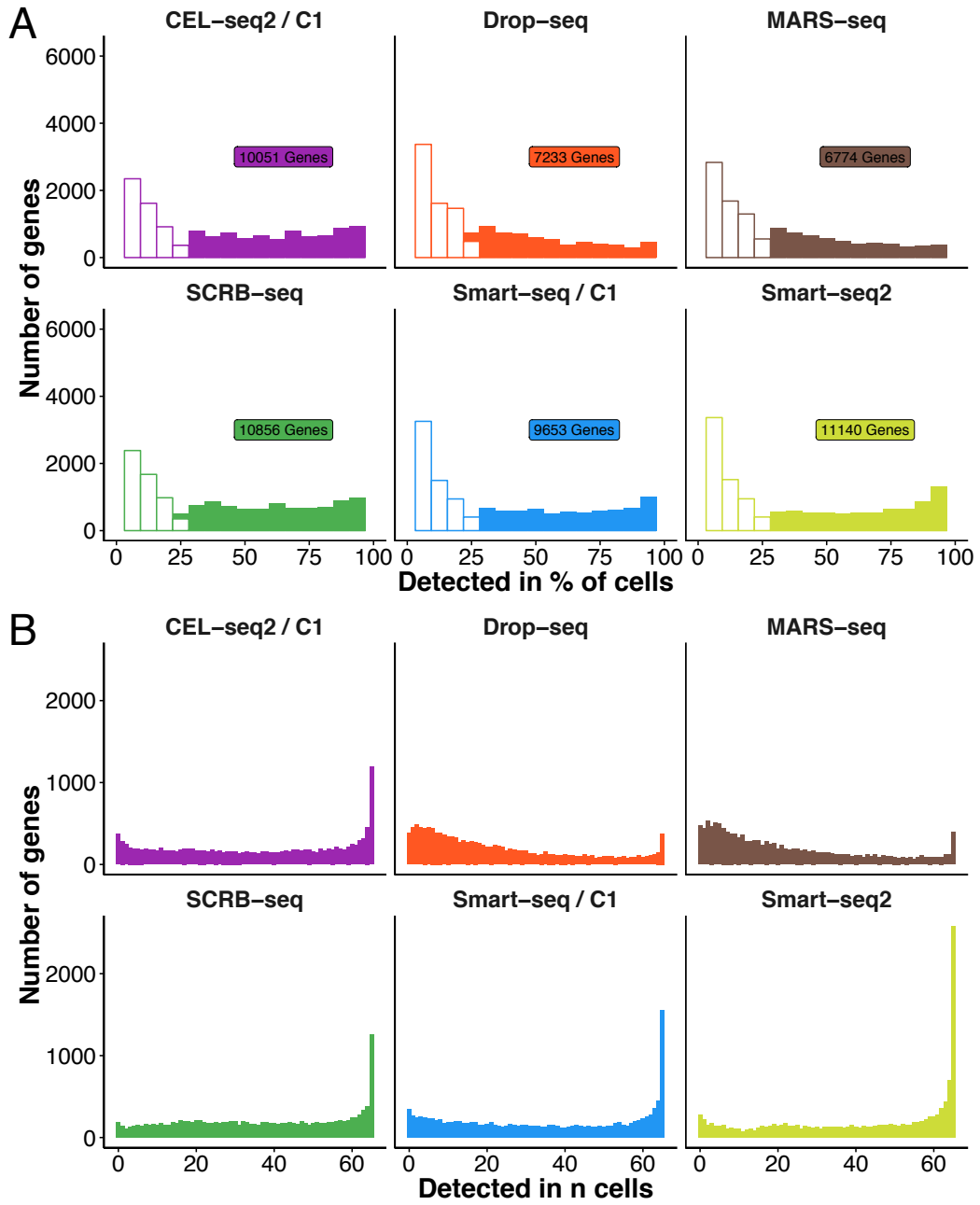


Figure S7

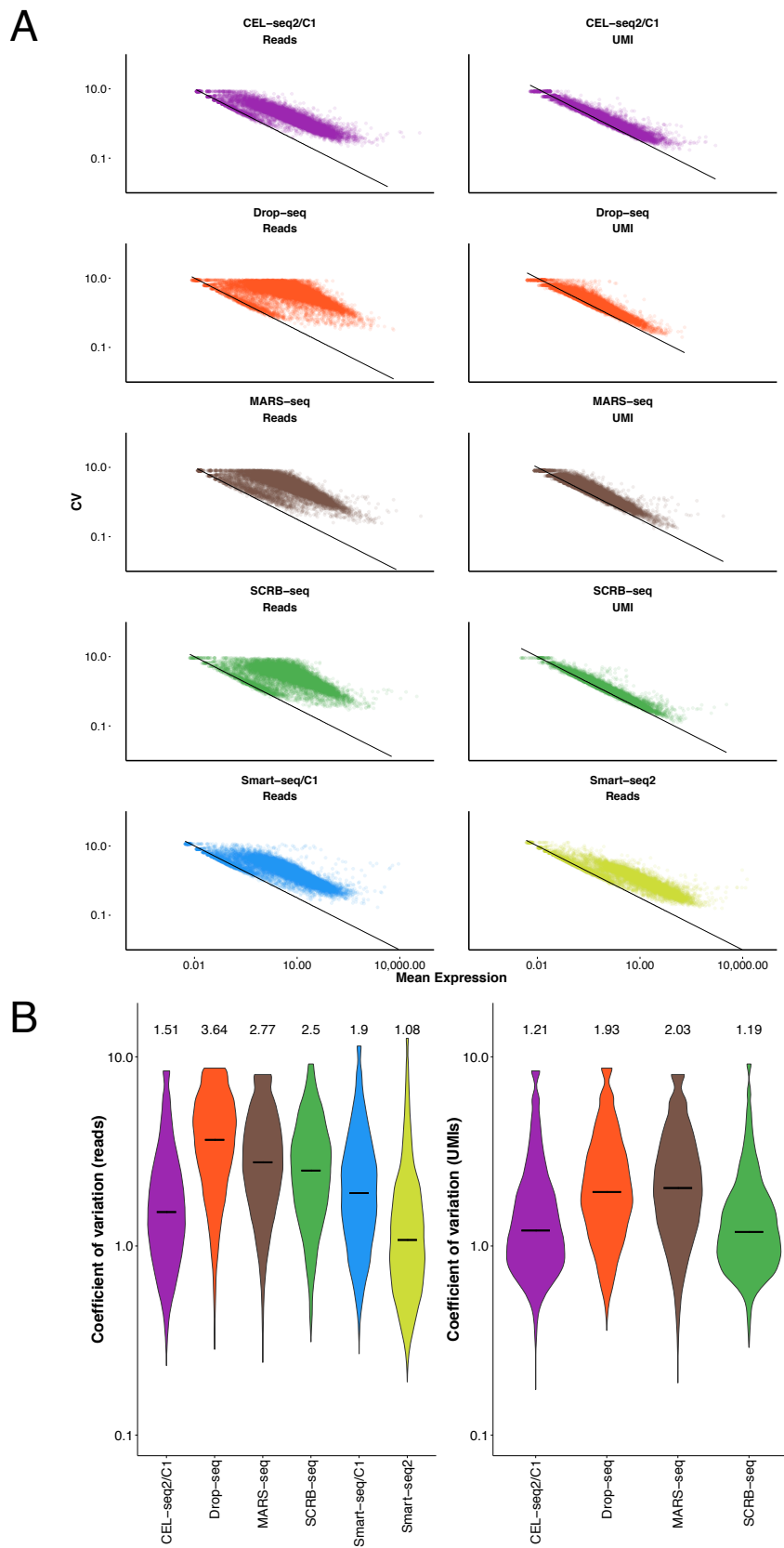


Figure S8

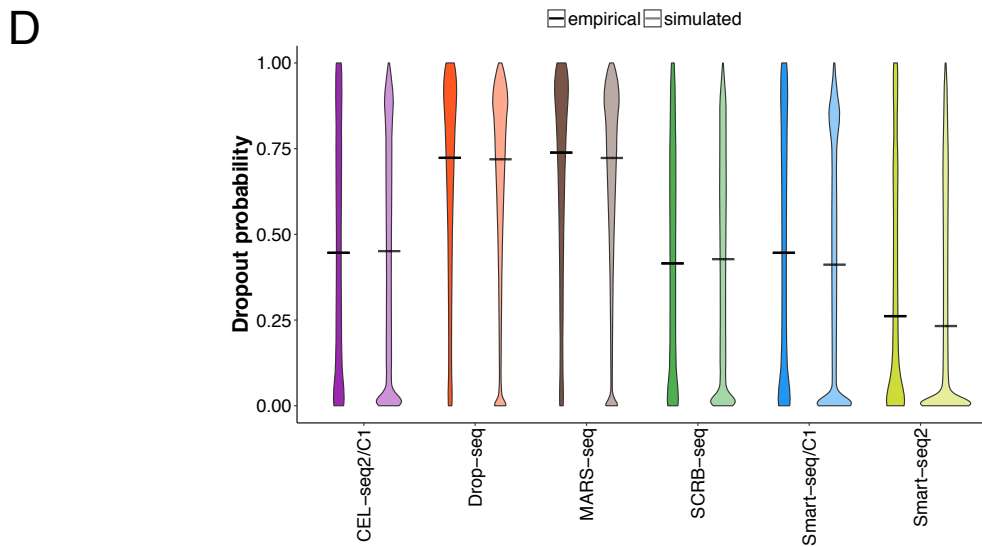
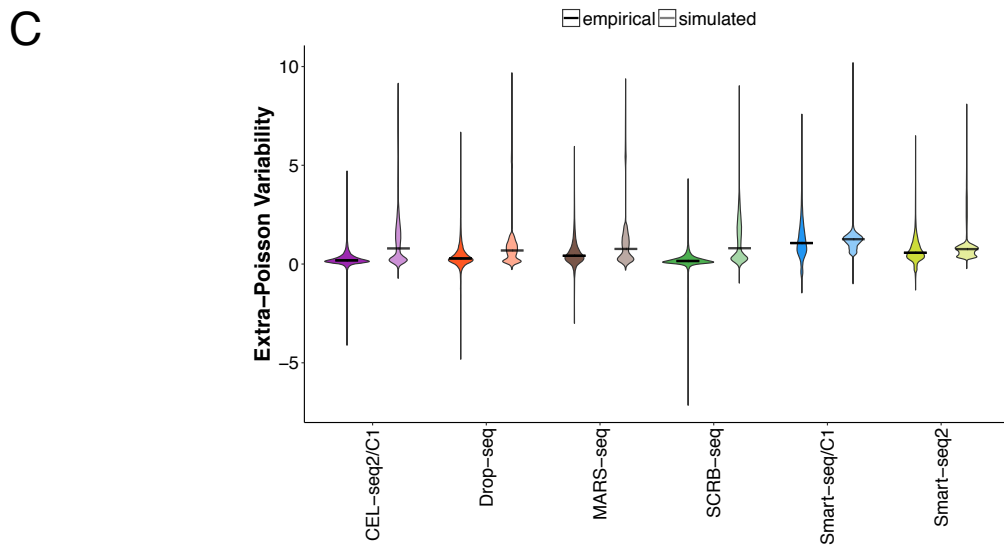
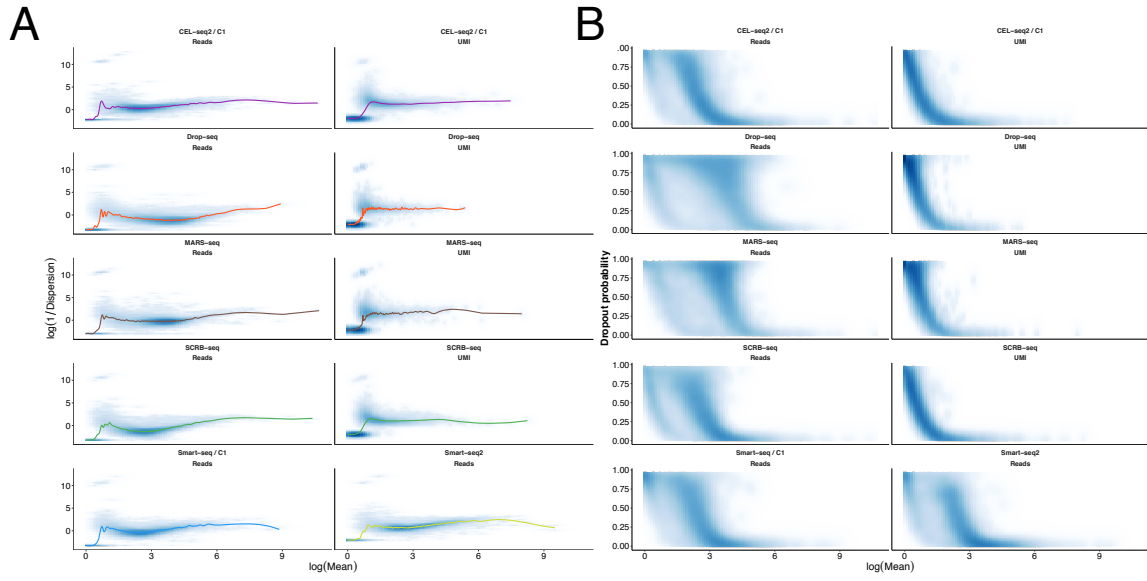


Figure S9

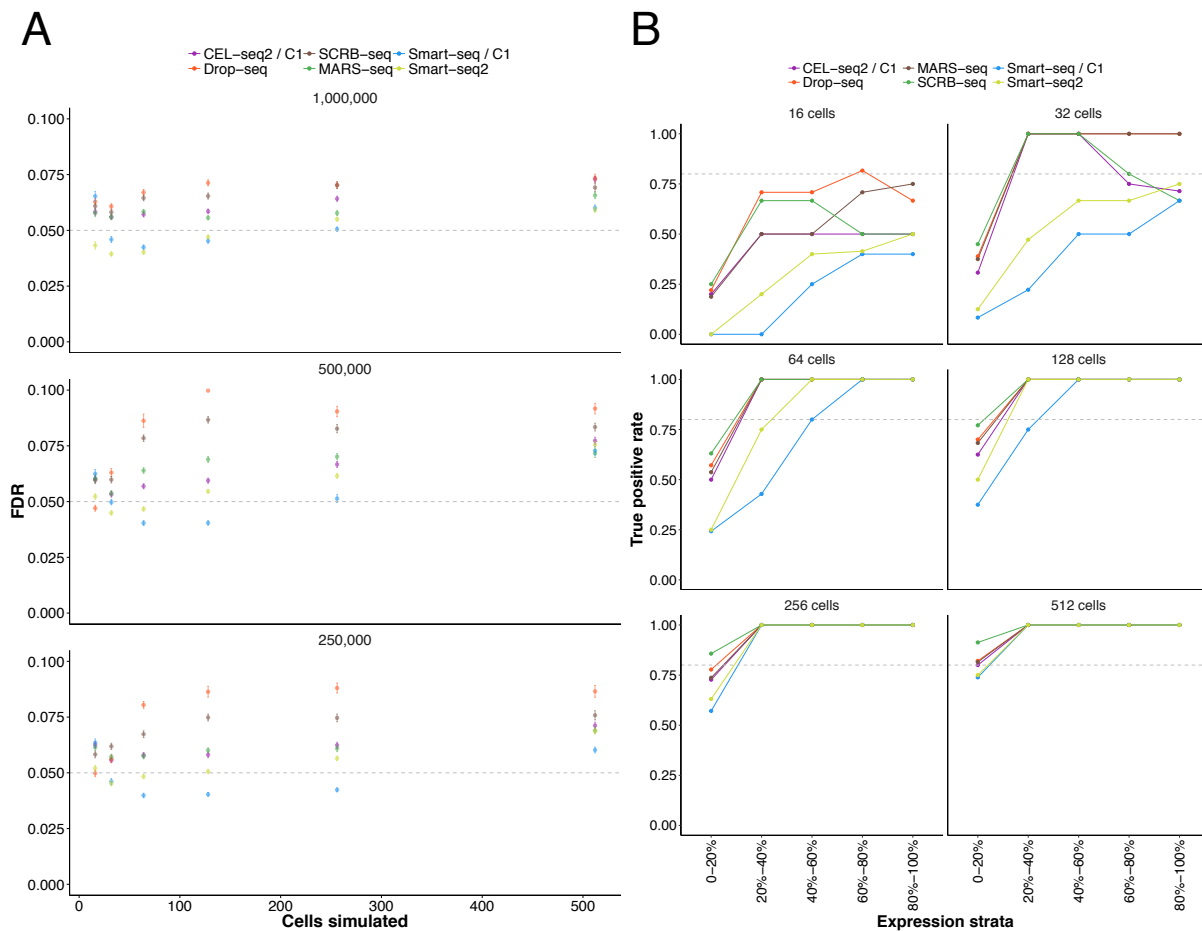
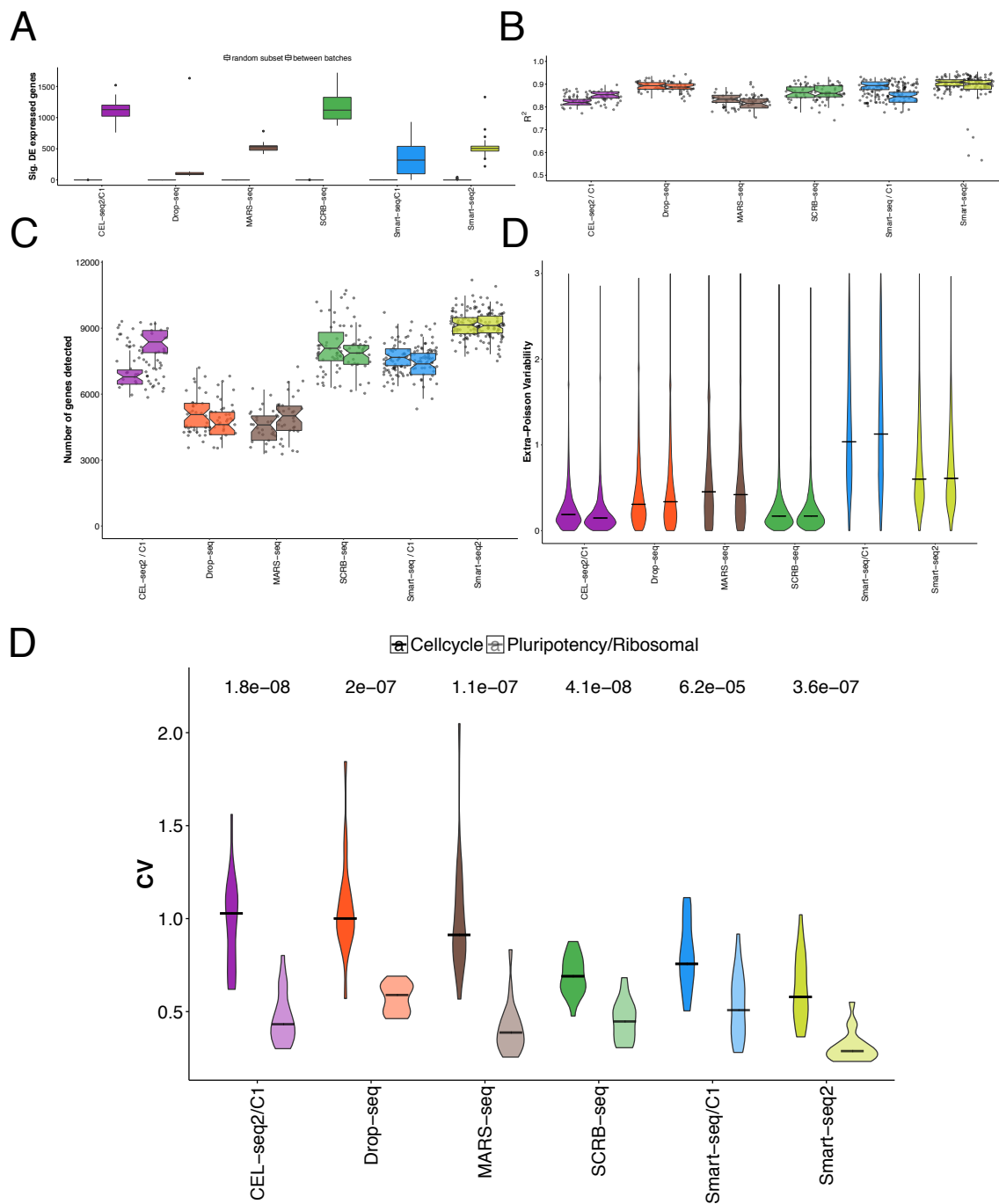


Figure S10



Supplementary Figure Legends

Figure S1 (related to Figure 1) | Quality control and filtering. **A** Drop-seq species mixing experiment using human and murine T-cells. For each cell-barcode human- and mouse read numbers are plotted. **B** Per-base quality scores were summarized using FastQC. Lines indicate median Phred quality score with upper and lower quartile shaded. **C** Total UMI content per cell, with the filter cutoff (two times mean) shown as black lines. Violin plots indicate the density of the UMI content distribution per replicate. **D** Nearest-neighbor filtering based on the maximum pairwise Spearman's rho for each cell. Violin plots indicate the density of rho distribution per replicate. Black lines indicate the employed cutoffs.

Figure S2 (related to Figure 1) | Downsampling of scRNA-seq libraries. **A** Detected genes (≥ 1 count) in relation to indicated sequencing depths. The ranges of the boxes indicate the upper and lower quartiles of cells and horizontal bars indicate the medians. **B** Boxplots of the number of detected genes in high-depth sequencing of Smart-seq2 libraries, showing a plateau above 1 million reads. **C** Boxplots of the number of detected UMIs per cell in relation to indicated sequencing depths.

Figure S3 (related to Figure 3) | Sensitivity **A** The overlap of detected genes (≥ 1 count) between methods for 65 random cells is displayed as a barplot. Colors indicate the level of overlap: Green (detected in all methods), dark blue (detected in five methods), yellow (detected in four methods), orange (detected in three methods), light blue (detected in two methods), grey (method-specific detection). **B** Gene body coverage (left to right equalling 5' to 3') of ~3000 genes detected by Smart-seq/C1 and/or Smart-seq2 (right panel) versus a random control set of 3000 genes detected by all methods. **C** Method-specific detected genes are shown as scatter plots with their rate of detection and mean counts over all cells. **D** For genes and their transcript variants of at least 2 kb length, we calculated the fraction of reads mapping to positions relative to the 3' end. For each method, we show mapping positions and a fit line per replicate. The dashed line indicates theoretical even distribution of reads across the 2.5 kb window. **(E)** Gene expression values were normalized as transcripts per million TPM or UMIs per million UPM. Principal component analysis was performed on the 1000 most variable genes to display the major variance between single cells. The 200 genes with the highest loading for PC1 were analysed and neither showed significant enrichment in GO categories (GORilla) nor in technical properties such as gene length or GC content.

Figure S4 (related to Figure 3) | Detection probabilities were estimated from ERCC dropouts, where the RNA molecule number is known. **A** Thick lines indicate the maximum-likelihood estimate of the detection probability with the thin lines showing the 95% confidence interval of the fit. **B** Shown are per-method maximum-likelihood estimates of mRNA detection probabilities. **C** Sensitivity per method estimated as the 50% probability to detect a transcript. The 95% confidence interval of estimate is displayed as error bars.

Figure S5 (related to Figure 4) | **A** Exemplary correlations of ERCC expression values (transcripts per million TPM or UMIs per million UPM) with annotated concentrations. For each method, we chose a representative cell/bead with a linear model correlation coefficient close to the median of all cells. **B** Detection of ERCC genes (≥ 1 count) in relation to sampling depth. Each boxplot represents the median, upper and lower quartile of all cells within each method. **C** Accuracy of scRNA-seq methods. ERCC expression values were correlated to their annotated molarity. Shown are the distributions of correlation coefficients (adjusted R^2 of linear regression model) across methods for bins of ERCC molarity. Each boxplot represents the median, first and third quartile for the R^2 in the indicated bin.

Figure S6 (related to Figure 5) | Gene detection sparsity. **A** For all detected genes (≥ 1 CPM) per method, we calculated the rate of detection. Histograms show this measure for detection sparsity. Filled bars represent the genes detected in at least 25% of cells of each method along with the number of these reproducibly detected genes. **B** For genes detected in at least 25% of cells of any method, we calculate the rate of detection in 65 random cells.

Figure S7 (related to Figure 5) | Variation in scRNA-seq data. **A** Gene-wise mean and coefficient of variation from all cells are shown as scatterplots for all methods. The black line indicates variance according to the poisson distribution. The two populations of genes seen for read-count data are unamplified genes (close to Poisson, one or very few reads per UMI) and amplified genes (higher CV for a given mean, several reads per UMI). **B** Gene-wise coefficient of variation (CV) of scRNA-seq data were calculated for all cells including detection dropouts. Violin plots are shown for UMI and read-count based quantification indicating the density of the distribution.

Figure S8 (related to Figure 6) | **A-B** Power simulation parameters estimated from 1 million reads per cell. **A** Mean expression and size parameters were estimated for each method and their functional relation was approximated by a smooth spline fit. **B** The dropout probability p_0 was calculated per gene and shown in relation to mean expression levels. We

fitted this relationship using a local polynomial regression. **C-D** Validation of power simulation framework. **C** Gene-wise Extra-Poisson Variability was calculated from empirical data and simulated data without addition of differentially expressed genes. Shown are the distributions with the black line indicating the median. **D** Gene-wise dropout rate distributions are shown from empirical data and simulated data. The black line indicates the median dropout rate.

Figure S9 (related to Figure 6 and Table 1) | **A** FDR. Simulations were performed using empirical mean, dispersion and dropout relationships (see Figure S8). For variable sample sizes of $n=16$, $n=32$, $n=64$, $n=128$, $n=256$ and $n=512$, we show points representing the mean FDR of 100 simulations with standard error. **B** | Stratified analysis of power. Shown are TPR for 1 million reads per cell for sample sizes $n=16$, $n=32$, $n=64$, $n=128$, $n=256$ and $n=512$ per group. Genes are grouped in five percentiles of mean expression with lines representing the median TPR of 100 simulations.

Figure S10 (related to Figure 6) | **A-D** Batch effects **A** For each method, we test for differential expression between random subsets of 25 cells per group (left box) and subsets of 25 cells of each batch (right box) in 20 permutations using limma. Shown are the number of significantly differentially expressed genes (FDR <0.01) as boxplots. **B** Sensitivity is shown as the number of detected genes (≥ 1 count) per batch. **C** Accuracy is shown per batch as the correlation coefficient of observed expression (TPM/UPM) to annotated ERCC molecule numbers. **D** Precision is shown per batch as the Extra-Poisson Variability for the common 13,361 genes. For 3' counting methods, UMI quantification is shown. The distribution was only shown between values of 0 and 3 to make differences more visible. **D** Cell cycle analysis. For each method, we show the coefficient of variation (CV) for a set of 19 cell cycle genes previously found to be variable in 2i/LIF cultured mESCs (Kolodziejczyk, 2015) (left violin) compared to 19 ribosomal and pluripotency genes. Numbers above the violins indicate p-values of a t-test between the two groups.

Supplementary Tables

Method	CEL-seq2/C1	Drop-seq	MARS-seq	SCRB-seq	Smart-seq/C1	Smart-seq2
Single-cell isolation	automated in the C1 system	droplets	FACS	FACS	automated in the C1 system	FACS
ERCC spike-ins	yes	no	yes	yes	yes	yes
UMI	6 bp	8 bp	8 bp	10 bp	no	no
Full-length coverage	no	no	no	no	yes	yes
1st strand synthesis	oligo-dT	oligo-dT	oligo-dT	oligo-dT	oligo-dT	oligo-dT
2nd strand synthesis	RNAseH / DNA Pol	template switching	RNAseH / DNA Pol	template switching	template switching	template switching
Amplification	IVT	PCR	IVT	PCR	PCR	PCR
Imaging of cells possible	yes	no	no	no	yes	no
Protocol usable for bulk	yes	no	yes	yes	yes	yes
Sequencing	paired-end	paired-end	paired-end	paired-end	single-end	single-end
Library cost /cell	~9.5€	~0.1€	~1.3€	~2€	~25€	~3/30*

Table S1 (related to Figure 2): Overview of single-cell RNA-seq methods.

* in-house produced Tn5 / commercial Tn5

powsimR: Power analysis for bulk and single cell RNA-seq experiments

This is a pre-copyedited, author-produced version of an article accepted for publication in *Bioinformatics* following peer review. The version of record “Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: Power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*. 2017” is available online at:

<https://doi.org/10.1093/bioinformatics/btx435>.

Gene Expression

powsimR: Power analysis for bulk and single cell RNA-seq experiments

Beate Vieth^{1,*}, Christoph Ziegenhain¹, Swati Parekh¹,
Wolfgang Enard¹ and Ines Hellmann^{1,*}

¹Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians University, Munich, Germany

*To whom correspondence should be addressed.

Associate Editor: Prof. Ivo Hofacker

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Power analysis is essential to optimize the design of RNA-seq experiments and to assess and compare the power to detect differentially expressed genes in RNA-seq data. PowsimR is a flexible tool to simulate and evaluate differential expression from bulk and especially single-cell RNA-seq data making it suitable for a priori and posterior power analyses. **Availability:** The R package and associated tutorial are freely available at <https://github.com/bvieth/powsimR>.

Contact: vieth@bio.lmu.de, hellmann@bio.lmu.de

Supplementary information: Supplementary data are available at *Bioinformatics*. online.

1 Introduction

RNA-sequencing (RNA-seq) is an established method to quantify levels of gene expression genome-wide (Mortazavi *et al.*, 2008). Furthermore, the recent development of very sensitive RNA-seq protocols, such as Smart-seq2 and CEL-seq (Picelli *et al.*, 2014; Hashimshony *et al.*, 2012) allows transcriptional profiling at single-cell resolution and droplet devices make single cell transcriptomics high-throughput, allowing to characterize thousands or even millions of single cells (Zheng *et al.*, 2017; Macosko *et al.*, 2015; Klein *et al.*, 2015).

Even though technical possibilities are vast, scarcity of sample material and financial consideration are still limiting factors (Ziegenhain *et al.*, 2017), so that a rigorous assessment of experimental design remains a necessity (Auer and Doerge, 2010; Conesa *et al.*, 2016). The number of replicates required to achieve the desired statistical power is mainly determined by technical noise and biological variability (Conesa *et al.*, 2016) and both are considerably larger if the biological replicates are single cells. Crucially, it is common that genes are detected in only a subset of cells and such dropout events are thought to be rooted in the stochasticity of single-cell library preparation (Kharchenko *et al.*, 2014). Thus dropouts in single-cell RNA-seq are not a pure sampling problem that can be solved by deeper sequencing (Bacher and Kendziorski, 2016). In order to model dropout rates it is absolutely necessary to model the mean-variance relationship inherent in RNA-seq data. Even though current power assessment tools use the negative binomial or similar models that have an inherent

mean-variance relationship, they do not explicitly estimate and model the observed relationship, but rather draw mean and variance separately (reviewed in Poplawski and Binder, 2017).

In powsimR, we have implemented a flexible tool to assess power and sample size requirements for differential expression (DE) analysis of single cell and bulk RNA-seq experiments. Even though powsimR does not evaluate clustering of cells, we believe that powsimR can provide information also for RNA-seq experiment with unlabeled cells: The power for cluster analysis should be proportional to the power to detect differentially expressed genes. For our read count simulations, we (1) reliably model the mean, dispersion and dropout distributions as well as the relationship between those factors from the data. (2) Simulate read counts from the empirical mean-variance- and dropout relations, while offering flexible choices of the number of differentially expressed genes, effect sizes and DE testing method. (3) Finally, we evaluate the power over various sample sizes. We use the embryonic stem cell data from Kolodziejczyk *et al.* (2015) to illustrate powsimR's utility to plan and evaluate RNA-seq experiments.

2 powsimR

2.1 Estimation of RNA-seq Characteristics

An important step in the simulation framework is the reliable representation of the characteristics of the observed data. In agreement with others (Grün *et al.*, 2014; Mi *et al.*, 2015; Lun *et al.*, 2016), we find that the read distribution for most genes is sufficiently captured by the negative binomial. We analyzed 18 single cell datasets using unique molecular identifiers (UMIs) to control for amplification duplicates and 20 without duplicate control. The negative binomial provides an adequate fit for 54% of the

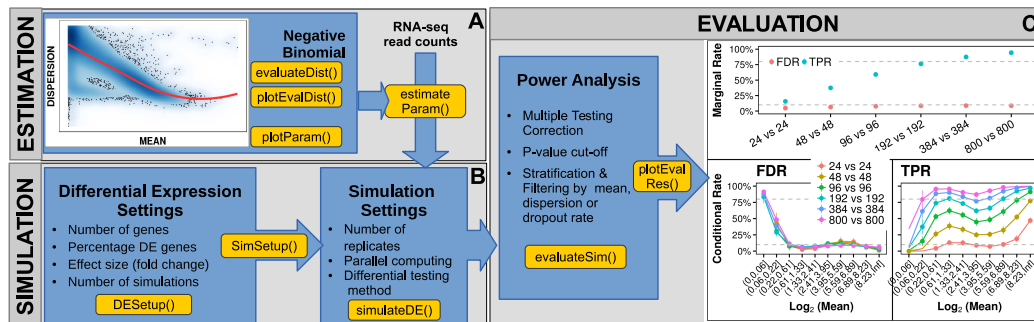


Fig. 1. powsimR schematic overview. A) The mean-dispersion relationship is estimated from RNA-seq data, which can be either single cell or bulk data. The user can provide their own count tables or one of our five example data sets and choose whether to fit a negative binomial or a zero-inflated negative binomial. The plot shows the mean-dispersion estimated, assuming a negative binomial for the Kolodziejczyk-data, the red line is the loess fit, that we later use for the simulations. B) These distribution parameters are then used to set-up the simulations. For better comparability, the parameters for the simulation of differential expression are set separately. C) Finally, the TPR and FDR are calculated. Both can be either returned as marginal estimates per sample configuration (top), or stratified according to the estimates of mean expression, dispersion or dropout-rate (bottom).

genes for the non-UMI-methods and 39% of the genes for UMI-methods, while the zero-inflated negative binomial was only adequate for 2.8% of the non-UMI-methods. In contrast, for the UMI-methods a simple Poisson distribution fits well for some studies (Ziegenhain *et al.*, 2017; Soumillon *et al.*, 2014) (Supplementary File S2). Furthermore, when comparing the fit of the other commonly used distributions, the negative binomial was most often the best fitting one for both non-UMI (57%) and UMI-methods (66%), while the zero inflated negative binomial improves the fit for only 19% and 1.6% (Supplementary Figure S4). Therefore the default sampling distribution in powsimR is the negative binomial (Figure 1), however the user has also the option to choose the zero-inflated negative binomial.

2.2 Simulation of Read Counts and Differential Expression

Simulations in powsimR can be based on provided data or on user-specified parameters. We first draw the mean expression for each gene. The expected dispersion given the mean is then determined using a locally weighted polynomial regression fit of the observed mean-dispersion relationship and to capture the variability of the observed dispersion estimates, a local variability prediction band ($\sigma = 1.96$) is applied to the fit (Figure 1A). Note, that using the fitted mean-dispersion spline is the feature that critically distinguishes powsimR from other simulation tools that draw the dispersion estimate for a gene independently of the mean. Our explicit model of mean and dispersion across genes allows us to reproduce the mean-variance as well as mean-dropout relationship observed (Supplementary Figure S2, Supplementary File S2).

To simulate DE genes, the user can specify the number of genes as well as the fraction of DE genes as \log_2 fold changes (LFC). Here, we assume that the grouping of samples is correct. For the Kolodziejczyk data, we found that a narrow gamma distribution mimicked the observed LFC distribution well (Supplementary Figure S3). The set-up for the expression levels and differential expression can be re-used for different simulation instances, allowing an easier comparison of experimental designs.

Finally, the user can specify the number of samples per group as well as their relative sequencing depth and the number of simulations. The simulated count tables are then directly used for DE analysis. In powsimR, we have integrated 8 R-packages for DE analysis for bulk and single cell data (limma (Ritchie *et al.*, 2015), edgeR (Robinson *et al.*, 2010), DESeq2 (Love *et al.*, 2014), ROTS (Seyednasrollah *et al.*, 2015), baySeq (Hardcastle, 2016), DSS (Wu *et al.*, 2013), NOISeq (Tarazona *et al.*, 2015), EBSec (Leng *et al.*, 2013)) and five packages that were specifically developed for

single-cell RNA-seq (MAST (Finak *et al.*, 2015), scde (Kharchenko *et al.*, 2014), BPSC (Vu *et al.*, 2016), scDD (Korthauer *et al.*, 2016), monocle (Qiu *et al.*, 2017)). For a review on choosing an appropriate method for bulk data, we refer to the work of others e.g. Schurch *et al.* (2016). Based on our analysis of the single-cell data from Kolodziejczyk *et al.* (2015), using standard settings for each tool we found that MAST performed best for this dataset given the same simulations as compared to results of other DE-tools.

2.3 Evaluating Statistical Power

Finally, powsimR integrates estimated and simulated expression differences to calculate marginal and conditional error matrices. To calculate these matrices, the user can specify nominal significance levels, methods for multiple testing correction and gene filtering schemes. Amongst the error matrix statistics, the power (True Positive Rate; TPR) and the False Discovery Rate (FDR) are the most informative for questions of experimental design. For easy comparison, powsimR plots power and FDR for a list of sample size choices either conditional on the mean expression (Wu *et al.*, 2014) or simply as marginal values (Figure 1). For example for the Kolodziejczyk data, 384 single cells for each condition would be sufficient to detect $> 80\%$ of the DE genes with a well controlled FDR of 5%. Given the lower sample sizes actually used in Kolodziejczyk *et al.* (2015), our power analysis suggests that only 60% of all DE genes could be detected.

3 Conclusion

In summary, powsimR can not only estimate sample sizes necessary to achieve a certain power, but also informs about the power to detect DE in a data set at hand. We believe that this type of posterior analysis will become more and more important, if results from different studies are compared. Often enough researchers are left to wonder why there is a lack of overlap in DE-genes when comparing similar experiments. powsimR will allow the researcher to distinguish between actual discrepancies and incongruities due to lack of power.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through LMUexcellent and the SFB1243 (Subproject A14/A15).

References

- Auer, P. L. and Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics*, **185**(2), 405–416.
- Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, **17**, 63.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.
- Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Plic, M., Linsley, P. S., and Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**(1), 1–13.
- Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat. Methods*, **11**(6), 637–640.
- Hardcastle, T. J. (2016). Generalized empirical bayesian methods for discovery of differential data in high-throughput biology. *Bioinformatics*, **32**(2), 195–202.
- Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, **2**(3), 666–673.
- Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**(7), 740–742.
- Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**(5), 1187–1201.
- Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C. H., Ilicic, T., Henriksson, J., Natarajan, K. N., Tuck, A. C., Gao, X., Bühler, M., Liu, P., Marioni, J. C., and Teichmann, S. A. (2015). Single cell RNA-Sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, **17**(4), 471–485.
- Korthauer, K. D., Chu, L.-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., and Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.*, **17**(1), 222.
- Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M. G., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). EBSeq: an empirical bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**(8), 1035–1043.
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**(12), 550.
- Lun, A. T. L., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**, 75.
- Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martnersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**(5), 1202–1214.
- Mi, G., Di, Y., and Schafer, D. W. (2015). Goodness-of-fit tests and model diagnostics for negative binomial regression of RNA sequencing data. *PLoS One*, **10**(3), e0119254.
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**(7), 621–628.
- Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, **9**(1), 171–181.
- Poplawski, A. and Binder, H. (2017). Feasibility of sample size calculation for RNA-seq studies. *Brief. Bioinform.*
- Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with census. *Nat. Methods*, **14**(3), 309–315.
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**(7), e47.
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.
- Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., and Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*.
- Seyednasrollah, F., Rantanen, K., Jaakkola, P., and Elo, L. L. (2015). ROTs: reproducible RNA-seq biomarker detector—prognostic markers for clear cell renal cell cancer. *Nucleic Acids Res.*, page gkv806.
- Soumillon, M., Cacchiarelli, D., Semrau, S., and others (2014). Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*.
- Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., and Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.*, **43**(21), e140.
- Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., and Pawitan, Y. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*, **32**(14), 2128–2135.
- Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., Mburu, F. M., Mantalas, G. L., Sim, S., Clarke, M. F., and Quake, S. R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, **11**(1), 41–46.
- Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, **14**(2), 232–243.
- Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
- Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, **65**(4), 631–643.e4.

powsimR: Power analysis for bulk and single cell RNA-seq
experiments

SUPPLEMENTARY FILE 2

by

Beate Vieth¹, Christoph Ziegenhain¹, Swati Parekh¹, Wolfgang Enard¹ and Ines Hellmann¹

¹Anthropology & Human Genomics, Department of Biology II,
Ludwig-Maximilians University, Munich, Germany

1 Single Cell RNA-sequencing Datasets

We analyzed RNA-sequencing data of 8 published studies that utilized 9 different RNA-seq library preparation methods (Table S1). One of the major differences between the methods is the use of Unique Molecular Identifiers (UMIs) that allow for confident removal of PCR-duplicates (Grün et al. (2014); Ziegenhain et al. (2017)). For all datasets, we evaluated the fit of 5 different distributions, namely the Poisson, negative binomial (NB), zero-inflated negative binomial (ZINB) and Poisson (ZIP) and Beta-Poisson (BP). For the vast majority the NB would be the distribution of choice. This is especially true for the UMI-methods: Here no zero-inflation is needed for modeling the gene expression distribution. On the contrary, also a simple Poisson often provides the best fit.

Study	Species	Cell type	Cell classification	No. cells	Library preparation	UMI	Coverage
1 Buettner et al. 2015	Mouse	Embryonic stem cells	G1 cell cycle stage	96	SmartSeq/C1	no	full length
2 Buettner et al. 2015	Mouse	Embryonic stem cells	G2M cell cycle stage	96	SmartSeq/C1	no	full length
3 Buettner et al. 2015	Mouse	Embryonic stem cells	S cell cycle stage	96	SmartSeq/C1	no	full length
4 Islam et al. 2014	Mouse	Embryonic stem cells		96	STRT-UMI	yes	end sequencing
5 Islam et al. 2011	Mouse	Embryonic stem cells		48	STRT	no	end sequencing
6 Islam et al. 2011	Mouse	Mouse embryonic fibroblast		48	STRT	no	end sequencing
7 Kolodziejczk et al. 2015	Mouse	Embryonic stem cells	alternative 2i media + LIF	194	SmartSeq/C1	no	full length
8 Kolodziejczk et al. 2015	Mouse	Embryonic stem cells	serum + LIF	242	SmartSeq/C1	no	full length
9 Kolodziejczk et al. 2015	Mouse	Embryonic stem cells	standard 2i media + LIF	433	Smart-seq/C1	no	full length
10 Pollen et al. 2014	Human	Primary epidermal keratinocytes		80	SmartSeq/C1	no	full length
11 Pollen et al. 2014	Human	Induced pluripotent stem cells		48	SmartSeq/C1	no	full length
12 Pollen et al. 2014	Human	Cultured primary human neurons		32	SmartSeq/C1	no	full length
13 Pollen et al. 2014	Human	HCC1954 breast cancer cells		44	SmartSeq/C1	no	full length
14 Pollen et al. 2014	Human	HCC1954 B lymphoblastoid cells		34	SmartSeq/C1	no	full length
15 Pollen et al. 2014	Human	HL-60 human promyelocytic leukemia cells		108	SmartSeq/C1	no	full length
16 Pollen et al. 2014	Human	K-562 myelogenous leukemia cells		178	SmartSeq/C1	no	full length
17 Pollen et al. 2014	Human	Neural progenitor cells obtained by differentiation of iPS line		30	SmartSeq/C1	no	full length
18 Pollen et al. 2014	Human	Primary human neurons		16	SmartSeq/C1	no	full length
19 Pollen et al. 2014	Human	BJ Human Fibroblasts	early passage, p6	74	Smart-seq/C1	no	full length
20 Soumillon et al. 2014	Human	adipose-derived stem cells	1 day post-differentiation	6197	SCRB-seq	yes	end sequencing
21 Soumillon et al. 2014	Human	adipose-derived stem cells	2 days post-differentiation	1599	SCRB-seq	yes	end sequencing
22 Soumillon et al. 2014	Human	adipose-derived stem cells	3 days post-differentiation	2068	SCRB-seq	yes	end sequencing
23 Zheng et al. 2017	Human	Peripheral Blood Mononuclear Cells	CD19+ B Cells	10085	10XGenomics	yes	end sequencing
24 Zheng et al. 2017	Human	Peripheral Blood Mononuclear Cells	CD14+ Monocytes	2612	10XGenomics	yes	end sequencing
25 Zheng et al. 2017	Human	Peripheral Blood Mononuclear Cells	CD34+ Cells	9232	10XGenomics	yes	end sequencing
26 Zheng et al. 2017	Human	Peripheral Blood Mononuclear Cells	CD4+ T Helper Cells	11213	10XGenomics	yes	end sequencing
27 Zheng et al. 2017	Human	Peripheral Blood Mononuclear Cells	CD56+ NK Cells	8385	10XGenomics	yes	end sequencing
28 Zheng et al. 2017	Human	Peripheral Blood Mononuclear Cells	CD8+ Cytotoxic T Cells	10209	10XGenomics	yes	end sequencing
29 Zheng et al. 2017	Human	Peripheral Blood Mononuclear Cells	CD4+/CD45RO+ Memory T Cells	10224	10XGenomics	yes	end sequencing
30 Zheng et al. 2017	Human	Peripheral Blood Mononuclear Cells	CD8+/CD45RA+ Naive Cytotoxic T Cells	11953	10XGenomics	yes	end sequencing
31 Zheng et al. 2017	Human	Peripheral Blood Mononuclear Cells	CD4+/CD45RA+/CD25- Naive T Cells	10479	10XGenomics	yes	end sequencing
32 Zheng et al. 2017	Human	Peripheral Blood Mononuclear Cells	CD4+/CD25+ Regulatory T Cells	10263	10XGenomics	yes	end sequencing
33 Ziegenhain et al. 2017	Mouse	Embryonic stem cells		71	CEL-seq2	yes	end sequencing
34 Ziegenhain et al. 2017	Mouse	Embryonic stem cells		76	Drop-seq	yes	end sequencing
35 Ziegenhain et al. 2017	Mouse	Embryonic stem cells		65	MARS-seq	yes	end sequencing
36 Ziegenhain et al. 2017	Mouse	Embryonic stem cells		84	SCRB-seq	yes	end sequencing
37 Ziegenhain et al. 2017	Mouse	Embryonic stem cells		287	Smart-seq/C1	no	full length
38 Ziegenhain et al. 2017	Mouse	Embryonic stem cells		157	Smart-seq2	no	full length

Table S1: Key properties of the single cell RNA-seq experiments for distribution evaluation.

2 Distributional Fitting per Dataset

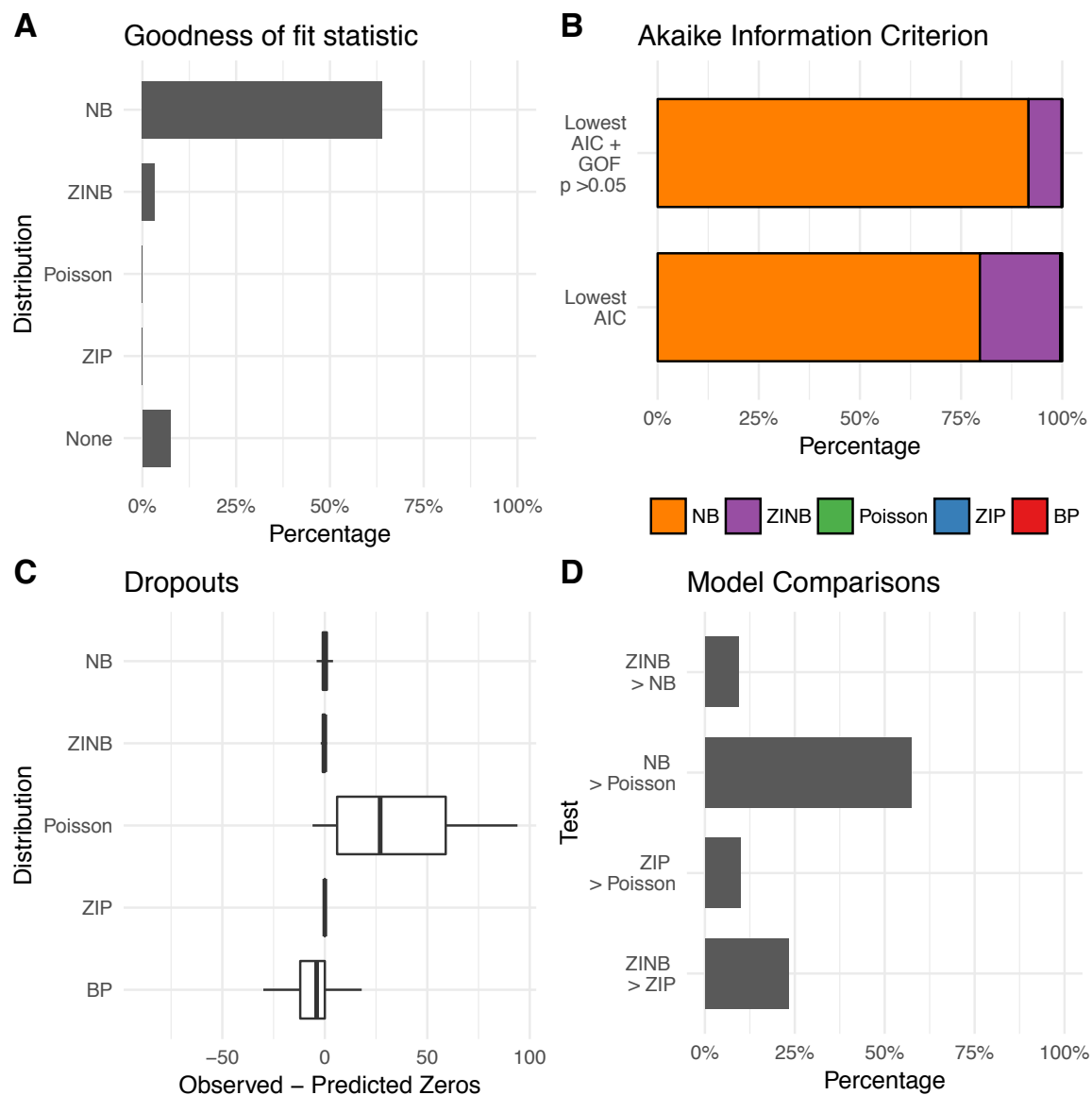


Figure S1: Buettner et al. 2015: Embryonic stem cells G1 cell cycle stage (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

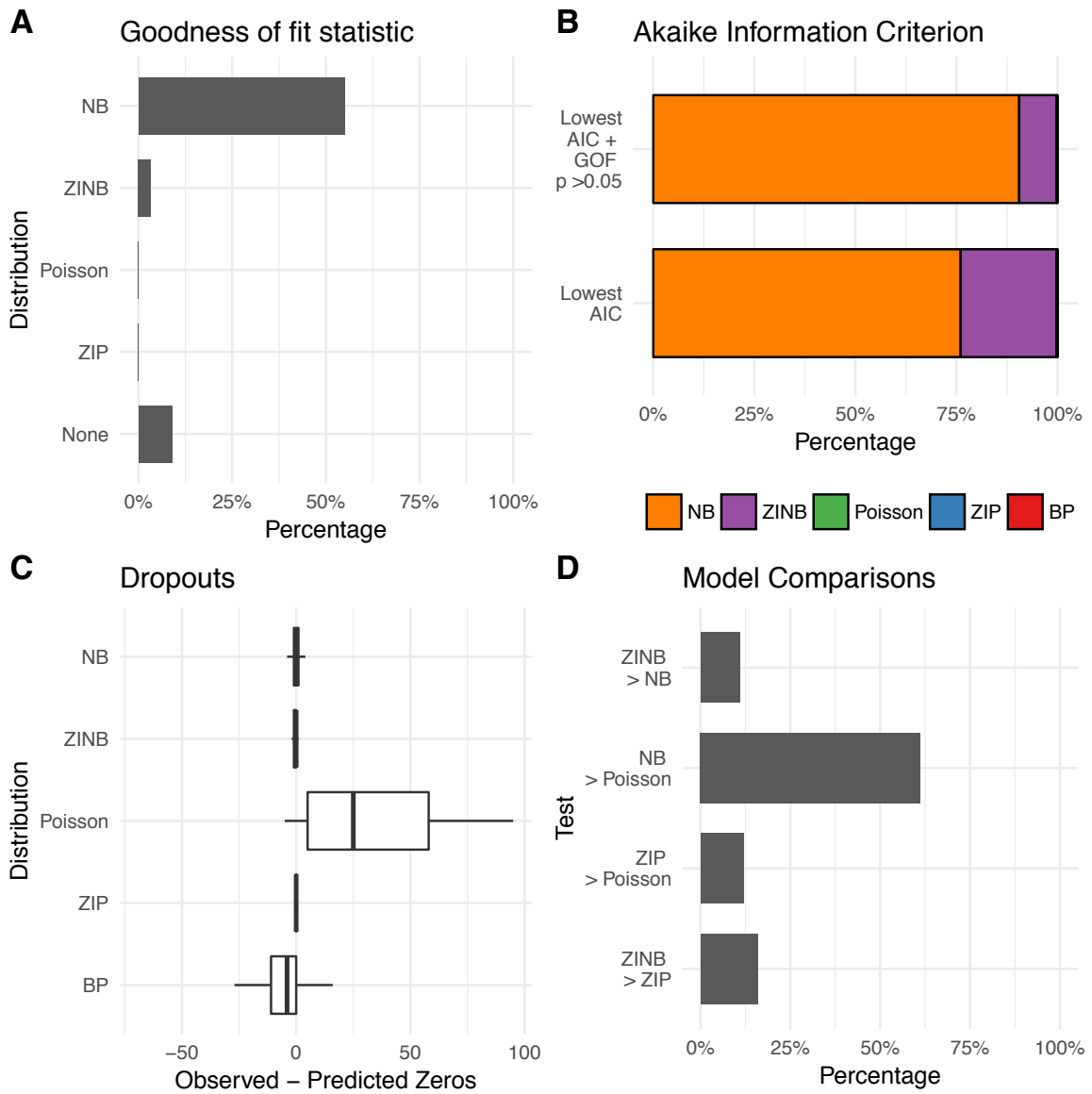


Figure S2: Buettner et al. 2015: Embryonic stem cells G2M cell cycle stage (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

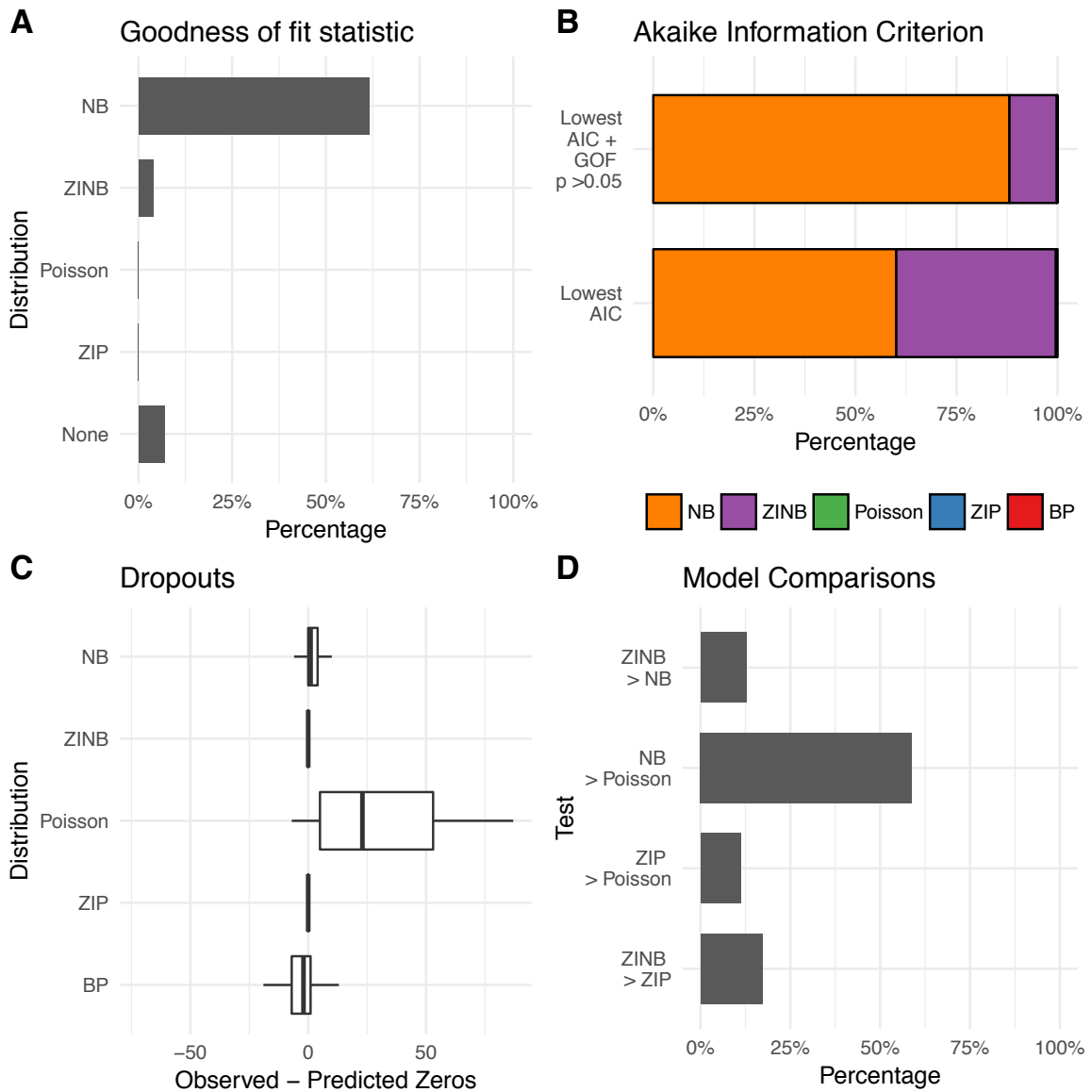


Figure S3: Buettner et al. 2015: Embryonic stem cells S cell cycle stage (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

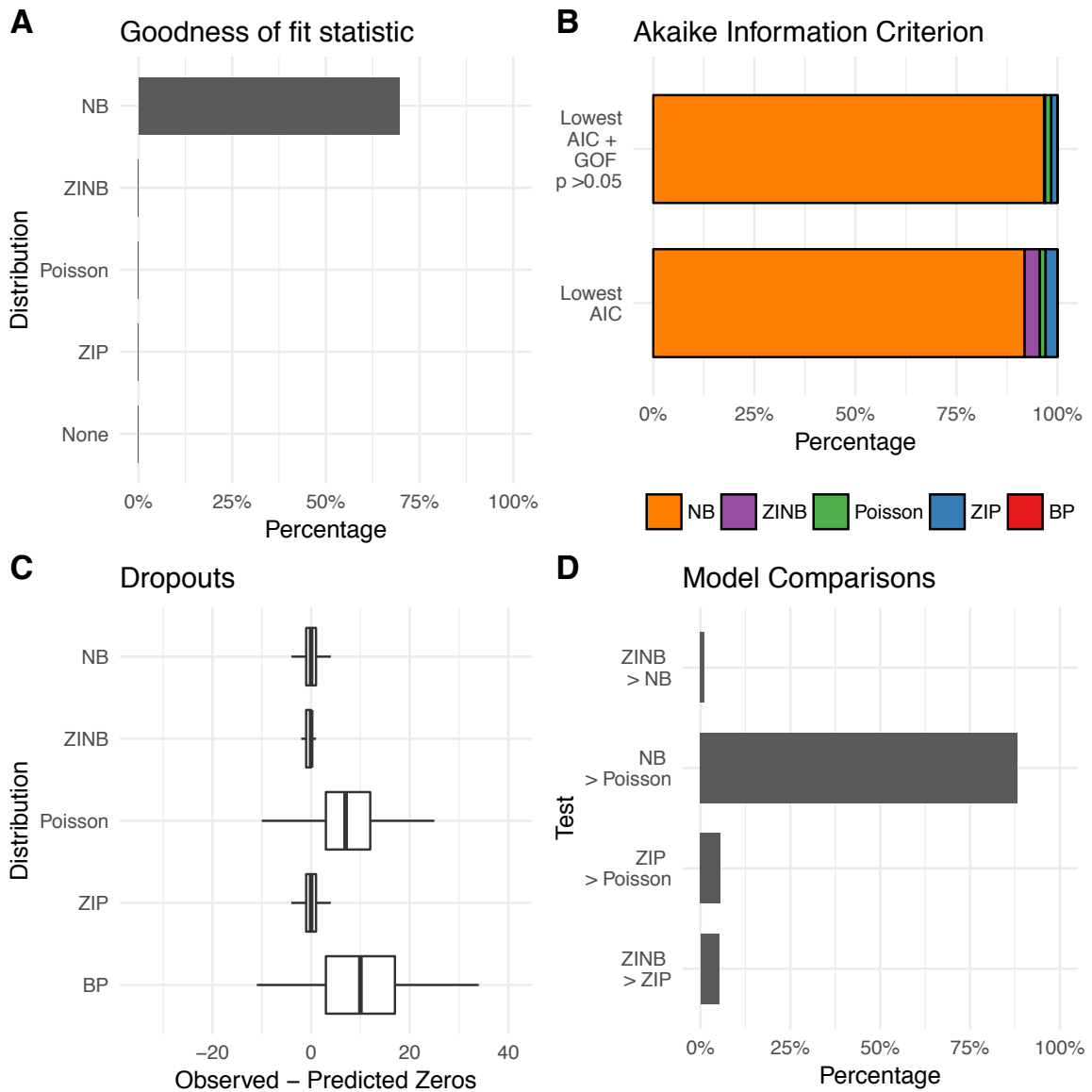


Figure S4: Islam et al. 2014: Embryonic stem cells (STRT-UMI). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

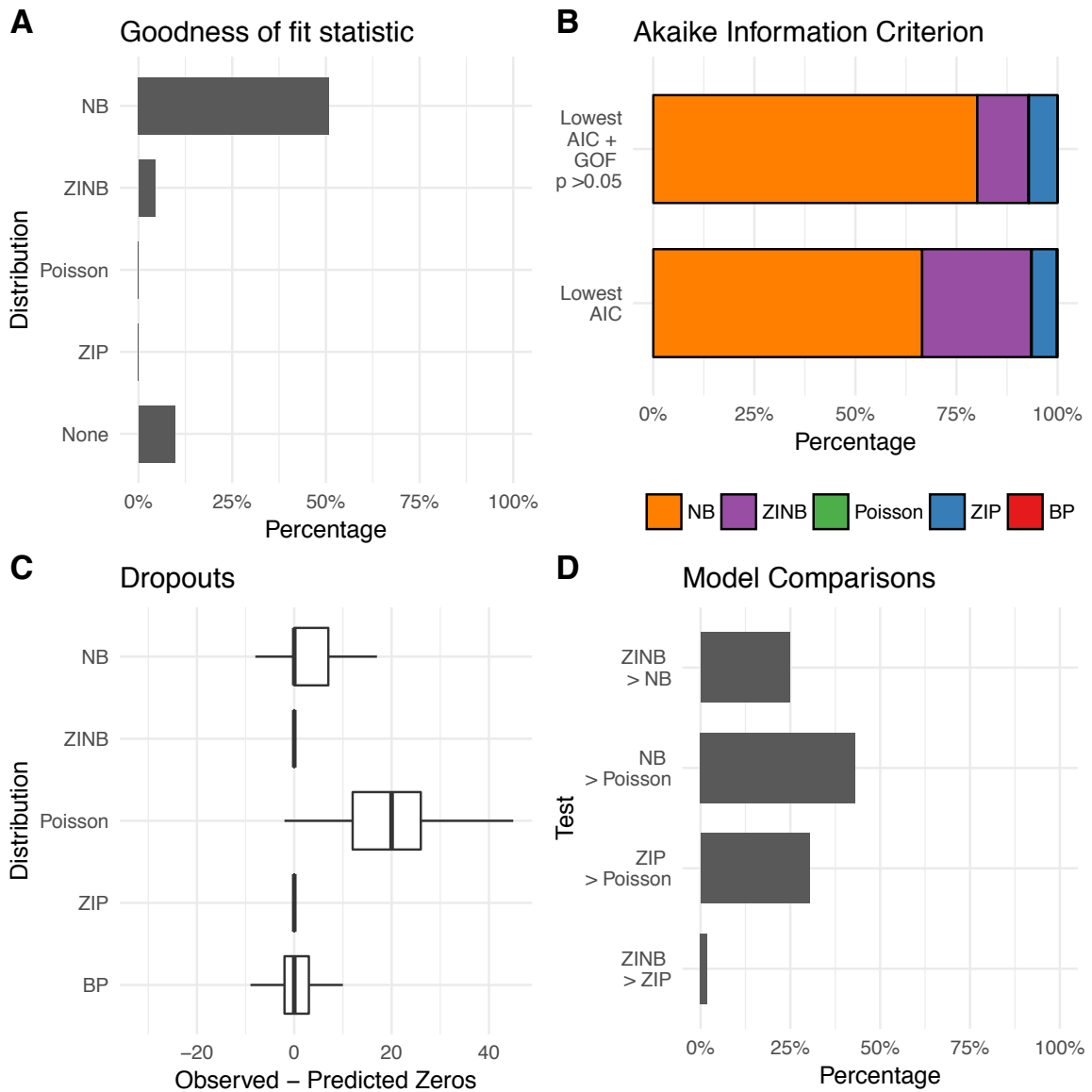


Figure S5: Islam et al. 2011: Embryonic stem cells (STRT). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

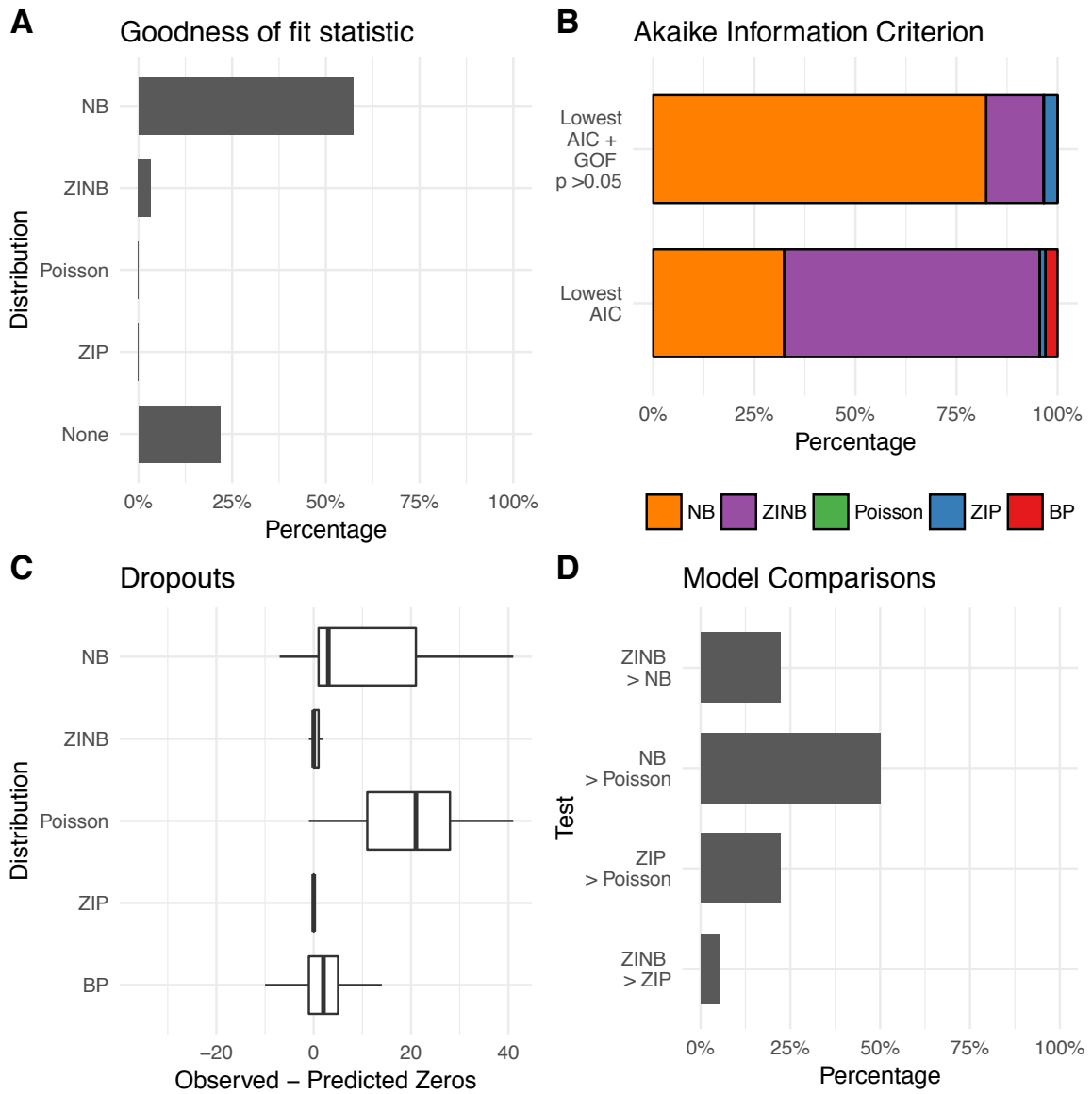


Figure S6: Islam et al. 2011: Mouse embryonic fibroblast (STRT). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

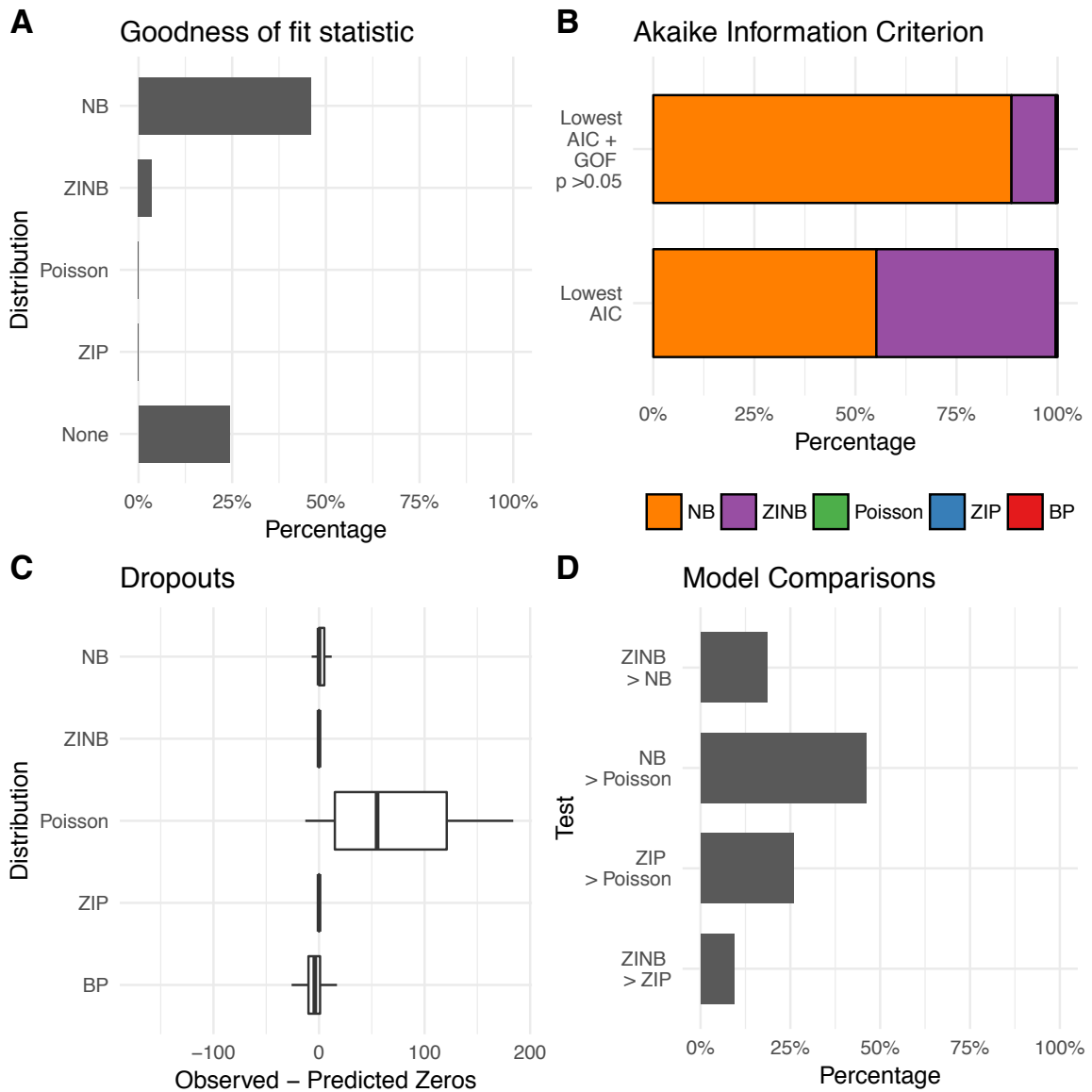


Figure S7: Kolodziejczk et al. 2015: Embryonic stem cells alternative 2i media + LIF (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

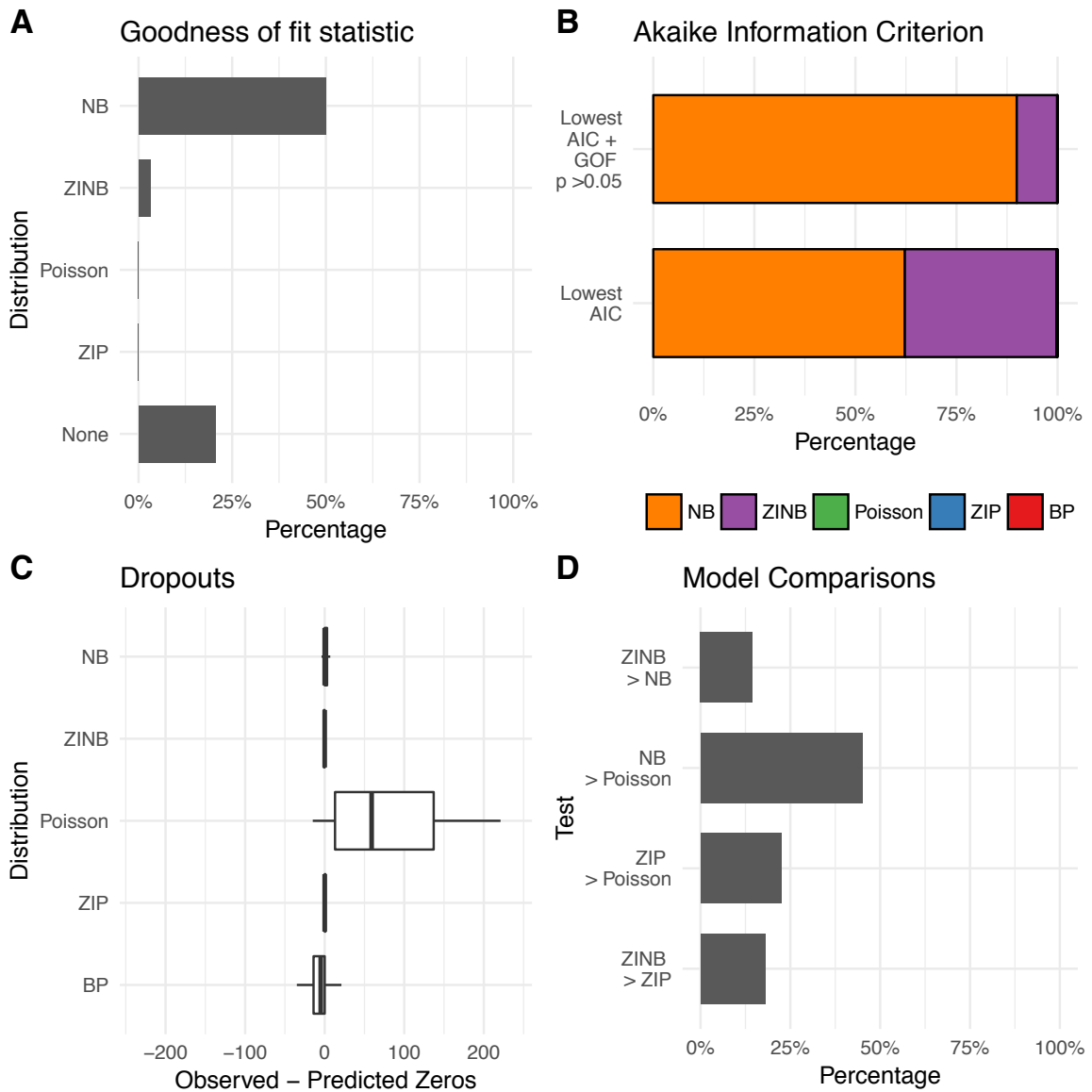


Figure S8: Kolodziejczk et al. 2015: Embryonic stem cells serum + LIF (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

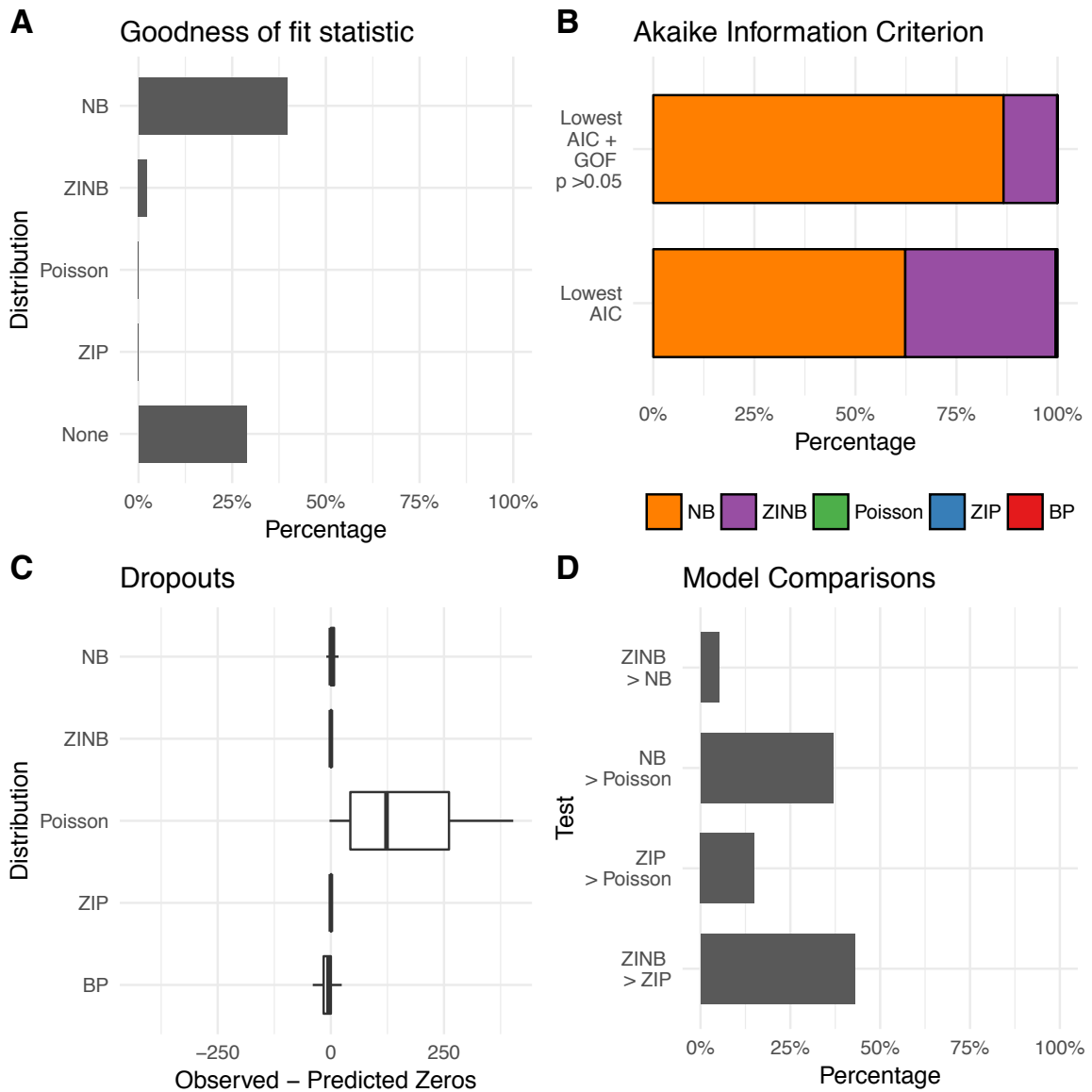


Figure S9: Kolodziejczk et al. 2015: Embryonic stem cells standard 2i media + LIF (Smart-seq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

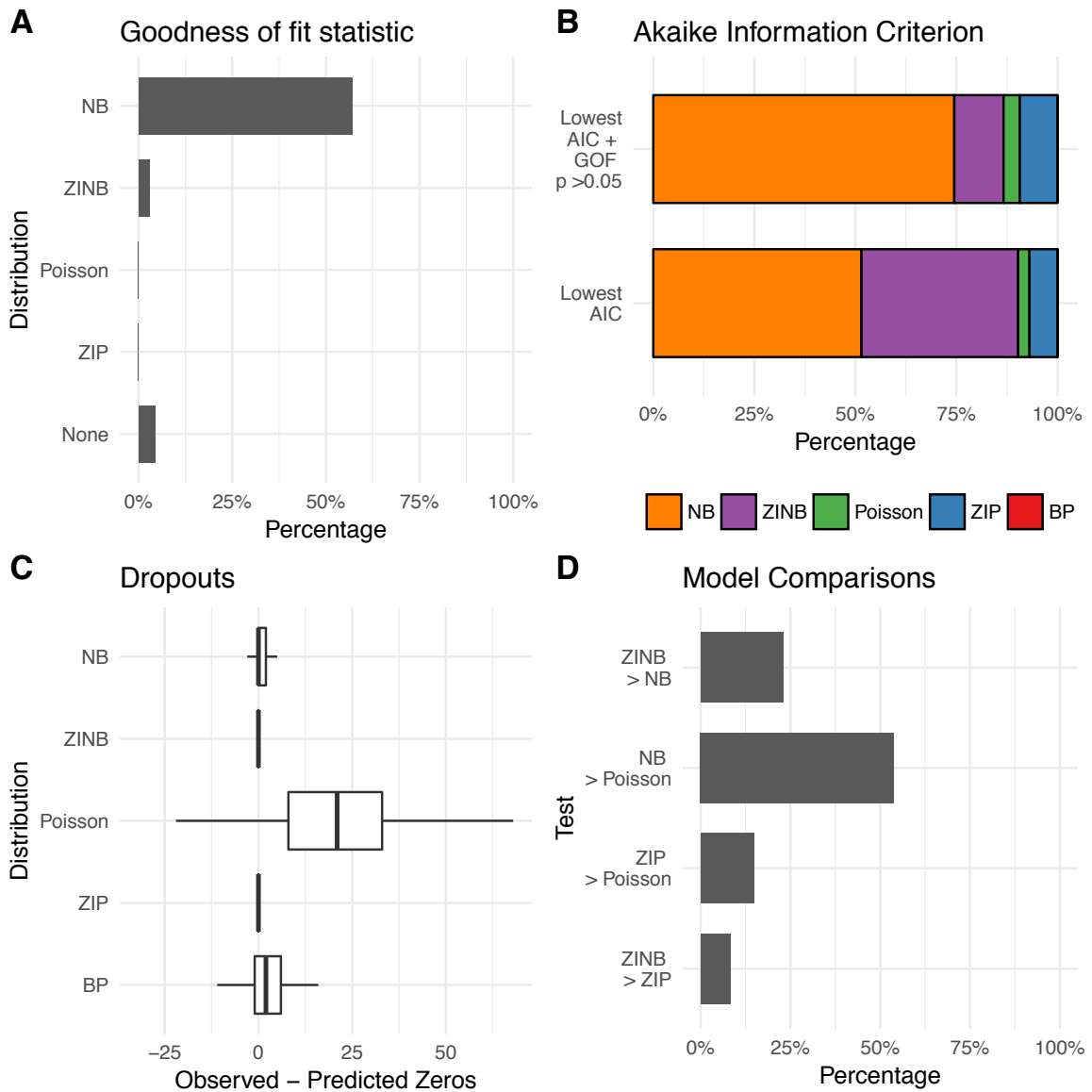


Figure S10: Pollen et al. 2014: Primary epidermal keratinocytes (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

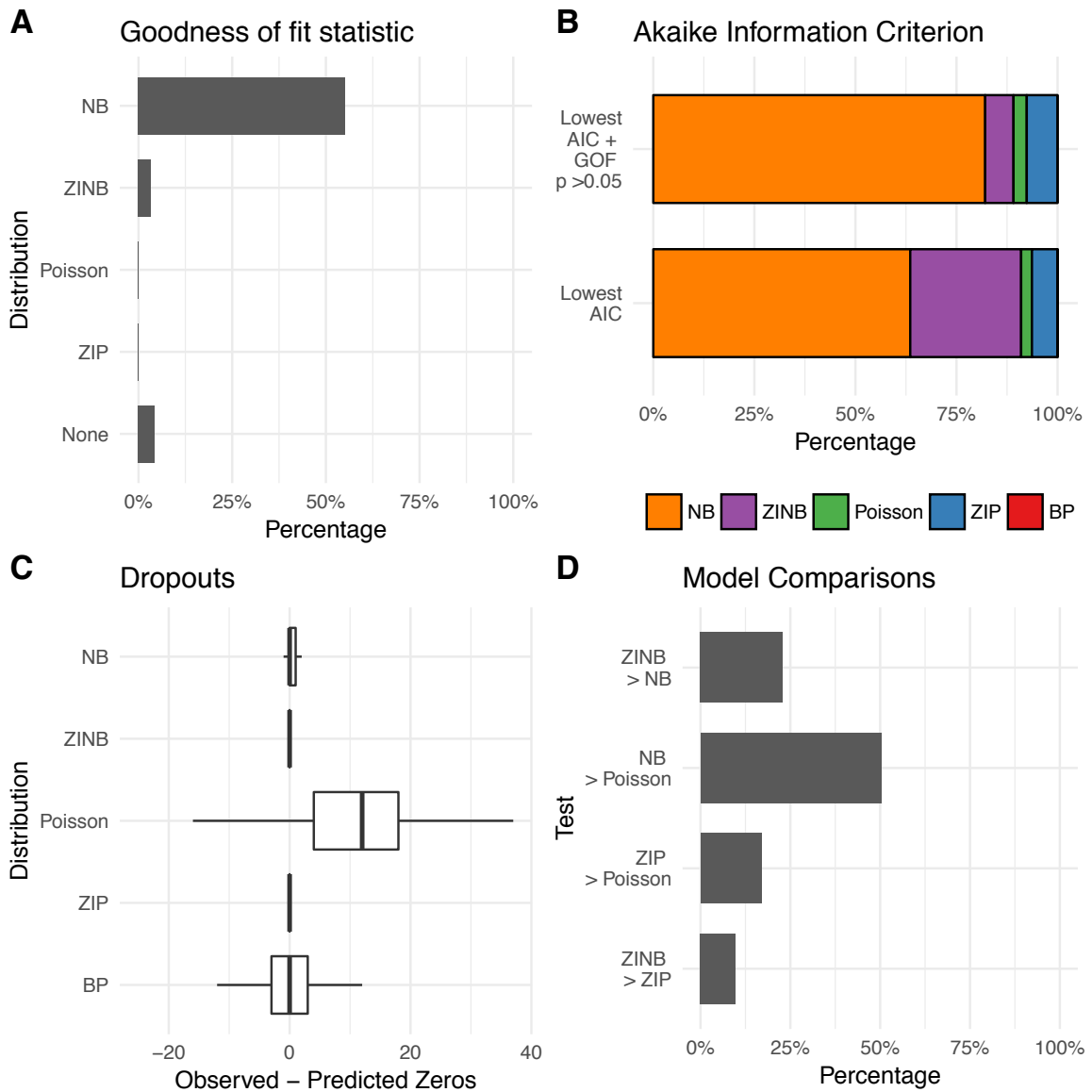


Figure S11: Pollen et al 2014: Induced pluripotent stem cells (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

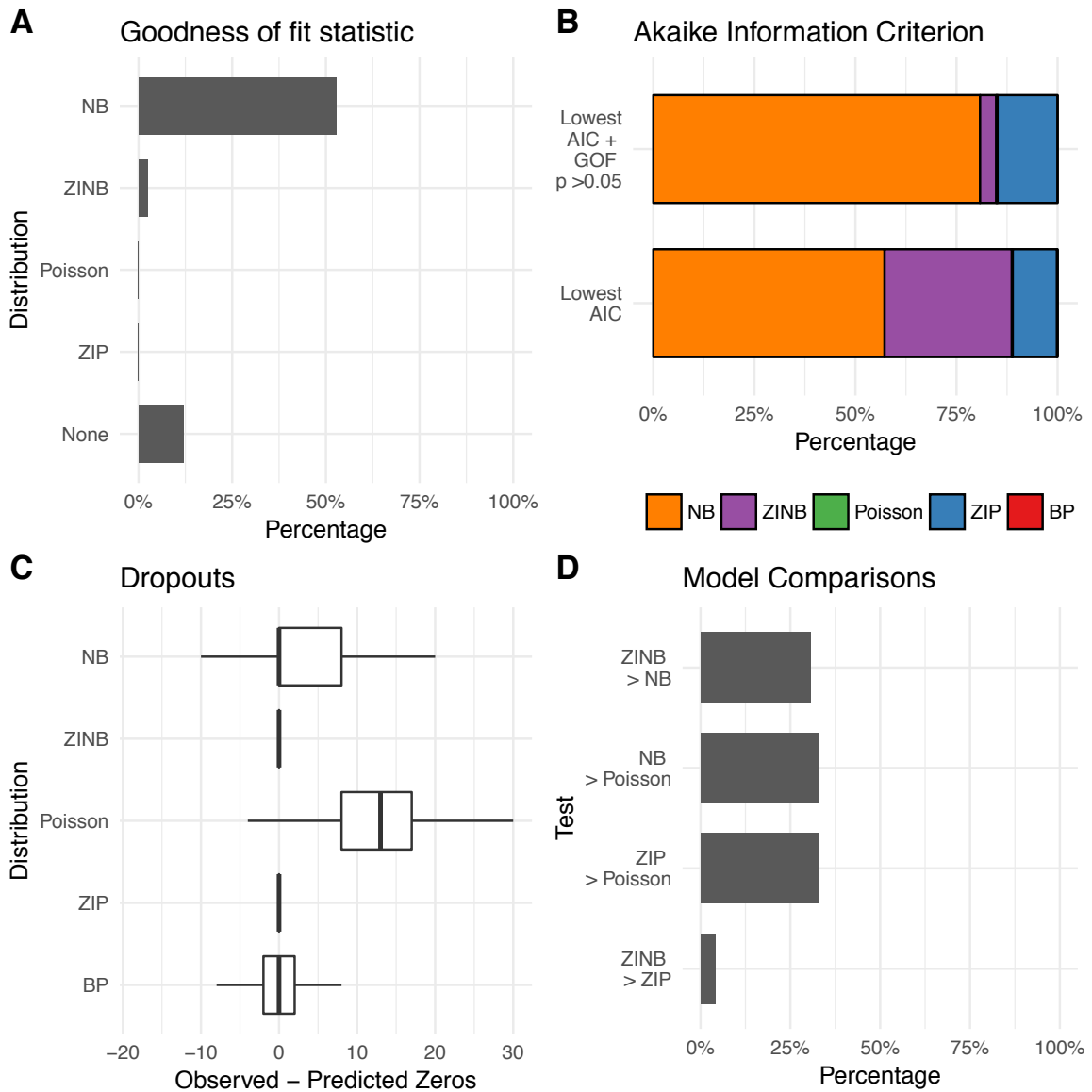


Figure S12: Pollen et al. 2014: Cultured primary human neurons (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

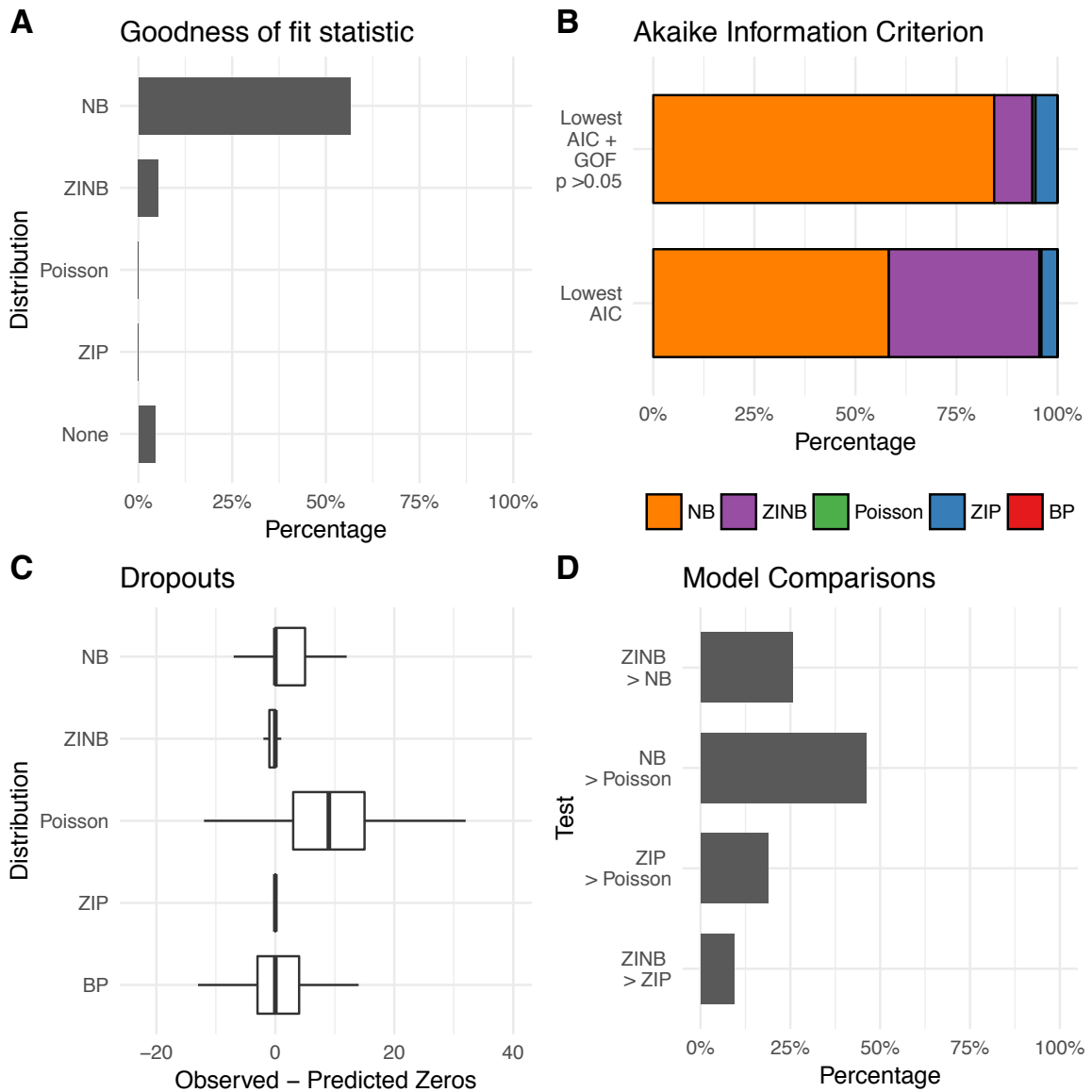


Figure S13: Pollen et al. 2014: HCC1954 breast cancer cells (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

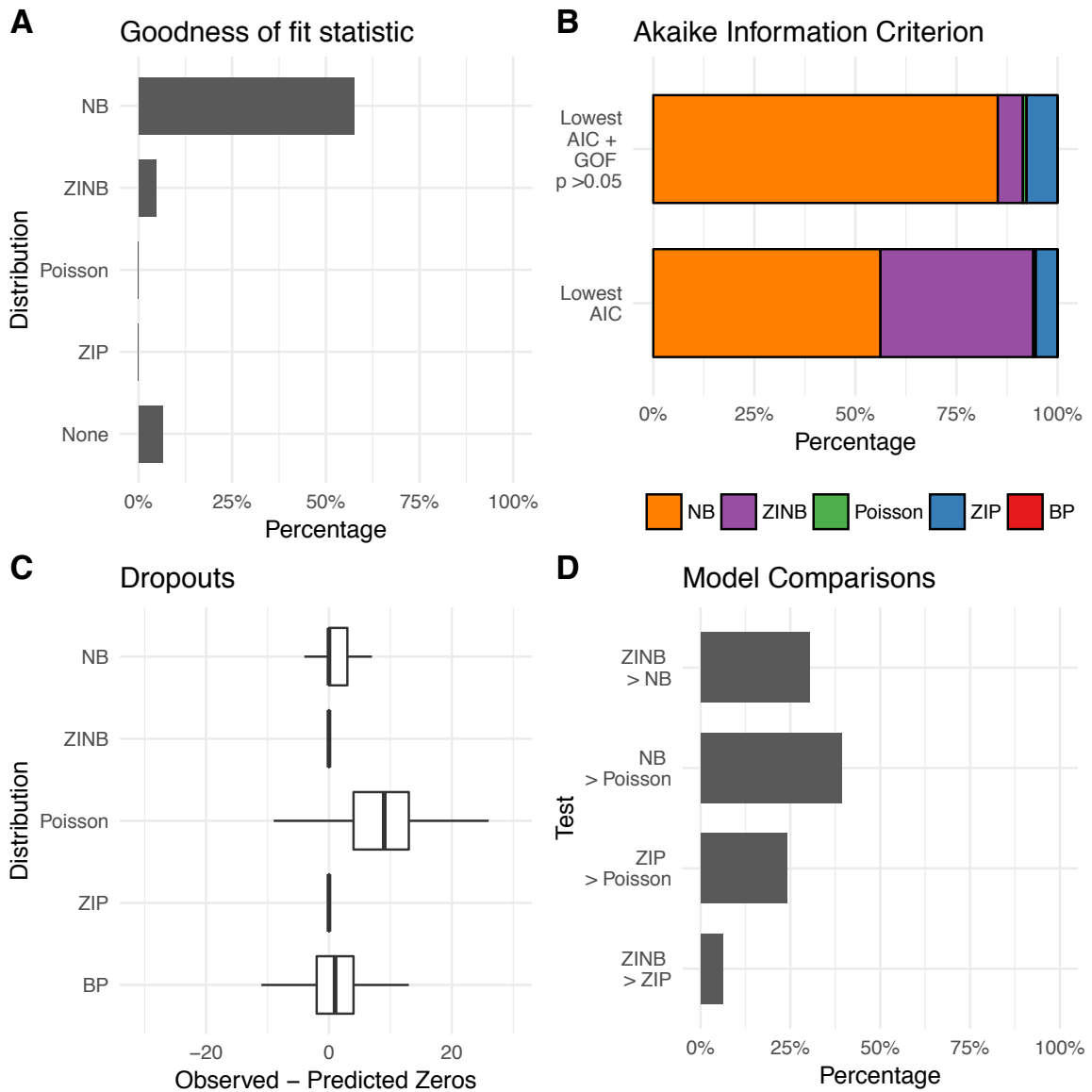


Figure S14: Pollen et al. 2014: HCC1954 B lymphoblastoid cells (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

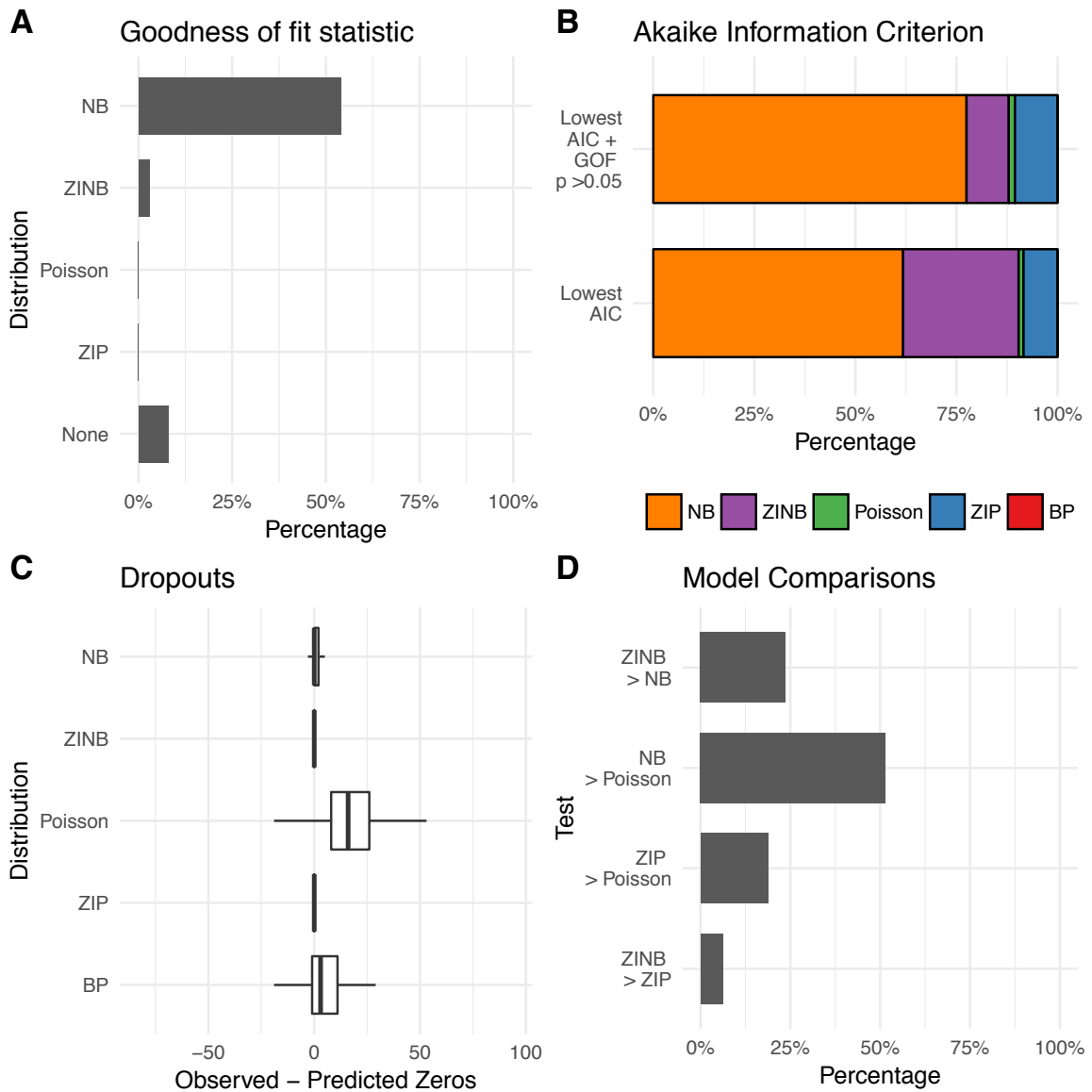


Figure S15: Pollen et al. 2014: HL-60 human promyelocytic leukemia cells (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

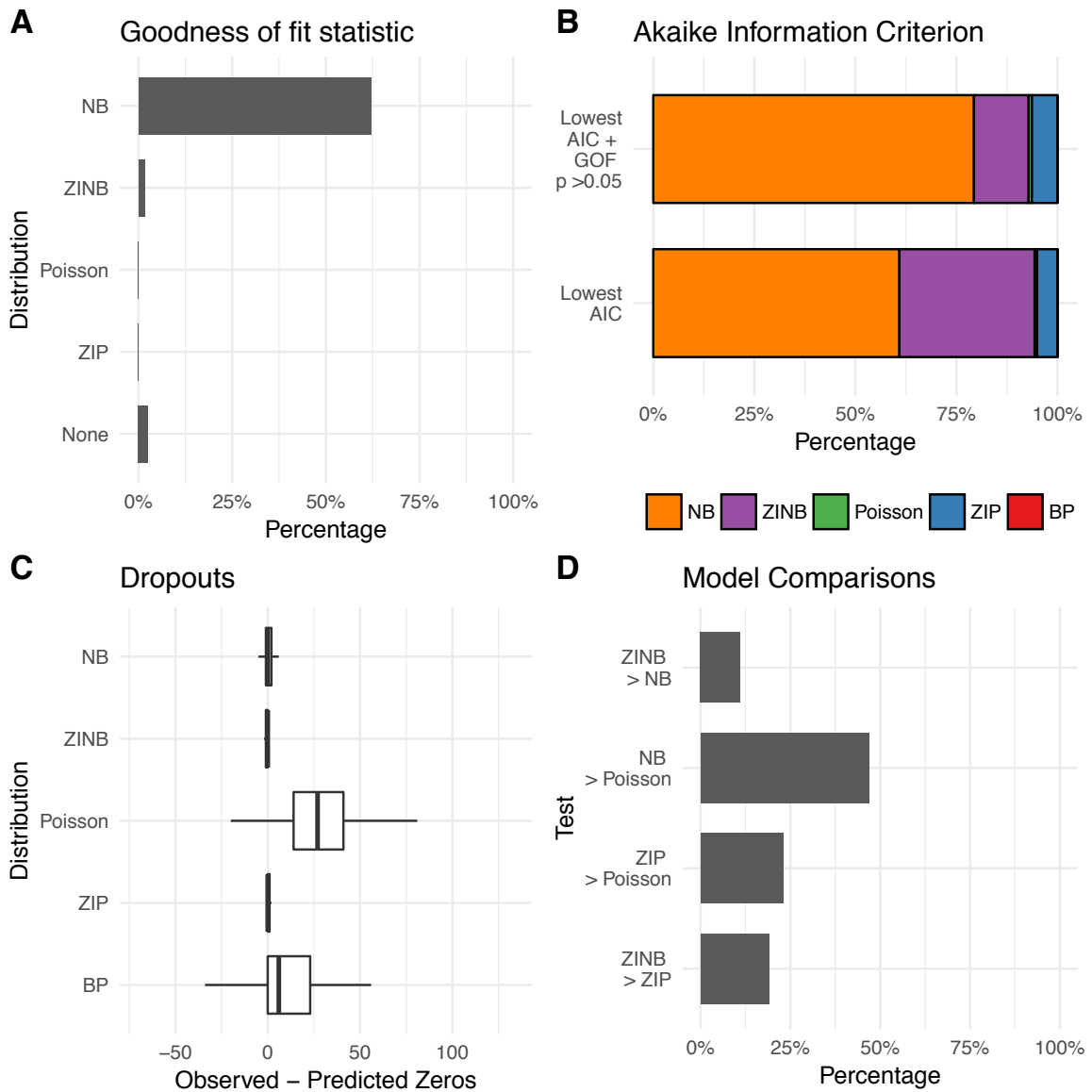


Figure S16: Pollen et al. 2014: K-562 myelogenous leukemia cells (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

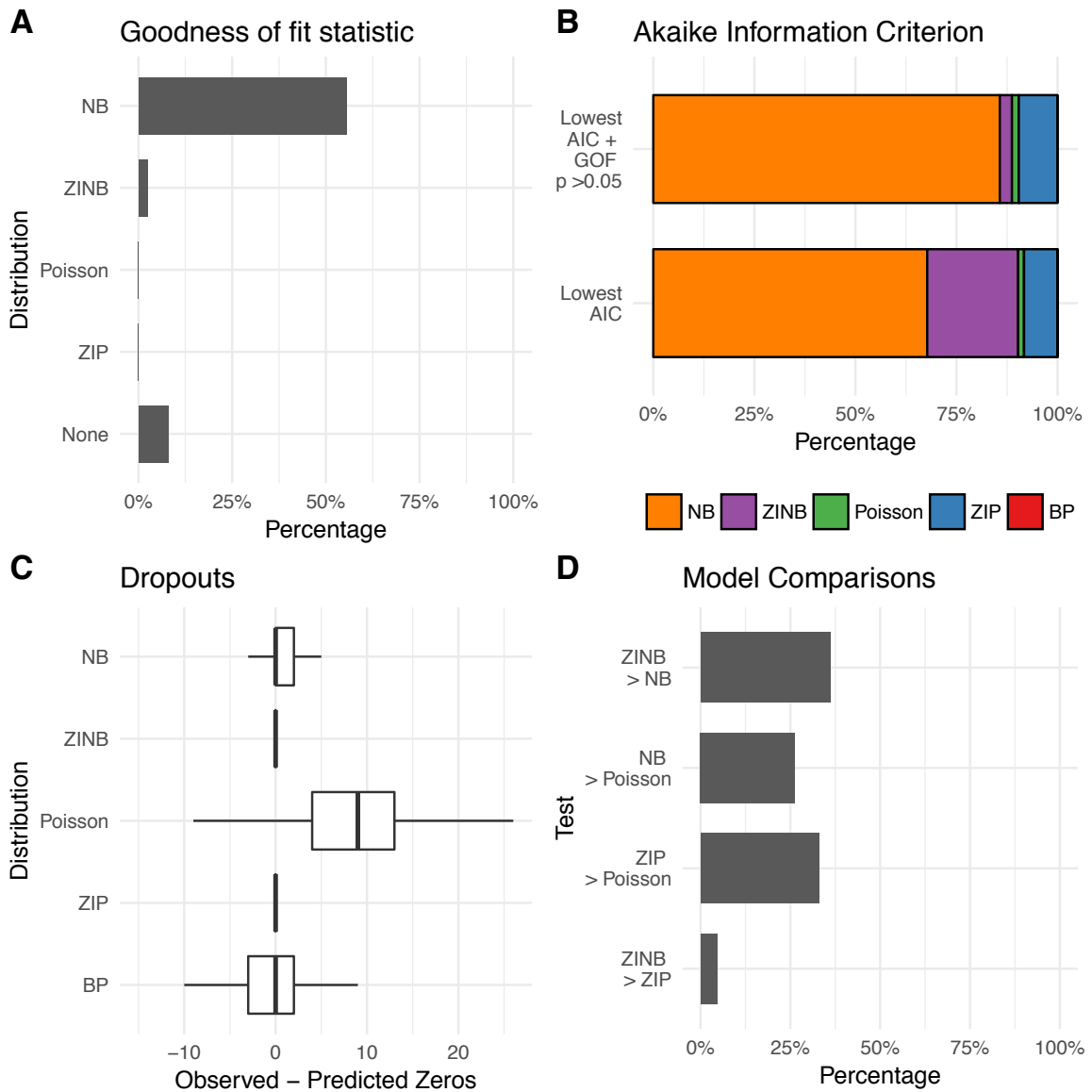


Figure S17: Pollen et al. 2014: Neural progenitor cells obtained by differentiation of iPS line (Smart-Seq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

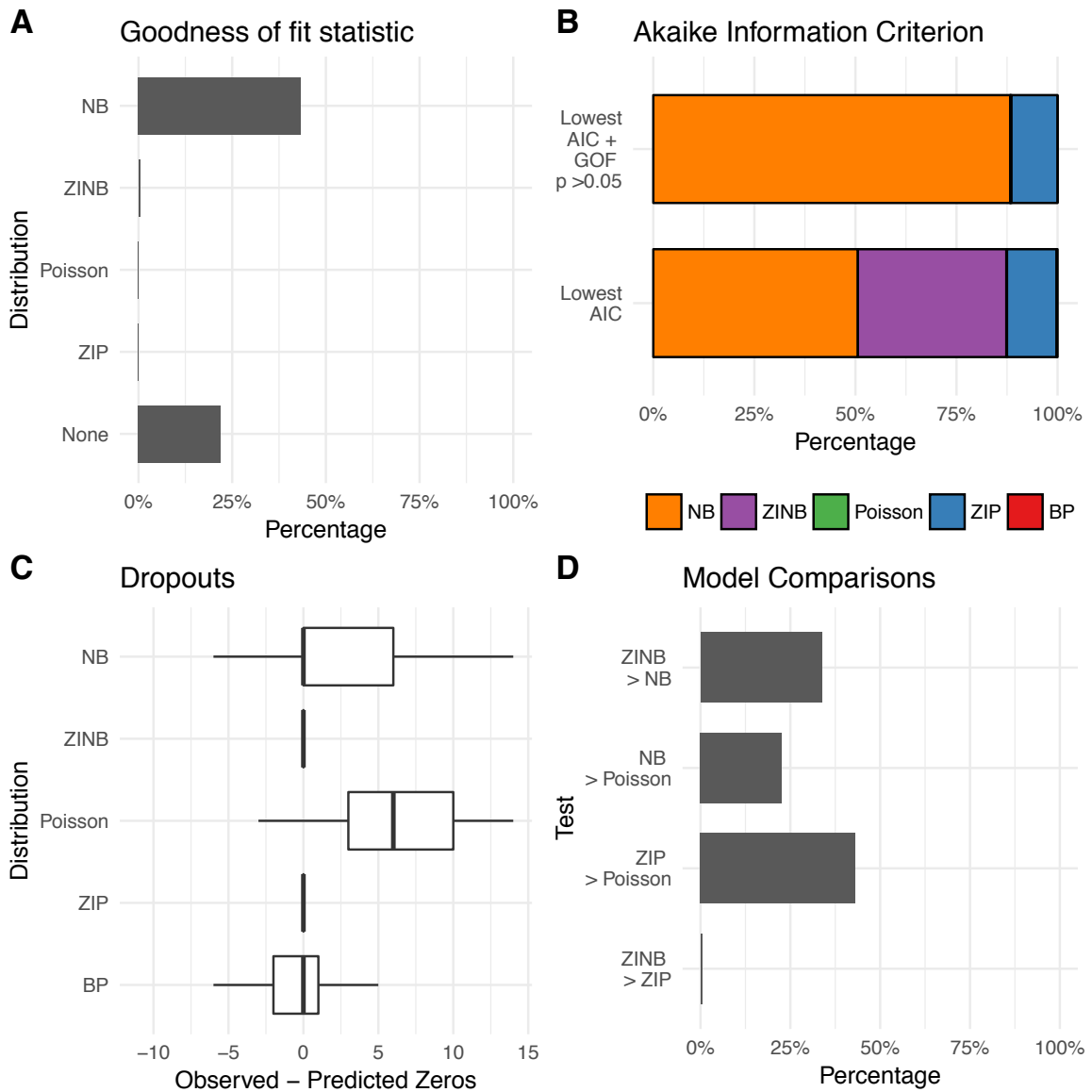


Figure S18: Pollen et al. 2014: Primary human neurons (SmartSeq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

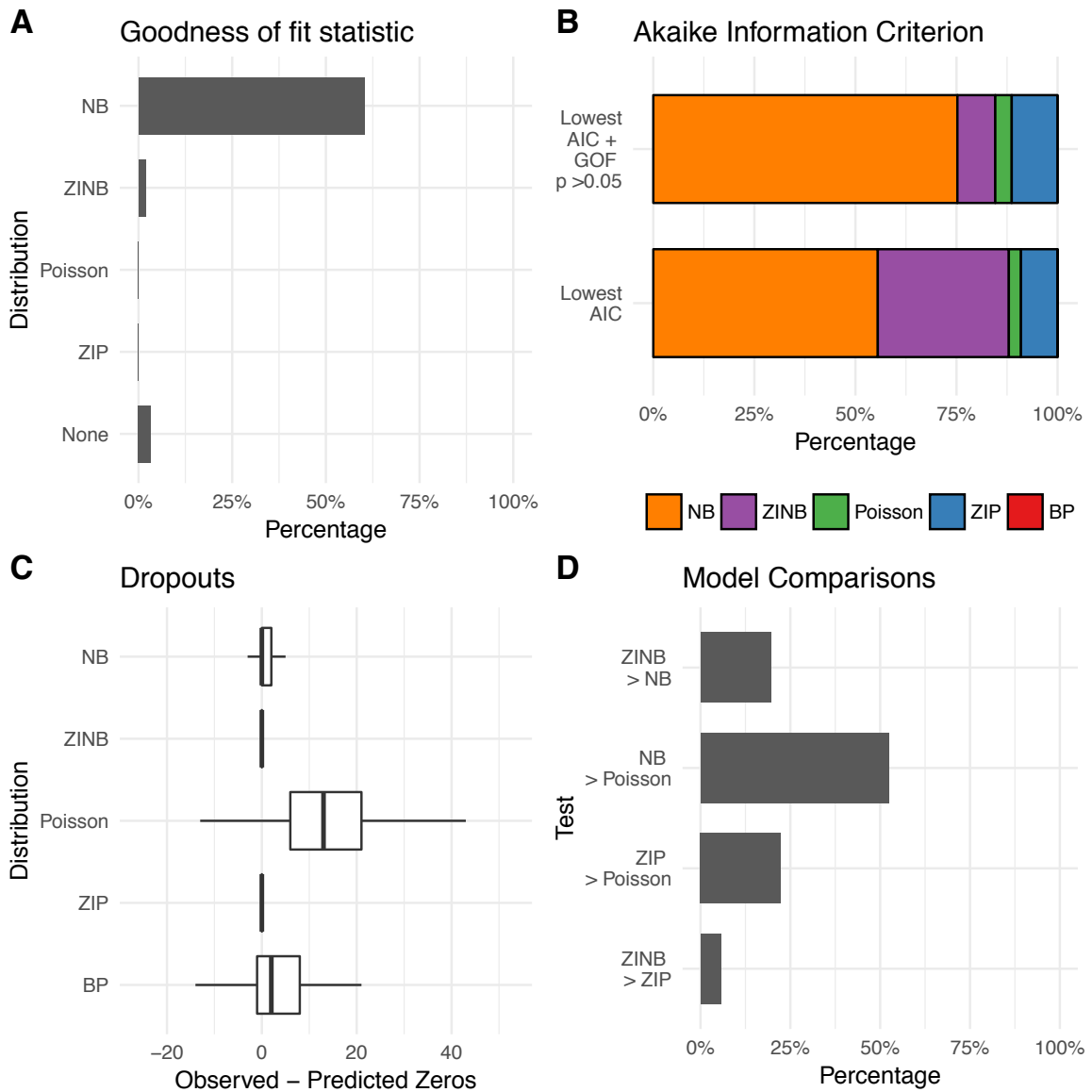


Figure S19: Pollen et al. 2014: BJ Human Fibroblasts early passage, p6 (Smart-seq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

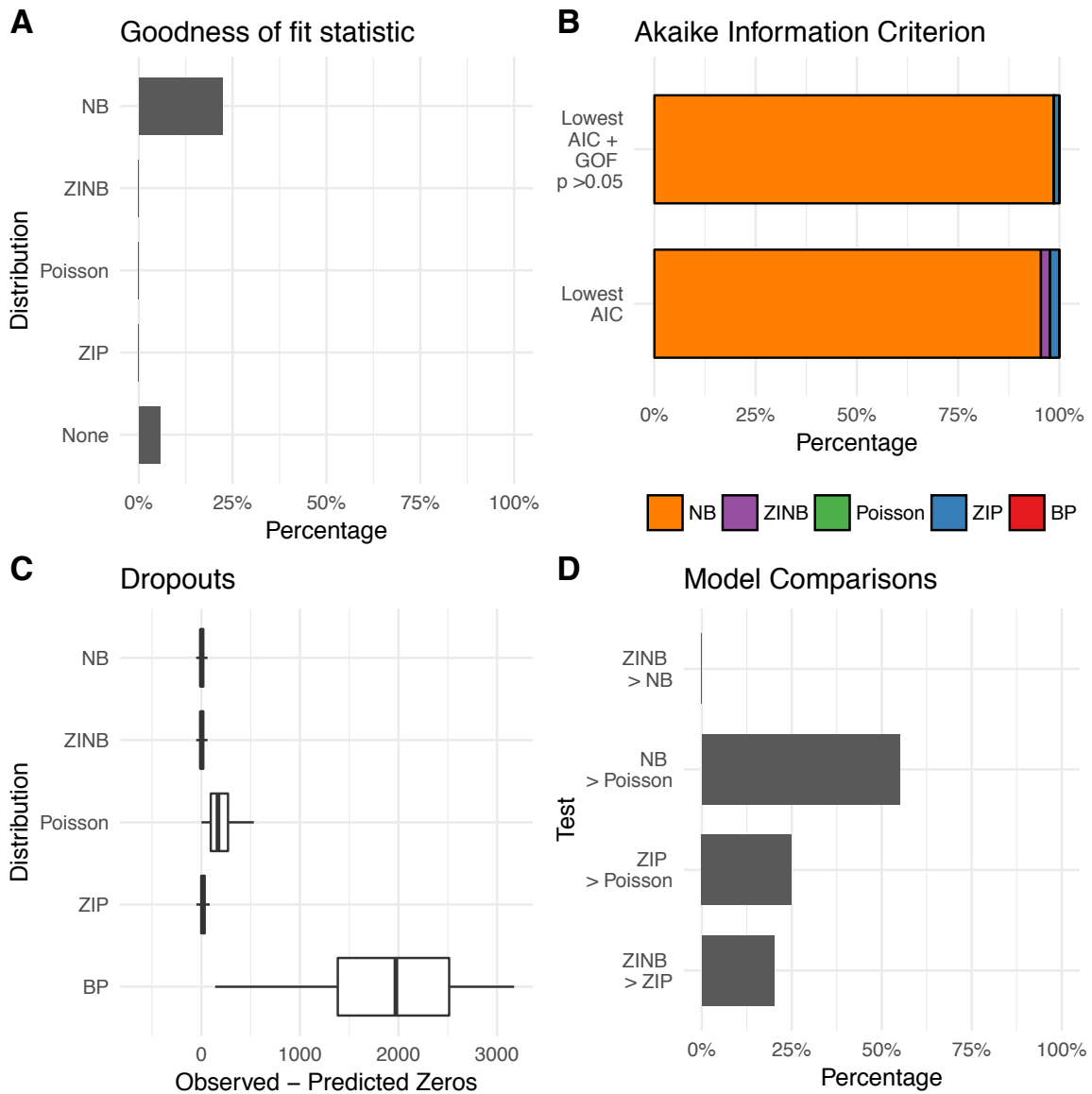


Figure S20: Soumillon et al. 2014: adipose-derived stem cells 1 day post-differentiation (SCR-seq). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

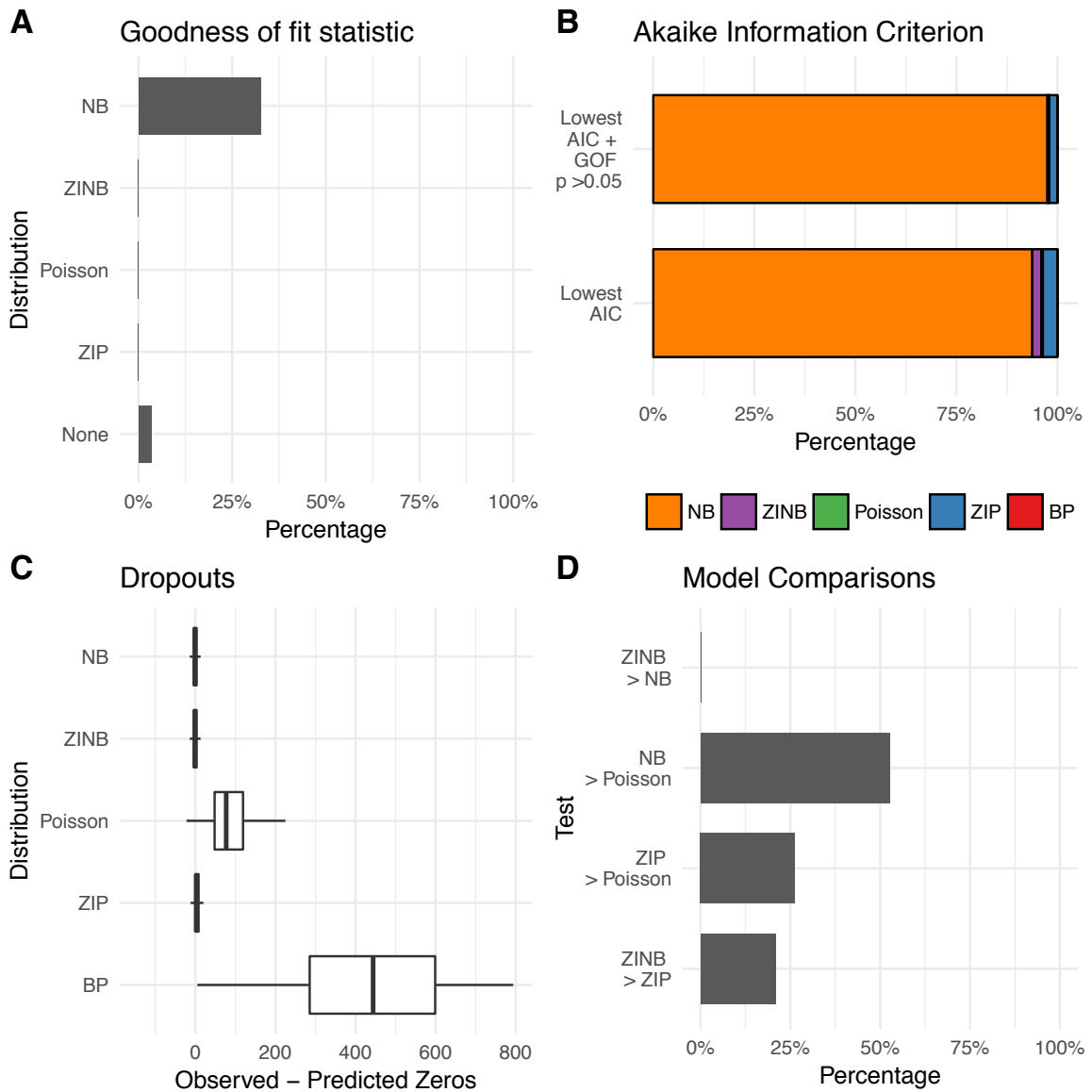


Figure S21: Soumillon et al. 2014: adipose-derived stem cells 2 days post-differentiation (SCR-seq). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

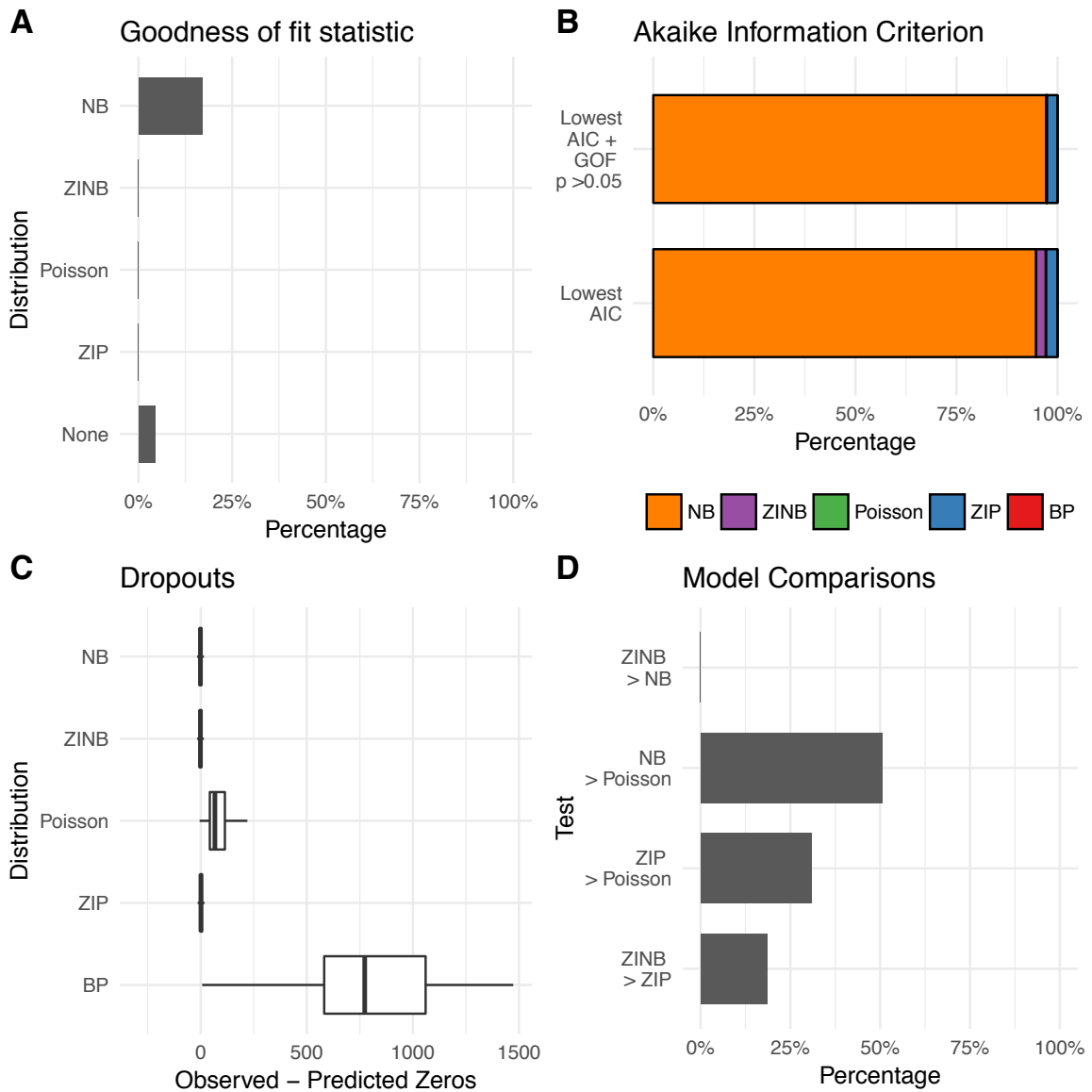


Figure S22: Soumillon et al. 2014: adipose-derived stem cells 3 days post-differentiation (SCR-seq). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

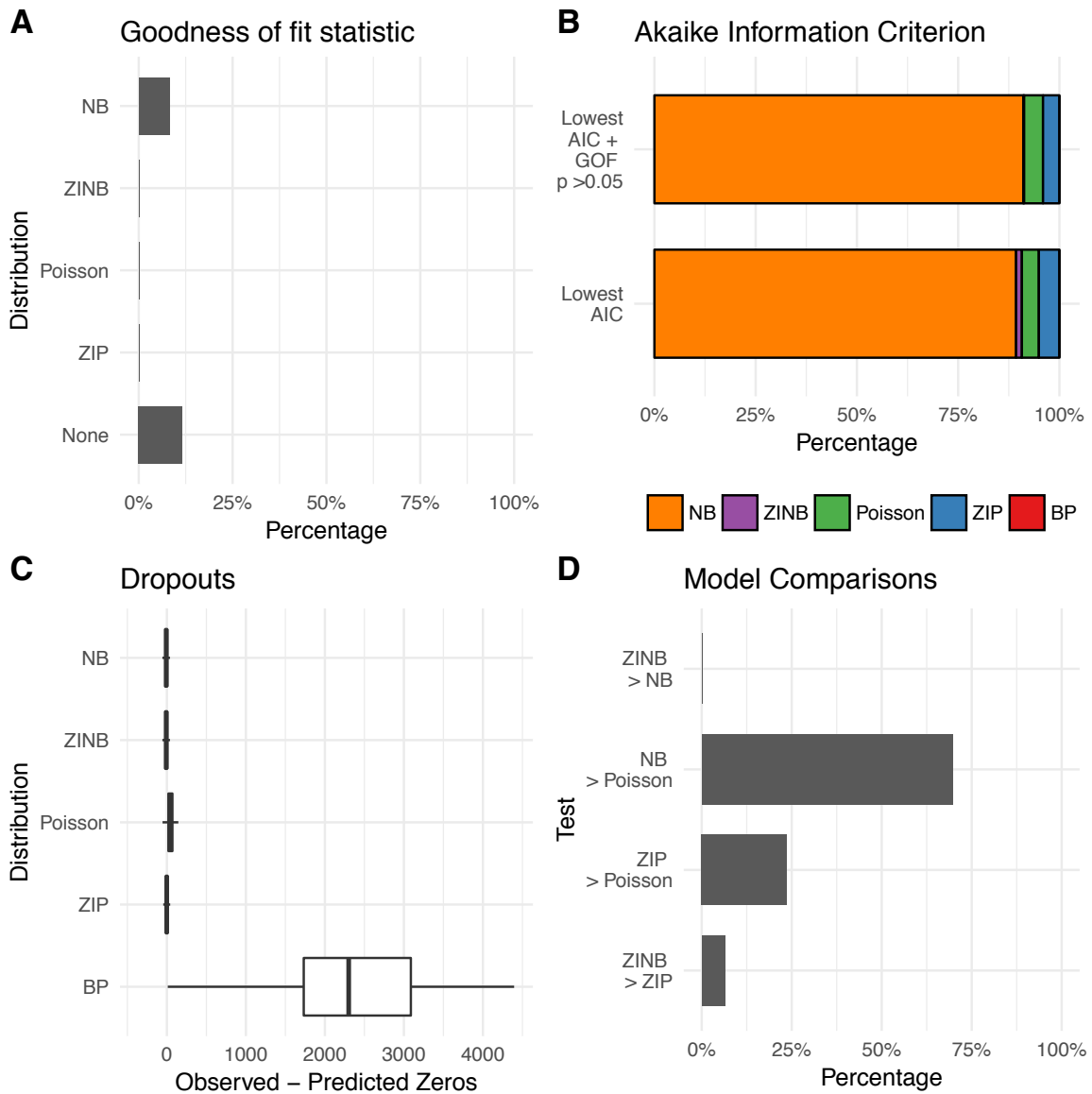


Figure S23: Zheng et al. 2017: Peripheral Blood Mononuclear Cells CD19+ B Cells (10XGenomics). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

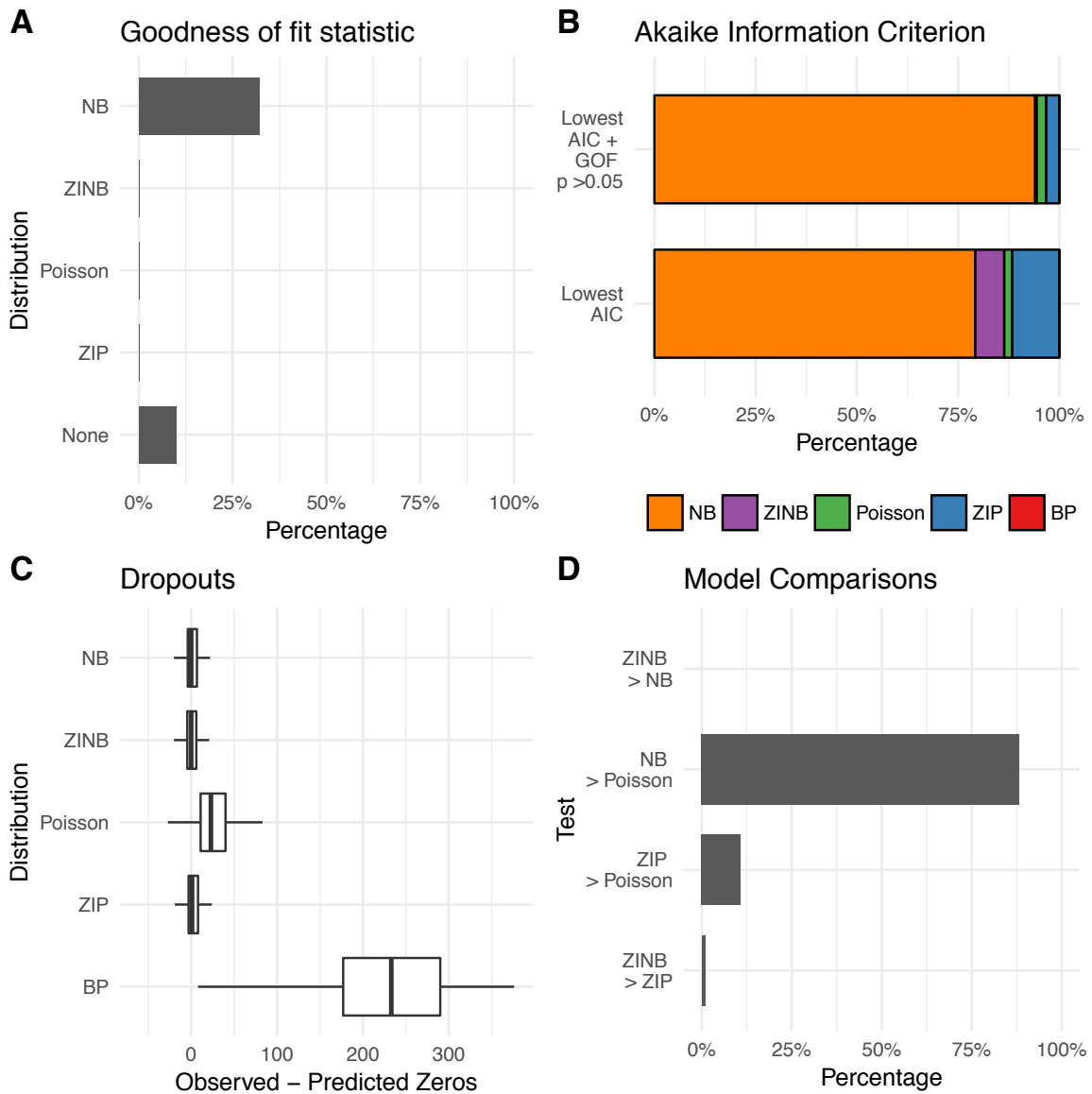


Figure S24: Zheng et al. 2017: Peripheral Blood Mononuclear Cells CD14+ Monocytes (10XGenomics). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

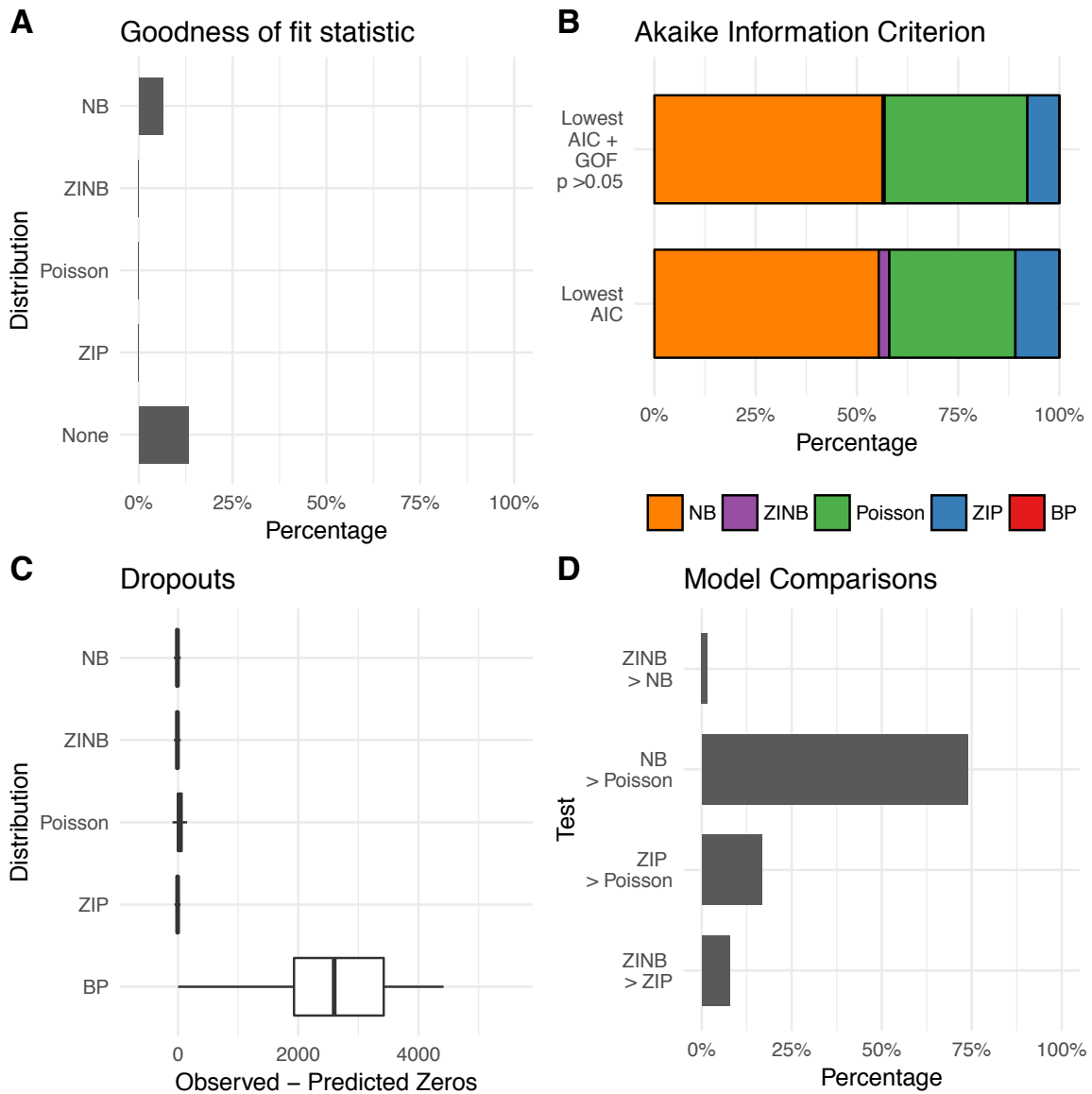


Figure S25: Zheng et al. 2017: Peripheral Blood Mononuclear Cells CD34+ Cells (10XGenomics). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

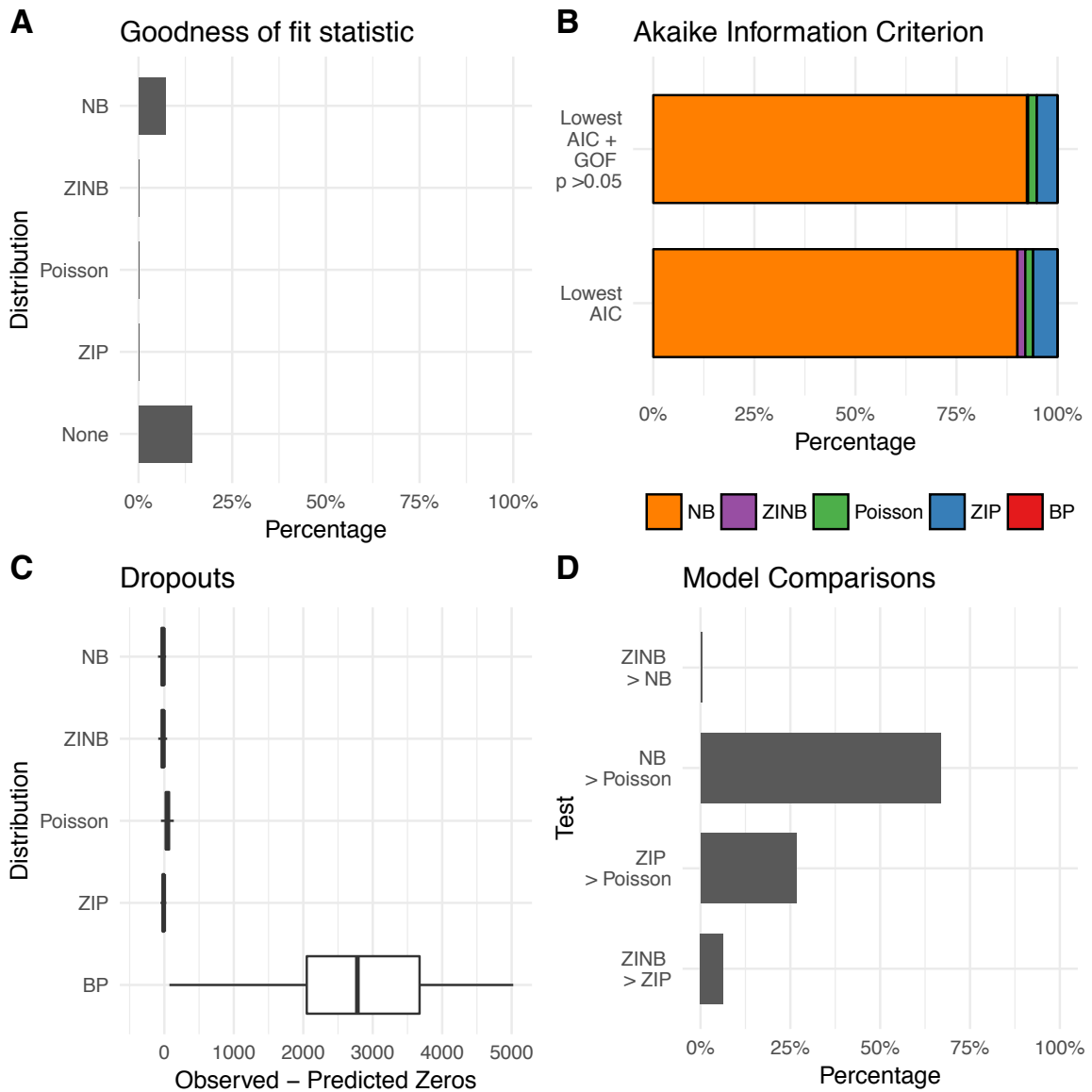


Figure S26: Zheng et al. 2017: Peripheral Blood Mononuclear Cells CD4+ T Helper Cells (10XGenomics). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

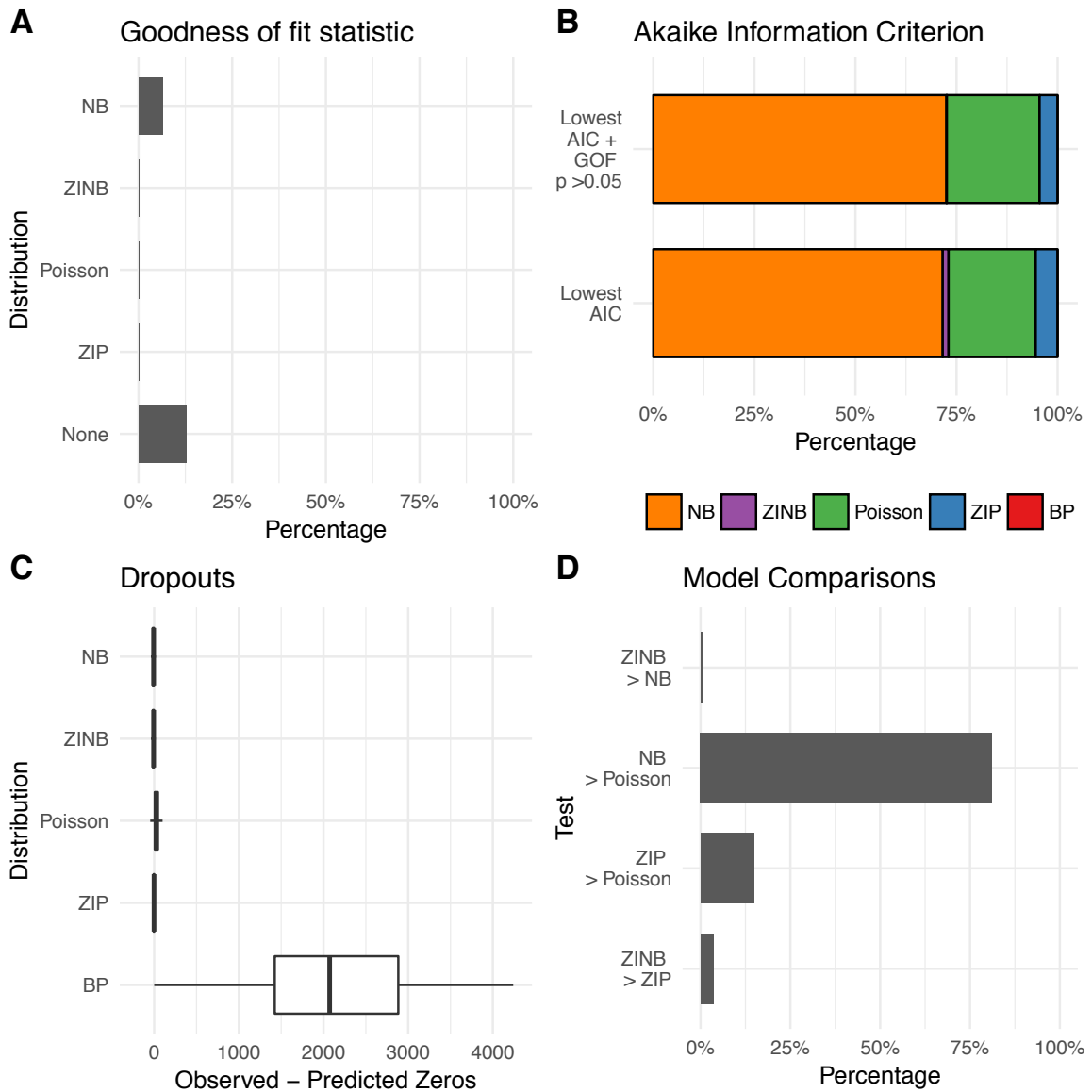


Figure S27: Zheng et al. 2017: Peripheral Blood Mononuclear Cells CD56+ NK Cells (10XGenomics). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

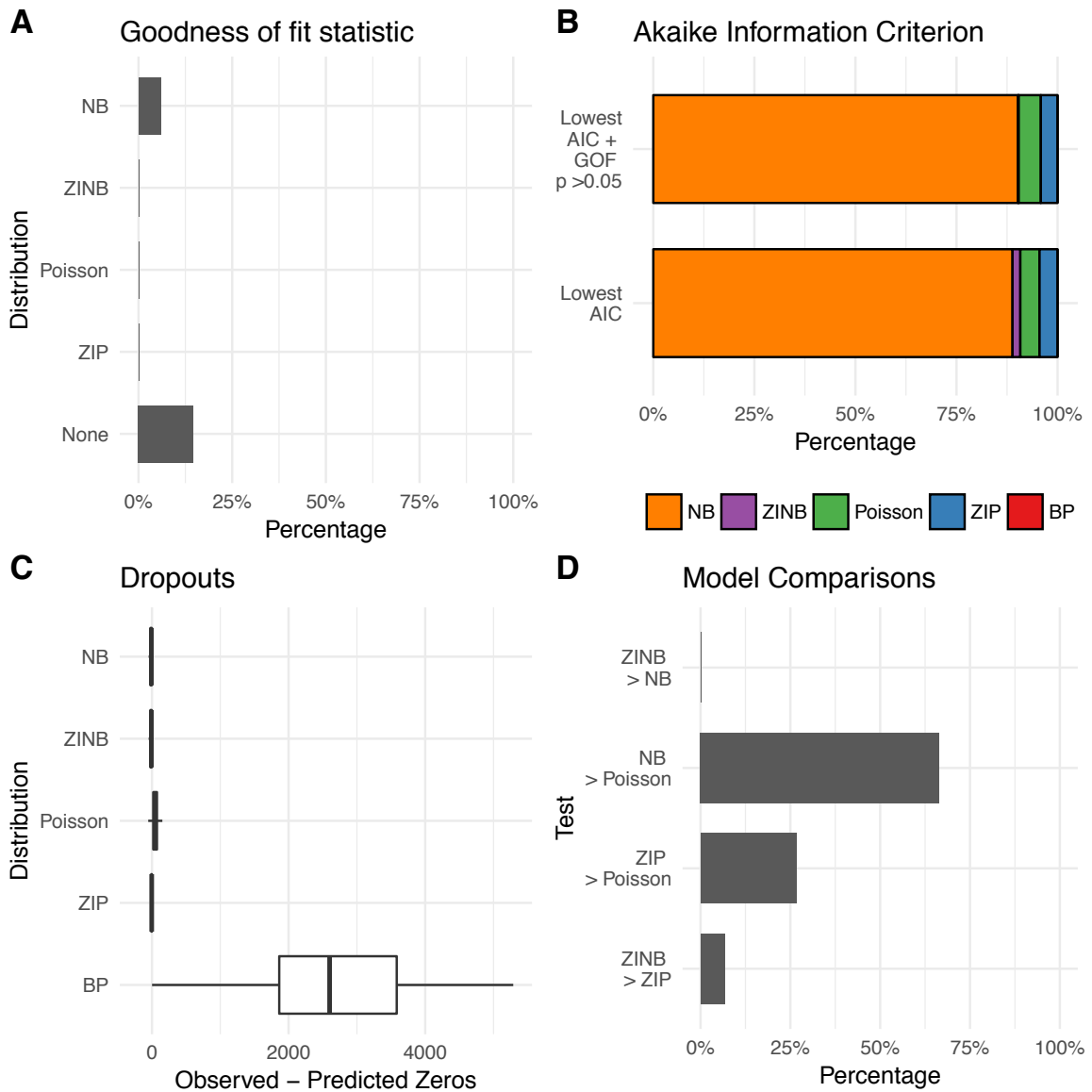


Figure S28: Zheng et al. 2017: Peripheral Blood Mononuclear Cells CD8+ Cytotoxic T Cells (10XGenomics). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

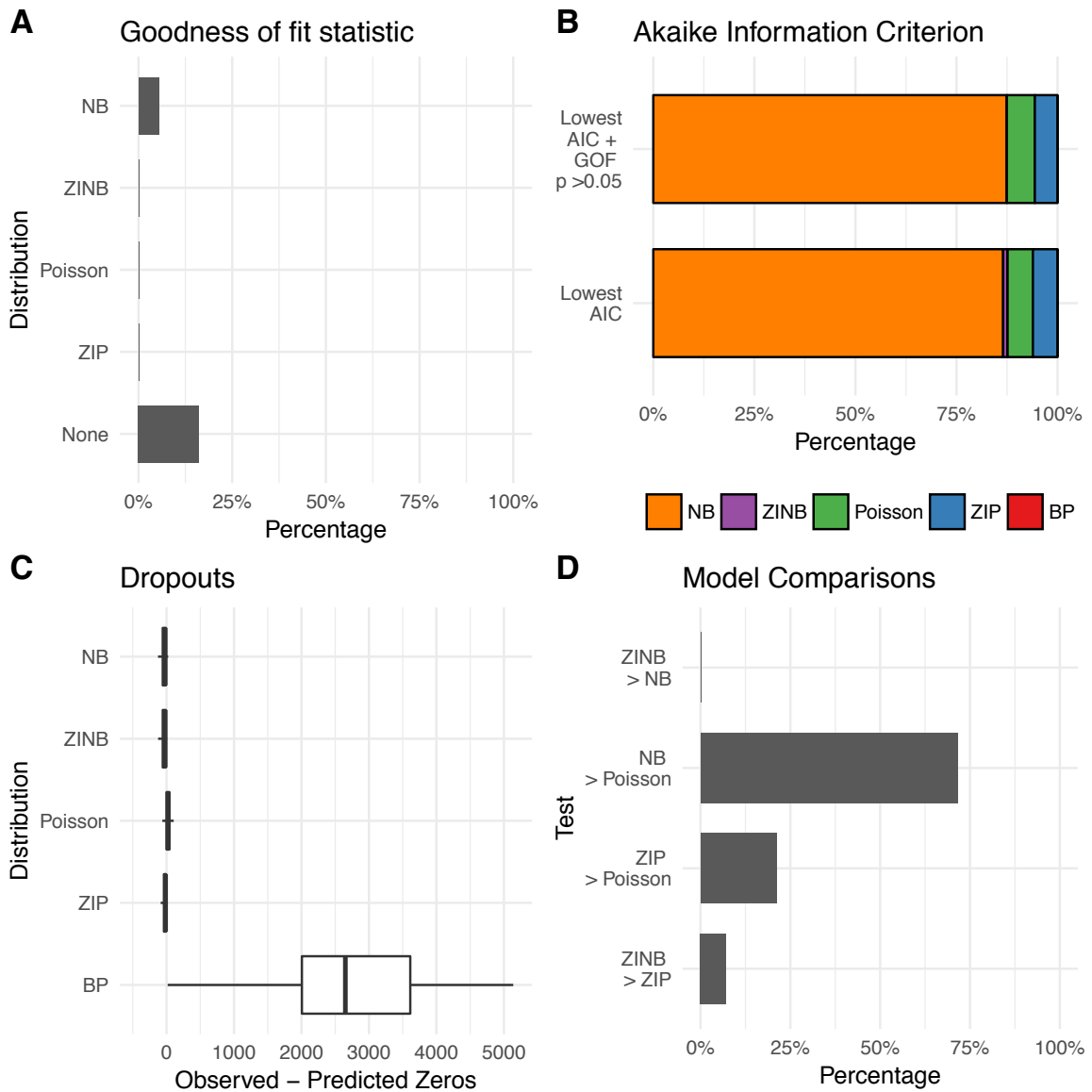


Figure S29: Zheng et al. 2017: Peripheral Blood Mononuclear Cells CD4+/CD45RO+ Memory T Cells (10XGenomics). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

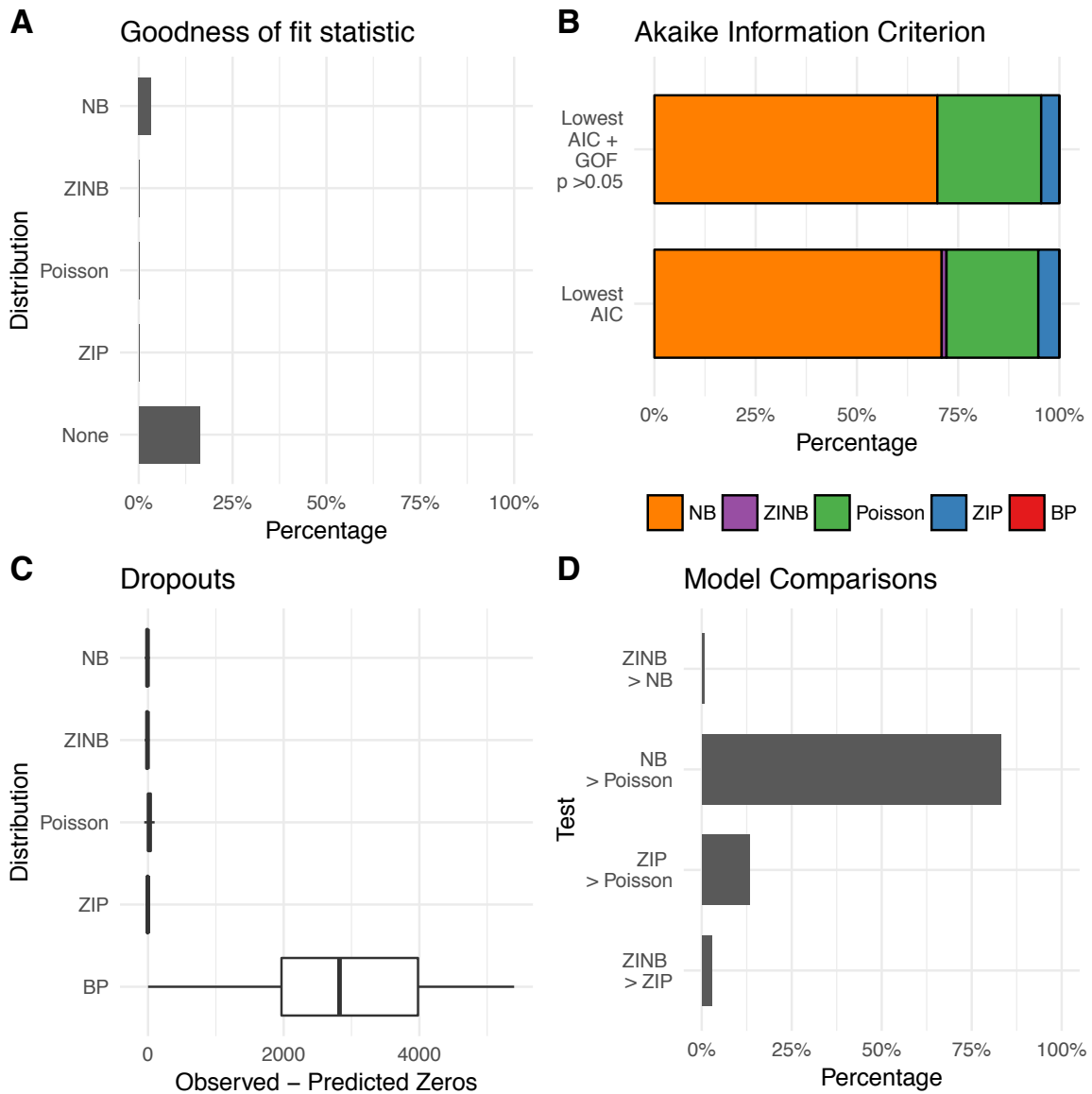


Figure S30: Zheng et al. 2017: Peripheral Blood Mononuclear Cells CD8+/CD45RA+ Naive Cytotoxic T Cells (10XGenomics). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

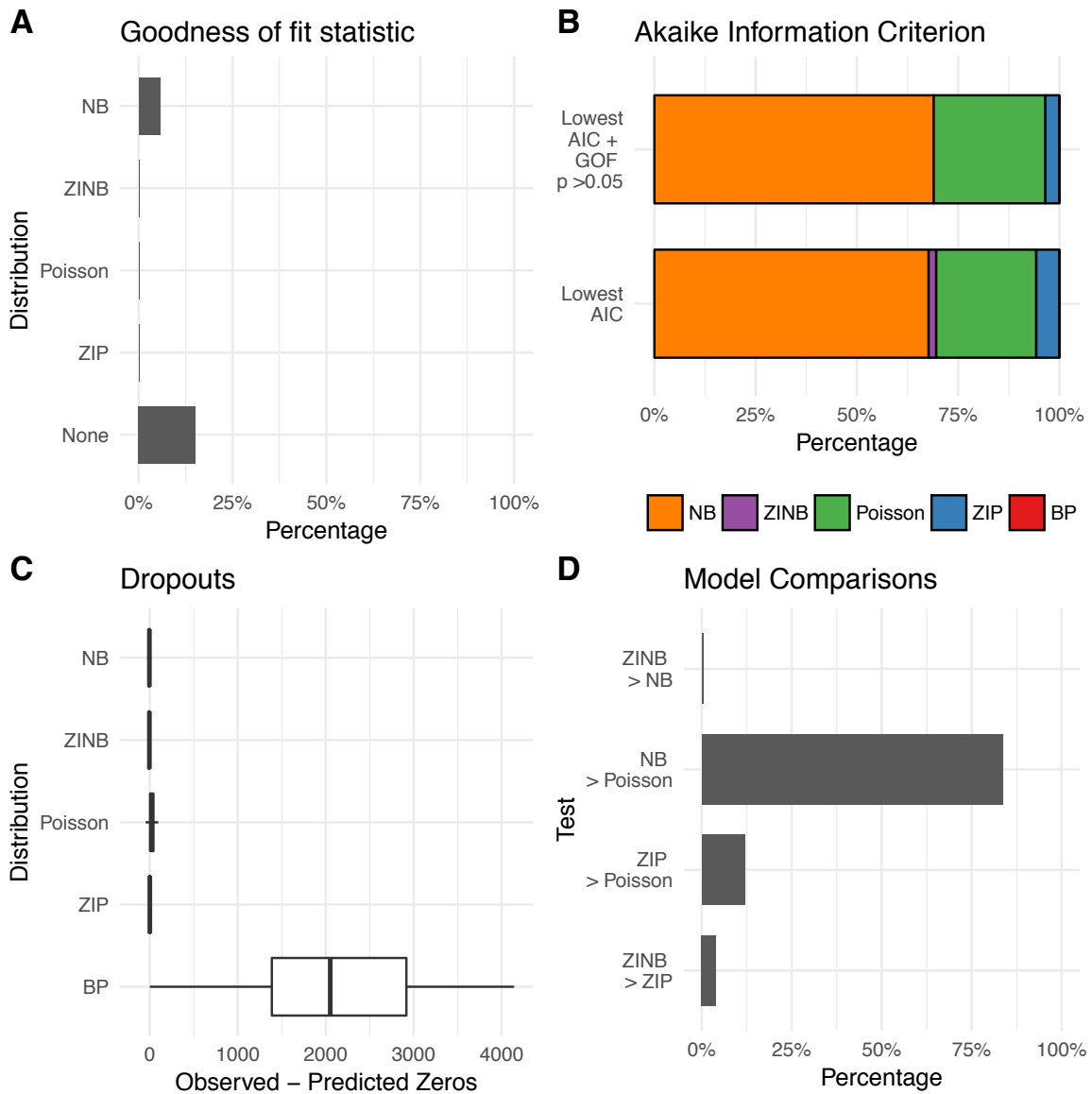


Figure S31: Zheng et al. 2017: Peripheral Blood Mononuclear Cells CD4+/CD45RA+/CD25- Naive T Cells (10XGenomics). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

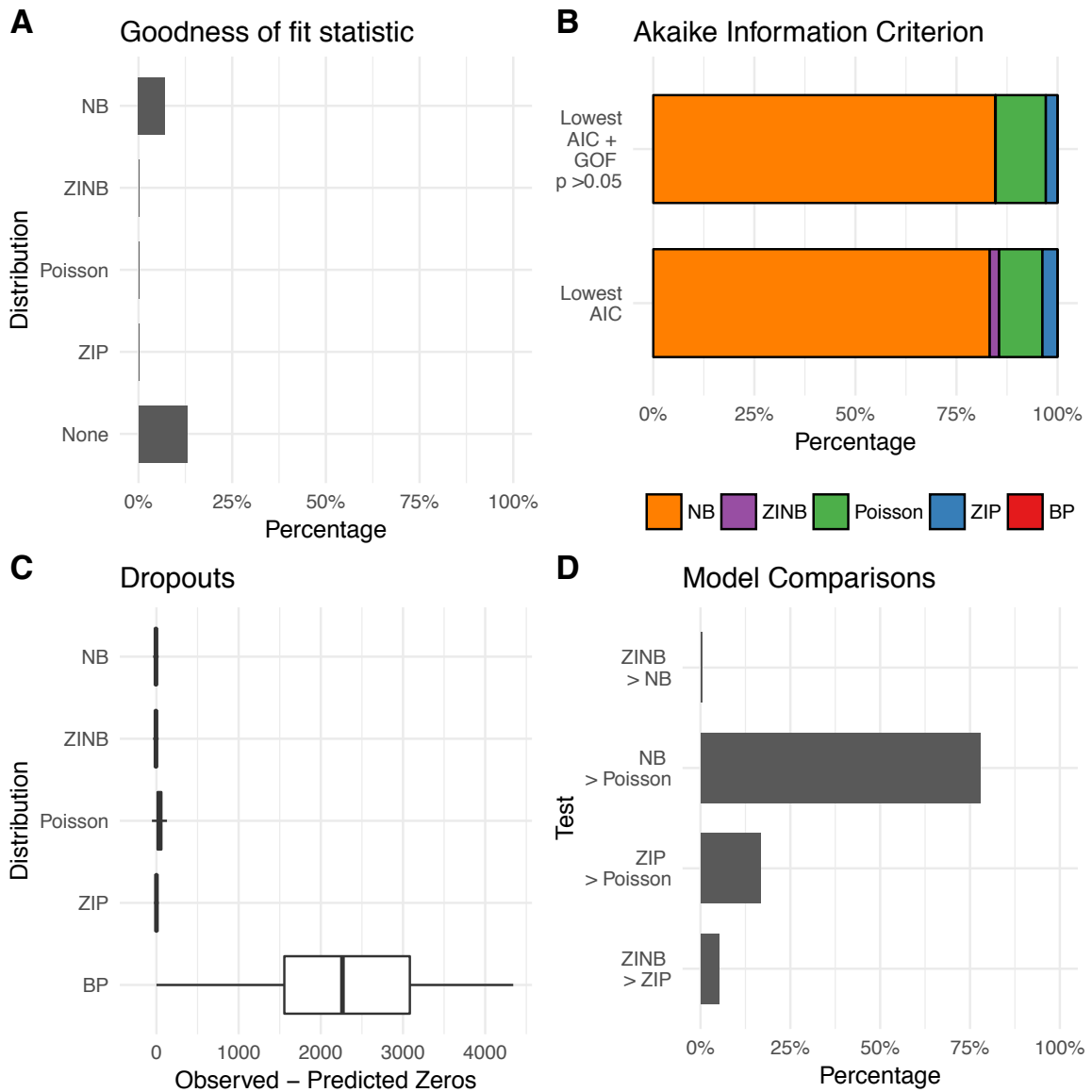


Figure S32: Zheng et al. 2017: Peripheral Blood Mononuclear Cells CD4+/CD25+ Regulatory T Cells (10XGenomics). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

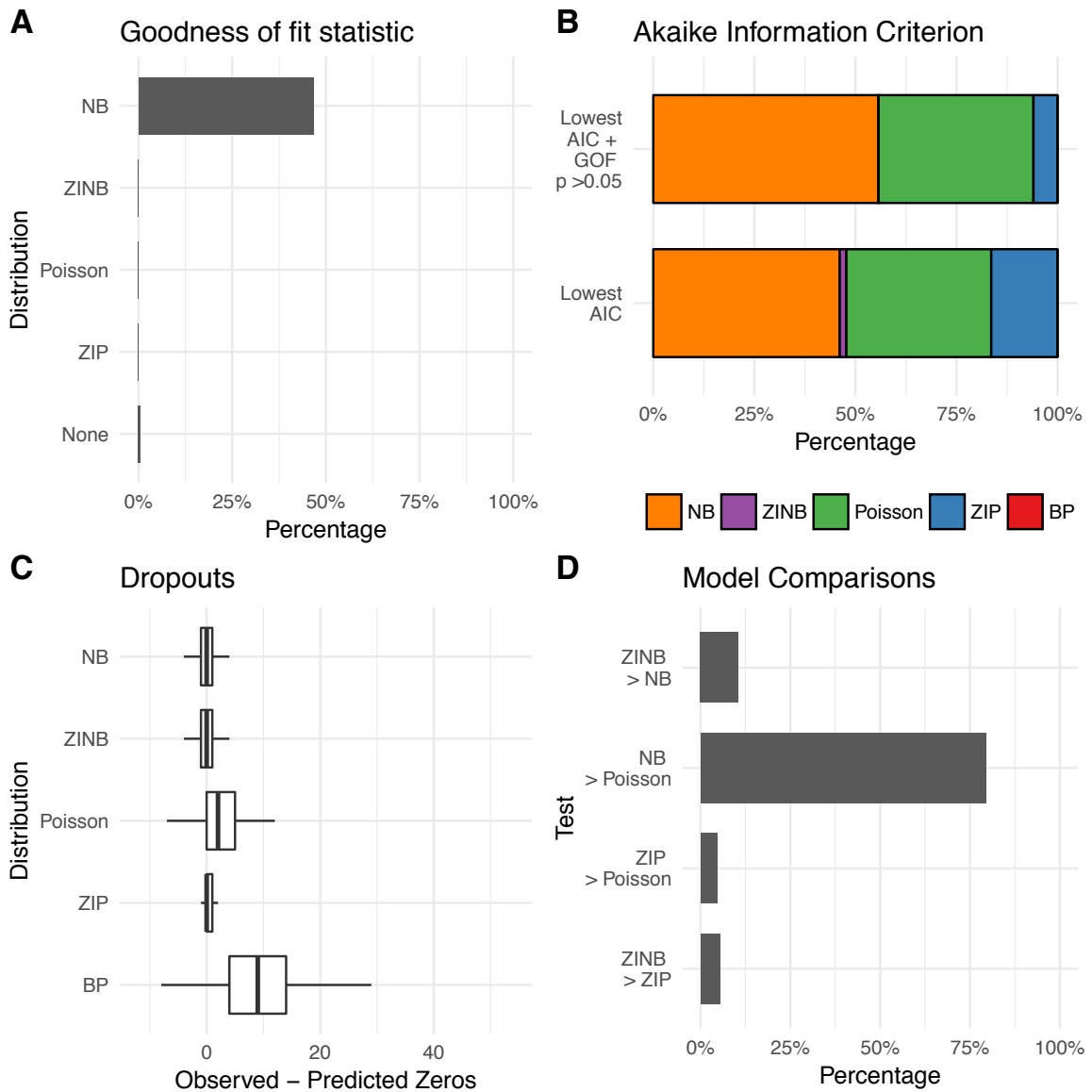


Figure S33: Ziegenhain et al. 2017: Embryonic stem cells (CEL-seq2). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

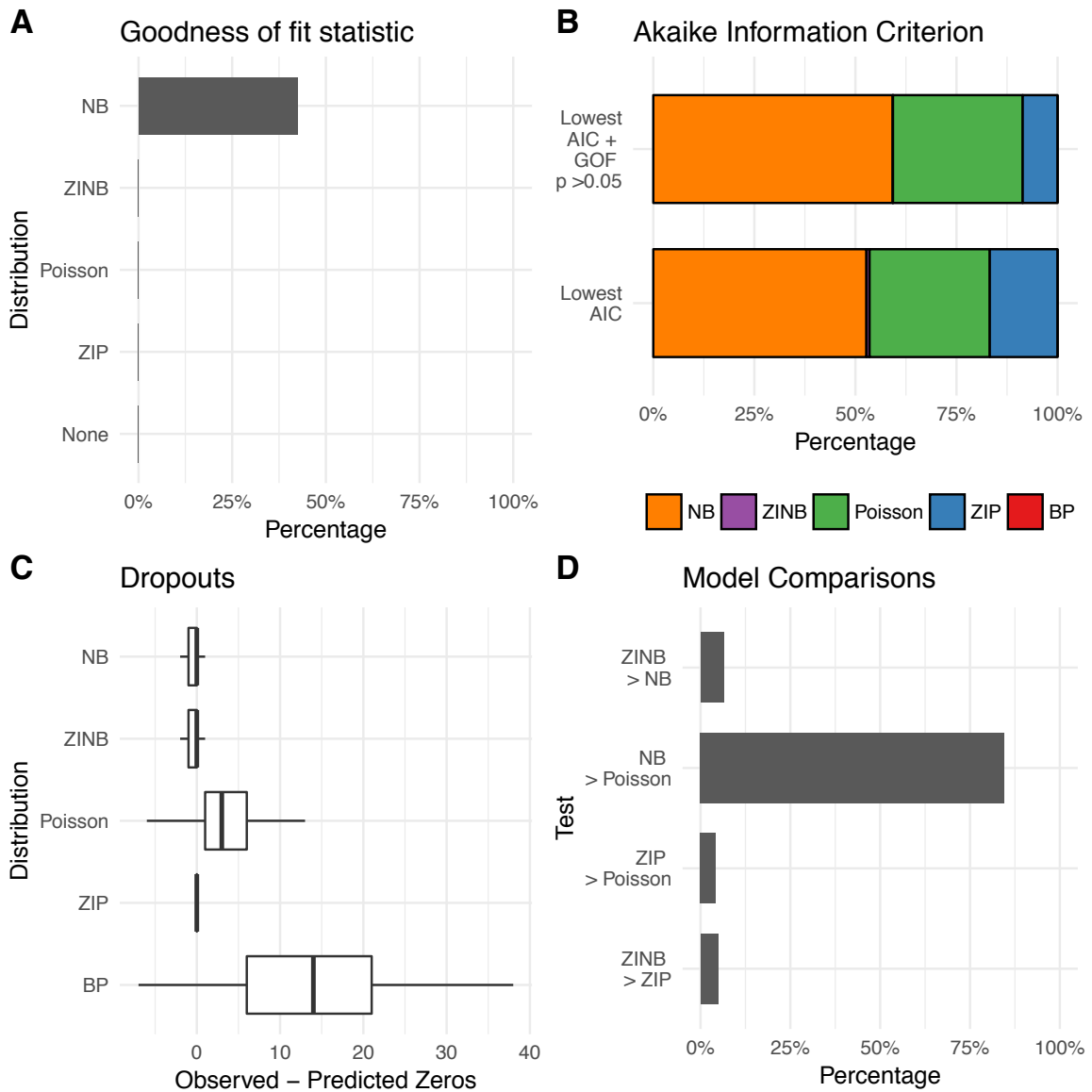


Figure S34: Ziegenhain et al. 2017: Embryonic stem cells (Drop-seq). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

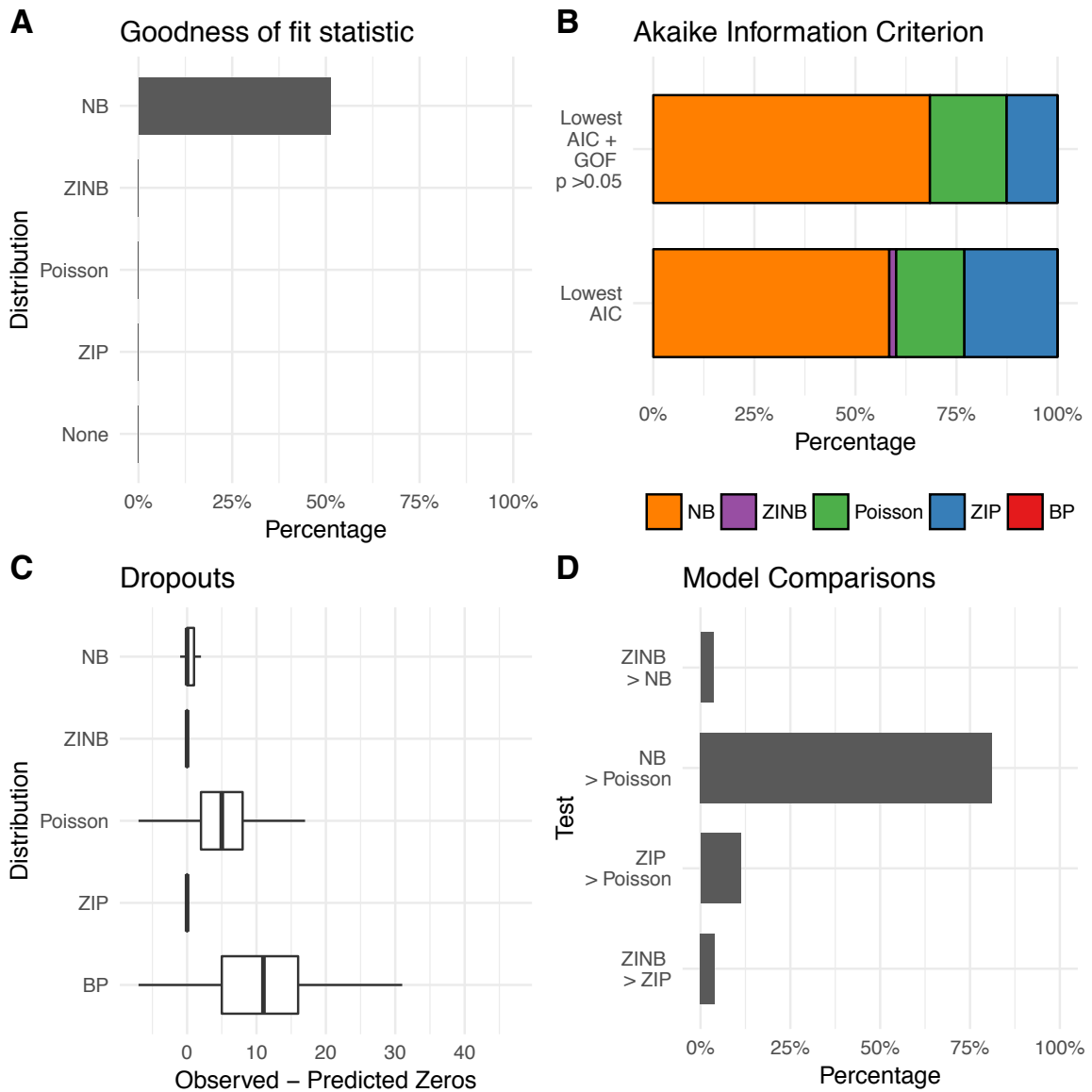


Figure S35: Ziegenhain et al. 2017: Embryonic stem cells (MARS-seq). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

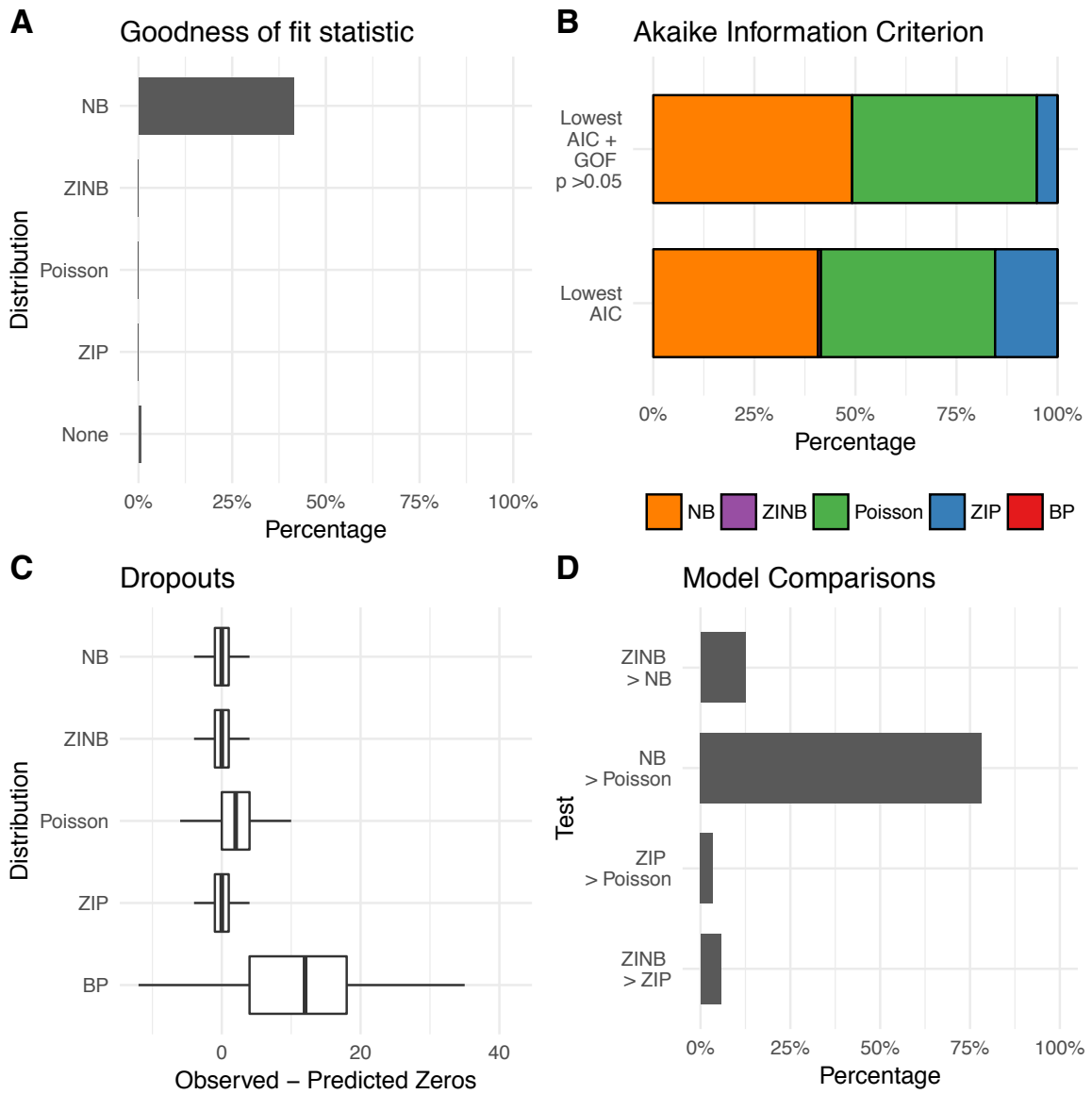


Figure S36: Ziegenhain et al. 2017: Embryonic stem cells (SCRB-seq). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

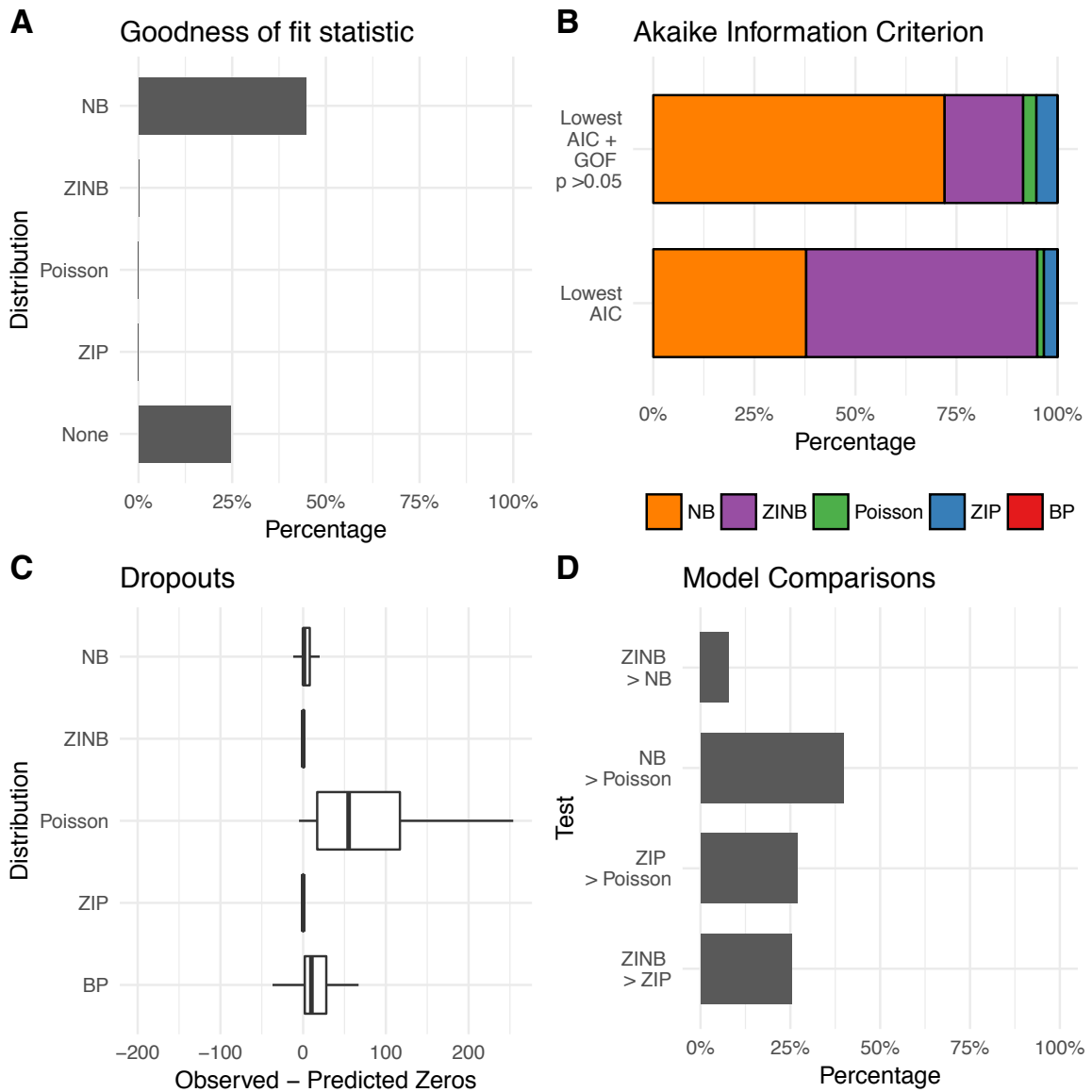


Figure S37: Ziegenhain et al. 2017: Embryonic stem cells (Smart-seq/C1). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

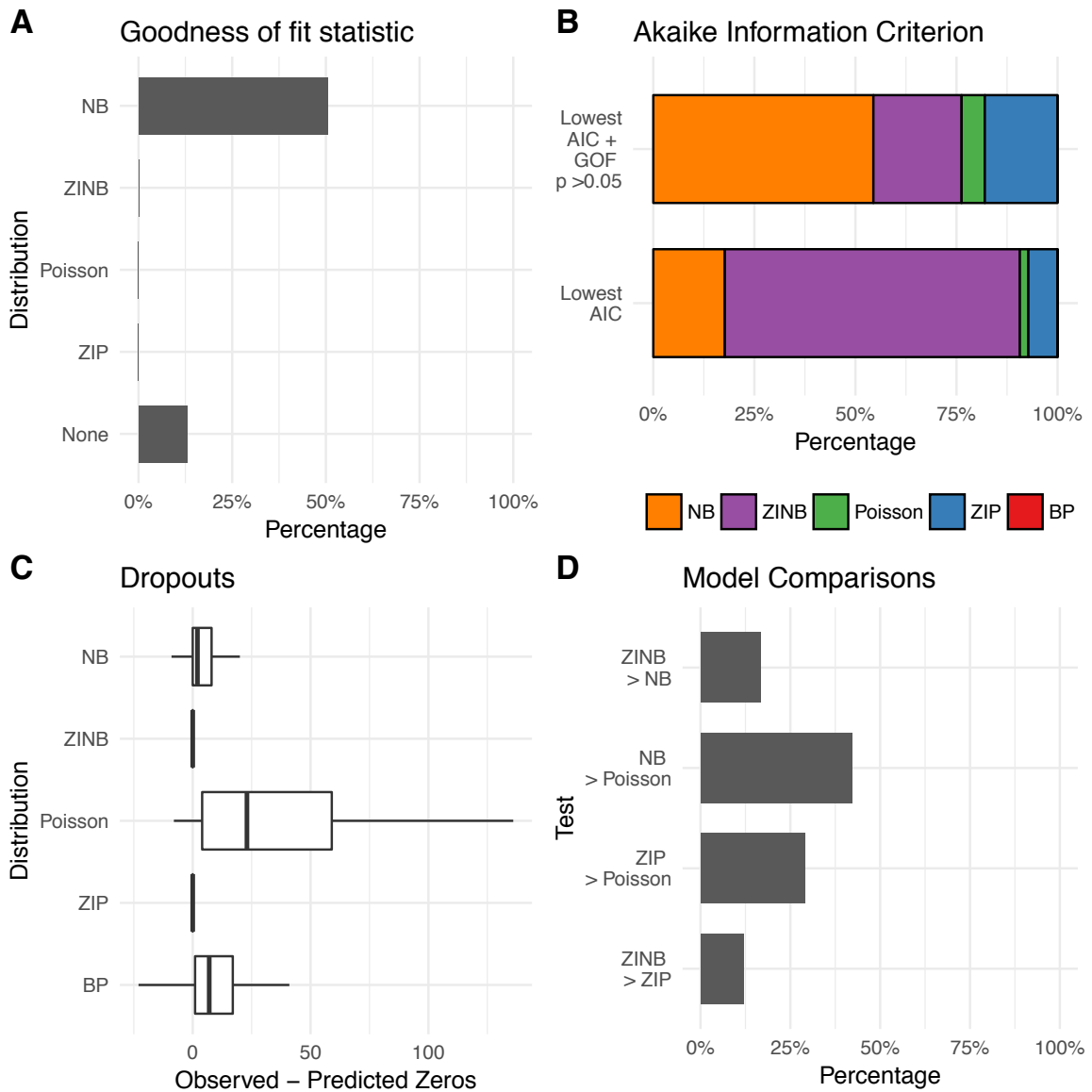


Figure S38: Ziegenhain et al. 2017: Embryonic stem cells (Smart-seq2). A) Goodness-of-fit of the model assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC. Model with the lowest AIC and passed goodness-of-fit statistic test. C) Observed versus predicted dropouts per model and gene plotted without outliers. D) Model Assessment based on Likelihood Ratio Test for nested models and Vuong Test for non-nested models. NB = Negative binomial; ZIP = Zero-inflated Poisson; ZINB = Zero-inflated negative binomial; BP = Beta-Poisson.

References

- Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, advance online publication.
- Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat. Methods*, 11(6):637–640.
- Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, 21(7):1160–1167.
- Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11(2):163–166.
- Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C. H., Ilicic, T., Henriksson, J., Natarajan, K. N., Tuck, A. C., Gao, X., Bühler, M., Liu, P., Marioni, J. C., and Teichmann, S. A. (2015). Single cell RNA-Sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17(4):471–485.
- Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., Li, N., Szpankowski, L., Fowler, B., Chen, P., Ramalingam, N., Sun, G., Thu, M., Norris, M., Lebofsky, R., Toppani, D., Kemp, I., Li, D. W., Wong, M., Clerkson, B., Jones, B. N., Wu, S., Knutsson, L., Alvarado, B., Wang, J., Weaver, L. S., May, A. P., Jones, R. C., Unger, M. A., Kriegstein, A. R., and West, J. A. A. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, 32(10):1053–1058.
- Soumillon, M., Cacchiarelli, D., Semrau, S., Oudenaarden, A. v., and Mikkelsen, T. S. (2014). Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*, page 003236.
- Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049.
- Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, 65(4):631–643.e4.

zUMIs: A fast and flexible pipeline to process RNA sequencing data with
UMIs

zUMIs – A fast and flexible pipeline to process RNA sequencing data with UMIs

Swati Parekh^{1,*}, Christoph Ziegenhain^{1,†}, Beate Vieth¹, Wolfgang Enard¹ and Ines Hellmann^{1,*}

¹Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians University, 82152 Martinsried, Germany

*parekh@bio.lmu.de; hellmann@bio.lmu.de

[†]Contributed equally.

Abstract

Single cell RNA-seq (scRNA-seq) experiments typically analyze hundreds or thousands of cells after amplification of the cDNA. The high throughput is made possible by the early introduction of sample-specific barcodes (BCs) and the amplification bias is alleviated by unique molecular identifiers (UMIs). Thus the ideal analysis pipeline for scRNA-seq data needs to efficiently tabulate reads according to both BC and UMI. *zUMIs* is such a pipeline, it can handle both known and random BCs and also efficiently collapses UMIs, either just for exon mapping reads or for both exon and intron mapping reads. Another unique feature of *zUMIs* is the adaptive downsampling function, that facilitates dealing with hugely varying library sizes, but also allows to evaluate whether the library has been sequenced to saturation. *zUMIs* flexibility allows to accommodate data generated with any of the major scRNA-seq protocols that use BCs and UMIs. To illustrate the utility of *zUMIs*, we analysed a single-nucleus RNA-seq dataset and show that more than 35% of all reads map to introns. We furthermore show that these intronic reads are informative about expression levels, significantly increasing the number of detected genes and improving the cluster resolution.

Availability: <https://github.com/sdparekh/zUMIs>

Key words: single-cell RNA sequencing, Digital gene expression, Unique Molecular Identifiers, Pipeline

Introduction

The recent development of increasingly sensitive protocols allows to generate RNA-seq libraries of single cells [1]. The throughput of such single-cell RNA-sequencing (scRNA-seq) protocols is rapidly increasing, enabling the profiling of tens of thousands of cells [2, 3] and opening exciting possibilities to analyse cellular identities [4, 5]. As the required amplification from such low starting amounts introduces substantial amounts of noise [6], many scRNA-seq protocols incorporate unique molecular identifiers (UMIs) to label individual cDNA molecules with a random nucleotide sequence before amplification [7]. This enables the computational removal of amplification noise and thus increases the power to detect expres-

sion differences between cells [8, 9]. To increase the throughput, many protocols also incorporate sample-specific barcodes (BCs) to label all cDNA molecules of a single cell with a nucleotide sequence before library generation [10, 2]. This allows for early pooling, which further decreases amplification noise [6]. Additionally, for cell types such as neurons it has been proven to be more feasible to isolate RNA from single nuclei rather than whole cells [11, 12]. This decreases mRNA amounts further, so that it has been suggested to count intron-mapping reads originating from nascent RNAs as part of single cell expression profiles [11]. However, the few bioinformatic tools that process RNA-seq data with UMIs and BCs have limitations. For example the Drop-seq pipeline is not open source

Key Points

- *zUMIs* processes UMI-based RNA-seq data from raw reads to count tables in one command.
- Unique features of *zUMIs*:
 - Automatic cell barcode selection
 - Adaptive downsampling
 - Counting of intron-mapping reads for gene expression quantification
- *zUMIs* is compatible with all major UMI based RNA-seq library protocols.

[10]. While Cell Ranger is open, it is exceedingly difficult to adapt the code to new or unknown sample barcodes and other library types. Other tools are specifically designed to work with one mapping algorithm and focus mainly on transcriptomes [13, 14]. Furthermore, to our knowledge, no UMI-RNA-seq pipeline provides the utility to also consider intron mapping reads [2, 15, 14, 13, 16]. Here, we present *zUMIs*, a fast and flexible pipeline that overcomes these limitations.

Findings

zUMIs is a pipeline that processes paired fastq files containing the UMI and BC reads and the cDNA sequence. Read pairs are filtered to remove reads with bad BCs or UMIs based on sequence quality and the remaining reads are then mapped to the genome (Figure 1). To allow the quantification of intronic reads that are generated from unspliced mRNAs, especially when using nuclei as input material, *zUMIs* generates separate UMI and read count tables for exons, introns and exon+introns. Another unique feature of *zUMIs* is that it allows for downsampling of reads before collapsing UMIs, uniquely enabling the user to assess whether a library was sequenced to saturation or whether deeper sequencing is necessary to depict the full mRNA complexity. Furthermore, *zUMIs* is flexible with respect to the length and sequences of the BC and UMIs, supporting protocols that have both sequences in one read [17, 18, 10, 14, 3, 2, 12] or split across several reads, as is the case in the InDrops v3 [19, 20] and STRT-2i [21] methods. Thus, *zUMIs* is compatible with all major UMI-based scRNA-seq protocols. Finally, *zUMIs* can be easily installed as an application on any unix machine or be conveniently deployed for cloud computing at Amazon's elastic compute service with a provided machine image.

Implementation and Operation

Pre-processing, Mapping and Counting

The input for *zUMIs* is a group of paired fastq files, where one file contains the cDNA sequence and the other file(s) the read(s) containing the BC and UMI. The exact location and length of UMI and BC are specified by the user, thus *zUMIs* can process sequences obtained from any scRNA-seq with UMIs. The first step in our pipeline is to filter reads that have low quality BCs according to a user-defined threshold, this should eliminate the bulk of spurious BCs. A similar sequence quality based cut-off can be applied to the UMI. Others have suggested to use edit distances and frequencies of the UMIs to collapse spurious counts due to errors [16]. However, in the data that we analyzed, quality filtering of UMIs had no significant impact on the power to detect differentially expressed genes (Figure 2), implying that the computationally expensive distance filter will be mostly unnecessary.

The remaining reads are then mapped to the genome using the splice-aware aligner STAR [22]. The user is free to customize mapping by using the options of STAR. Furthermore, if

the user wishes to use a different mapper, it is also possible to provide *zUMIs* with an aligned bam-file instead of the fastq-file with the cDNA sequence, with the sole requirement that only one mapping position per read is reported in the bam-file. Next, reads are assigned to genes and to exons or introns based on the provided gtf file, while ensuring introns are not overlapping with any exon. `Rsubread featureCounts` [23] is used to first assign reads to exons and afterwards to check whether the remaining reads fall into introns. The output is then read into R using `data.table` [24] count tables for UMIs and reads per gene per BC are generated. Only identical UMI sequences that were mapped either to the exon or intron of the same gene are collapsed. Note that only the processing of intron and exon reads together allows to properly collapse UMIs that can be sampled from the intronic as well as from the exonic part of the same nascent mRNA molecule.

Cell Barcode Selection

In order to be compatible with well-based and droplet-based scRNA-seq methods, *zUMIs* needs to be able to deal with known as well as random BCs. As default behavior, *zUMIs* infers which barcodes mark good cells from the data (Figure 3 A,B). To this end, we fit a k-dimensional multivariate normal distribution [25, 26] for the number of reads/BC, and reason that only the kth normal distribution with the largest mean contains barcodes that identify reads originating from intact cells. We exclude all barcodes that fall in the lower 1% tail of this distribution. The HEK dataset used in this paper contains 96 cells with known barcodes and *zUMIs* identifies 99 barcodes as intact, including all the 96 known barcodes. Also for the single-nucleus RNA-seq from Habib et al. [12] *zUMIs* identified a reasonable number of cells: Habib et al. report 10,877 nuclei and *zUMIs* identified 11,013 intact nuclei. However, if the number of barcodes or barcode sequences are known, it is preferable to use this information. In the case that *zUMIs* is either given the number of BCs or is provided with a list of BC sequences, it will use this information and forgo automatic inference.

Downsampling

scRNA-seq library sizes can vary by orders of magnitude, which complicates normalization [27, 28]. A straight-forward solution for this issue is to downsample over-represented libraries [29]. *zUMIs* has an inbuilt function for downsampling datasets to a user-specified number of reads or a range of reads. By default, *zUMIs* downsamples all selected barcodes to be within three absolute deviations from the median number of reads per barcode (Figure 3 C). Alternatively, the user can provide a target sequencing depth and *zUMIs* will downsample to the specified read number or omit the sample from the downsampled count table. Furthermore, *zUMIs* also allows to specify multiple target read number at once for downsampling. This feature is helpful, if the user wishes to determine whether the RNA-seq library was sequenced to saturation or whether further sequencing would increase the number of detected genes or UMIs enough to justify the extra cost. In our HEK-cell exam-

ple dataset the number of detected genes starts leveling off at one million reads, sequencing double that amount would only increase the number of detected genes from 9,000 to 10,600, when counting exon reads (Figure 3D). The saturation curve of exon+intron reads runs parallel to the one for exon reads, both indicating that a sequencing depth of one million reads per cell is sufficient for these libraries.

Output and Statistics

zUMIs outputs three UMI and three read count tables: gene-wise counts for traditional exon mapping, one for intron and one for exon+intron counts. If a user chooses the downsampling option, 6 additional count-tables per target read count are provided. To evaluate library quality *zUMIs* summarizes the mapping statistics of the reads. While exon and intron mapping reads likely represent mRNA quantities, a high fraction of intergenic and unmapped reads indicates low-quality libraries. Another measure of RNA-seq library quality is the complexity of the library, for which the number of detected genes and the number of identified UMIs are good measures (Figure 1). We processed 227 million reads with *zUMIs* and quantified expression levels for exon and intron counts on a unix machine using up to 16 threads, which took barely 3 hours. Increasing the number of reads increases the processing time approximately linearly, where filtering, mapping and counting each take up roughly one third of the total time (Figure 3 E). We also observe that the peak RAM usage for processing datasets of 227, 500 and 1000 million pairs was 42 Gb, 89 Gb and 172 Gb, respectively. Finally, *zUMIs* could process the largest scRNA-seq dataset reported to date with around 1.3 million brain cells and 25 billion read pairs generated with 10xGenomics Chromium https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons on a 22-core Intel Xeon E5-2699 processor in only 7 days.

Intron Counting

Assuming that intron mapping reads originate from nascent mRNAs, *zUMIs* also counts and collapses intron mapping reads with other reads mapping to the same gene with the same UMI. To assess the information gain from intronic reads to estimate gene expression levels, we analysed a publicly available DroNc-seq mouse brain dataset ([12], https://portals.broadinstitute.org/single_cell). For the ~ 11,000 single nuclei of this dataset, the fraction of intron mapping reads of all reads goes up to 61%. Thus, if intronic reads are considered, the mean number of detected genes per cell increases significantly from 1041 for exon reads to 1995 for exon+intron reads (Welch two sample t-test: p-value < 2.2e-16). To assess the impact of intronic reads on the inference of differential expression, we performed power simulations using empirical mean and dispersion distributions from this dataset [9]. The simulations assumed a balanced two-group comparison of variable sample sizes with 10% of the genes differentially expressed between groups. We observed a 0.5% decrease of the marginal false discovery rate (FDR) for exon+intron relative to exon counts for group sample sizes of < 250 cells, while the power to detect differentially expressed genes was similar for exon and exon+intron counts. Next, we investigated whether exon+intron counting improves the identification of cell types, as suggested in [11]. Following the Seurat pipeline [30], we clustered the cells of the DroNc-seq dataset based on the exon as well as our exon+intron counts. The KNN-clustering reported 24 distinct clusters for the exon+intron counts, while we could only discriminate 15 clusters using exon counts (Figure 4). This analysis shows, that the additional genes that were detected by also counting intron-mapping reads are not spurious, but carry biological meaning.

Conclusion

zUMIs is a fast and flexible pipeline processing raw reads to obtain count tables for RNA-seq data using UMIs. To our knowledge it is the only open source pipeline that has a barcode and UMI quality filter, allows intron counting and has an integrated downsampling functionality. These features ensure that *zUMIs* is applicable to most experimental designs of RNA-seq data, including single nucleus sequencing techniques, droplet-based methods where the BC is unknown, as well as plate-based UMI-methods with known BCs. Finally, *zUMIs* is computationally efficient, user-friendly and easy to install.

Availability of Source Code and Requirements

- Project name: *zUMIs*
- Project home page: <https://github.com/sdparekh/zUMIs>
- Operating system(s): UNIX
- Programming language: shell, R, perl
- Other requirements: STAR >= 2.5.3a, R >= 3.4, pigz >= 2.3 & samtools >= 1.1
- License: GNU GPLv3.0

Availability of supporting data and materials

All data that were generated for this project were submitted to GEO under accession GSE99822.

Declarations

List of Abbreviations

scRNA-seq - single-cell RNA-sequencing
UMI - Unique Molecular Identifier
BC - Barcode
MAD - Median Absolute Deviation

Competing Interests

The author(s) declare that they have no competing interests.

Funding

This work has been supported by the DFG through SFB1243 sub-projects A14/A15.

Author's Contributions

SP and CZ designed and implemented the pipeline. BV tested the pipeline and helped in power simulations. SP, CZ, WE and IH wrote the manuscript. All authors read and approved the final manuscript.

References

1. Sandberg R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat Methods* 2014; Jan;11(1):22-24.
2. 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; 16 Jan;8:14049.

3. Rosenberg AB, Roco C, Muscat RA, Kuchina A, Mukherjee S, Chen W, et al. Scaling single cell transcriptomics through split pool barcoding. *bioRxiv* 2017 2 Feb;p. 105163.
4. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* 2016 8 Nov;34(11):1145–1160.
5. Regev A, Teichmann S, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. *bioRxiv* 2017 8 May;p. 121202.
6. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. *Sci Rep* 2016 9 May;6:25533.
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* 2012 Jan;9(1):72–74.
8. 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 16 Feb;65(4):631–643.e4.
9. Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: Power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* 2017 Jul;.
10. Macosko EZ, Basu A, Satija R, Nemes J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 2015 21 May;161(5):1202–1214.
11. Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* 2016 24 Jun;352(6293):1586–1590.
12. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* 2017 Oct;14(10):955–958.
13. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* 2017 6 Mar;.
14. Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* 2016 28 Apr;17(1):77.
15. Macosko EZ, Basu A, Satija R, Nemes J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 2015 21 May;161(5):1202–1214.
16. Smith TS, Heger A, Sudbery I. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* 2017 18 Jan;.
17. Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv* 2014 5 Mar;.
18. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretzky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 2014 14 Feb;343(6172):776–779.
19. 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 21 May;161(5):1187–1201.
20. Zilionis R, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc* 2017 Jan;12(1):44–73.
21. Hochgerner H, Lännerberg P, Hodge R, Mikes J, Heskol A, Hubschle H, et al. STRT-seq-2i: dual-index 5' single cell and nucleus RNA-seq on an addressable microwell array. *bioRxiv* 2017 20 Apr;p. 126268.
22. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013 1 Jan;29(1):15–21.
23. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014 1 Apr;30(7):923–930.
24. Dowle M, Srinivasan A. data.table: Extension of 'data.frame'; 2017, <https://CRAN.R-project.org/package=data.table>, r package version 1.10.4.
25. Fraley C, Raftery AE. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *J Am Stat Assoc* 2002 Jun;97(458):611–631.
26. Fraley C, Raftery AE, Brendan Murphy T, Scrucca L. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation 2012;.
27. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* 2017 Jun;14(6):565–571.
28. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform* 2017 27 Feb;.
29. Grün D, van Oudenaarden A. Design and Analysis of Single-Cell Sequencing Experiments. *Cell* 2015 5 Nov;163(4):799–810.
30. Butler A, Satija R. Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. *bioRxiv* 2017 Jul;p. 164889.

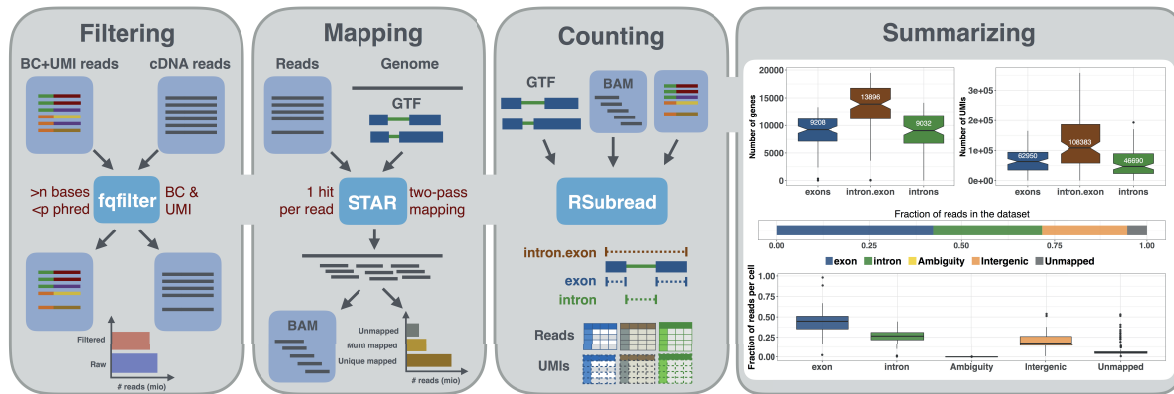


Figure 1. Schematic of the zUMIs pipeline. Each of the grey panels from left to right depicts a step of the zUMIs pipeline. First, fastq files are filtered according to user-defined barcode (BC) and unique molecular identifier (UMI) quality thresholds. Next, the remaining cDNA reads are mapped to the reference genome using STAR. Gene-wise read and UMI count tables are generated for exon, intron and exon+intron overlapping reads. To obtain comparable library sizes, reads can be downsampled to a desired range during the counting step. In addition, zUMIs also generates data and plots for several quality measures, such as the number of detected genes/UMIs per barcode and distribution of reads into mapping feature categories.

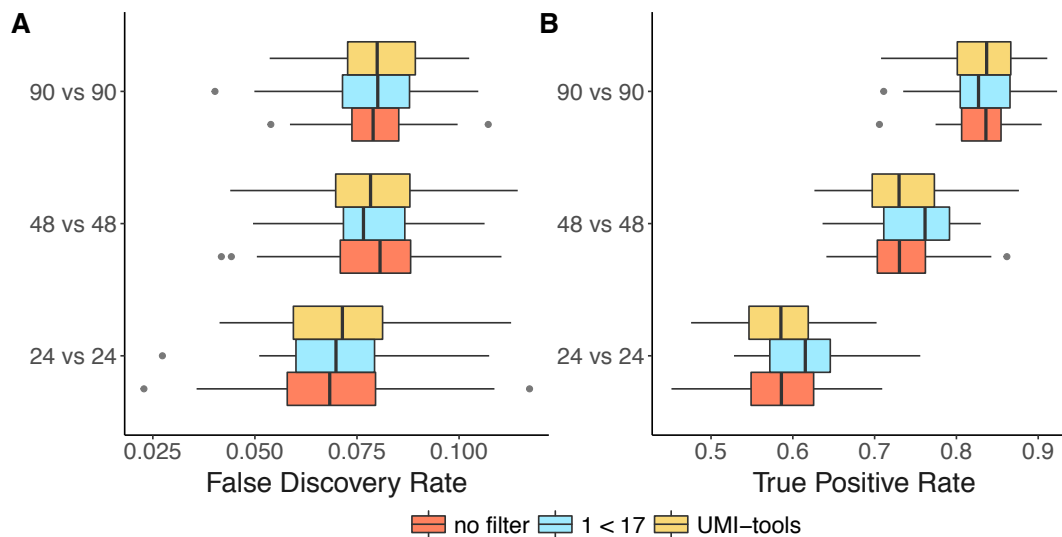


Figure 2. Impact of UMI quality filtering on Differential Gene Expression. We estimated the mean expression and dispersion of genes across the cells from our HEK dataset without any UMI quality filters (red); reads where the UMI has at least one base with a quality score < 17 (blue) and using the directional-adjacency method implemented in UMI-tools[16] (yellow), that collapses UMIs based on their distance in a sequence graph also considering the frequency. The resulting count matrices were then used for power simulations using powsimR [9] with balanced sample sizes of n in each group. We performed 50 simulations with 9000 genes where 10% of the genes are differentially expressed with \log_2 fold changes drawn from a normal distribution $N(\mu = 0, \sigma = 1.5)$. We report here A) false discovery rate (FDR) and B) true positive rate (TPR) to detect differential expression for each filtering criterion.

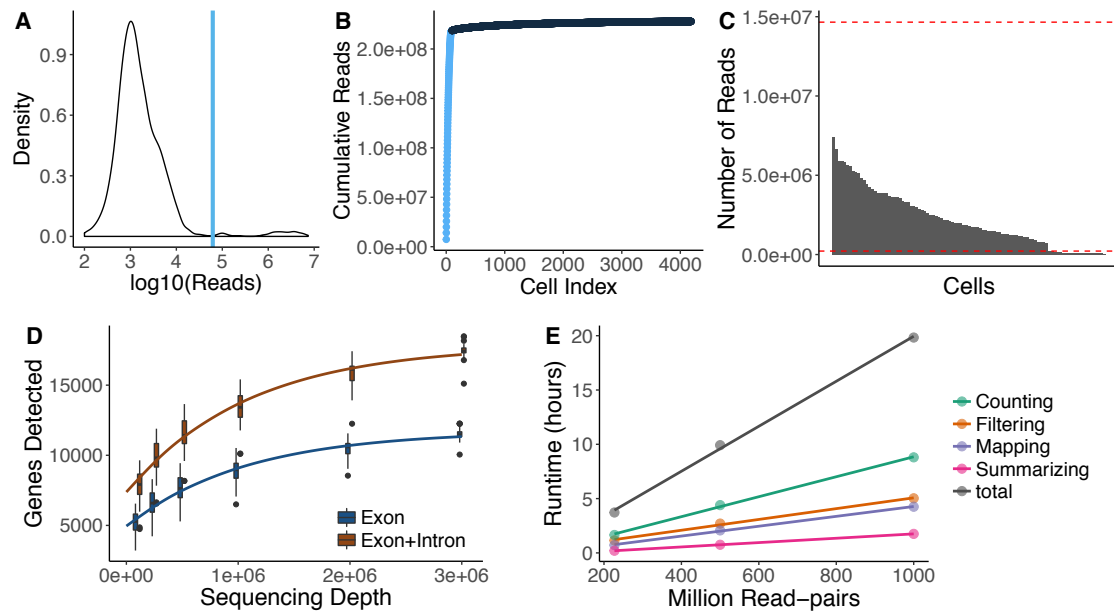


Figure 3. Utilities of zUMIs. Each of the panels shows the utilities of zUMIs pipeline. The plots from A–D are the results from the example HEK dataset used in the paper. A) The plot shows a density distribution of reads per barcode. Cell barcodes with reads above the blue line are selected. B) The plot shows the cumulative read distribution in the example HEK dataset where the barcodes in light blue are the selected cells. C) The barplot shows the number of reads per selected cell barcode with the red lines showing upper and lower MAD (Median Absolute Deviations) cutoffs for adaptive downsampling. Here, the cells below the lower MAD have very low coverage and are discarded in downsampled count tables. D) Cells were downsampled to six depths from 100,000 to 3,000,000 reads. For each sequencing depth the genes detected per cell is shown. E) Runtime for three datasets with 227, 500 and 1000 million read-pairs. The runtime is divided in the main steps of the zUMIs pipeline: Filtering, Mapping, Counting and Summarizing. Each dataset was processed using 16 threads ("–p 16").

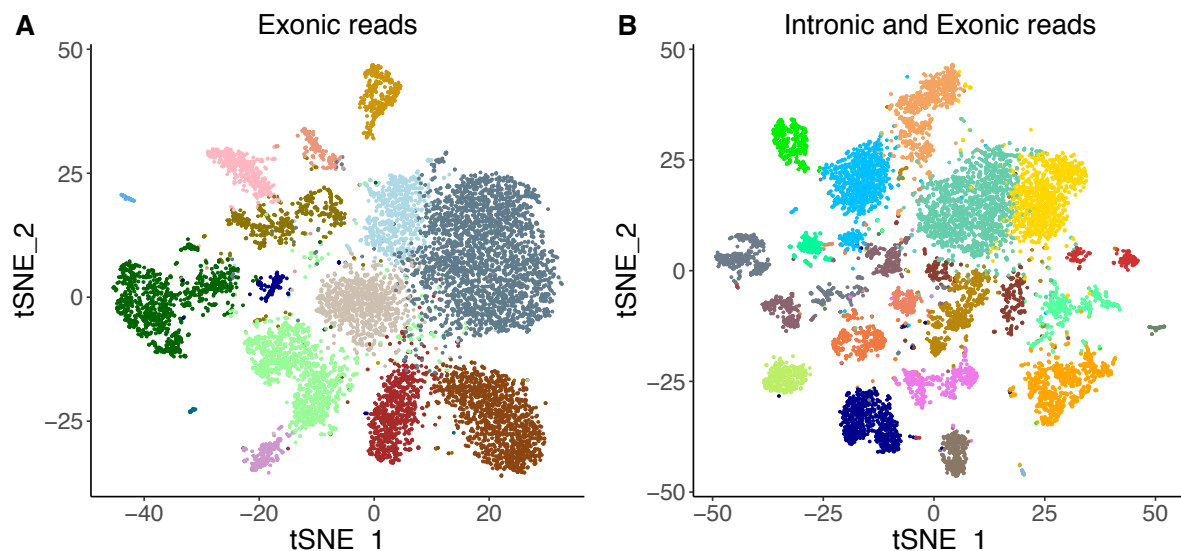


Figure 4. Contribution of intron reads in scRNA-seq. We analyse published single-nucleus RNA-seq data[12] to assess the utility of counting intron reads. We processed the raw data with zUMIs to obtain a count table with exon reads as well as exon+intron reads. We follow the Seurat pipeline[30] for filtering, normalising and clustering of cells for exon and exon+intron count tables and find 15 and 24 clusters, respectively. The t-SNE plot in panel (A) is colored by cluster identity of exon reads and panel (B) colored by cluster identity from exon+intron reads.

zUMIs: a fast and flexible pipeline to process RNA sequencing data
with UMIs

SUPPLEMENTARY INFORMATION

by

Swati Parekh^{1,2*}, Christoph Ziegenhain^{1,*}, Beate Vieth¹, Wolfgang Enard¹ and Ines
Hellmann^{1,2}

¹Anthropology & Human Genomics, Department of Biology II,
Ludwig-Maximilians University, Munich, Germany

*Contributed equally

²Corresponding author

1 Characterization of zUMIs

To demonstrate the utility of *zUMIs*, we processed data generated from 96 HEK cells using the SCRBS-seq protocol [2, 3].

227 million read-pairs of sequencing data were processed on a linux workstation running at light load using up to 16 threads. The processing was complete after 173 minutes (Figure S1). We observe that runtime for *zUMIs* scales linearly, as does RAM usage. The peak RAM usage for processing datasets of 227, 500 and 1000 million pairs was 42 Gb, 89 Gb and 172 Gb, respectively.

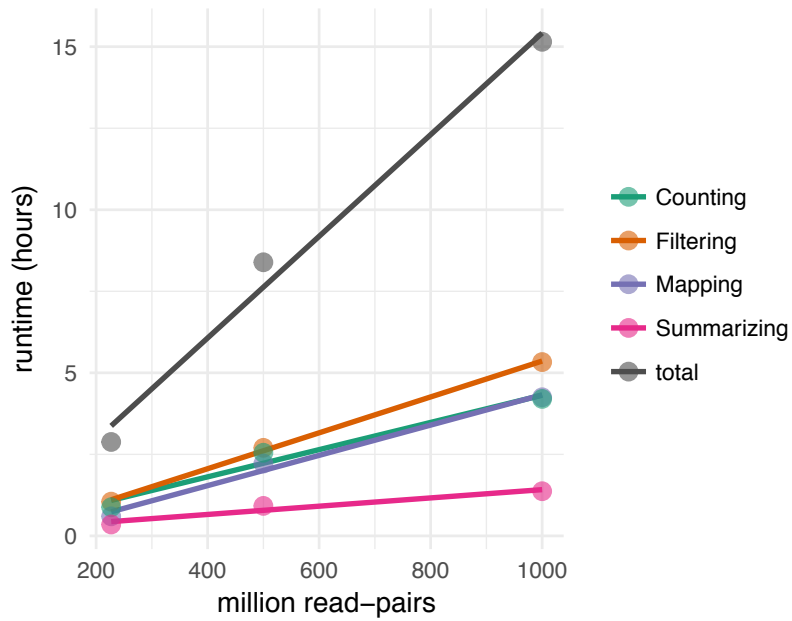


Figure S1: *zUMIs* runtime for three datasets with 227, 500 and 1000 million read-pairs. The runtime is divided in the main steps of the *zUMIs* pipeline: Filtering, Mapping, Counting and Summarizing. Each dataset was processed using up to 16 threads ("p 16").

2 zUMIs example dataset

At the end of each run, *zUMIs* optionally generates statistical output and plots. Shown here are the generated plots for the exemplary HEK cell dataset (Figure S2 and S3).

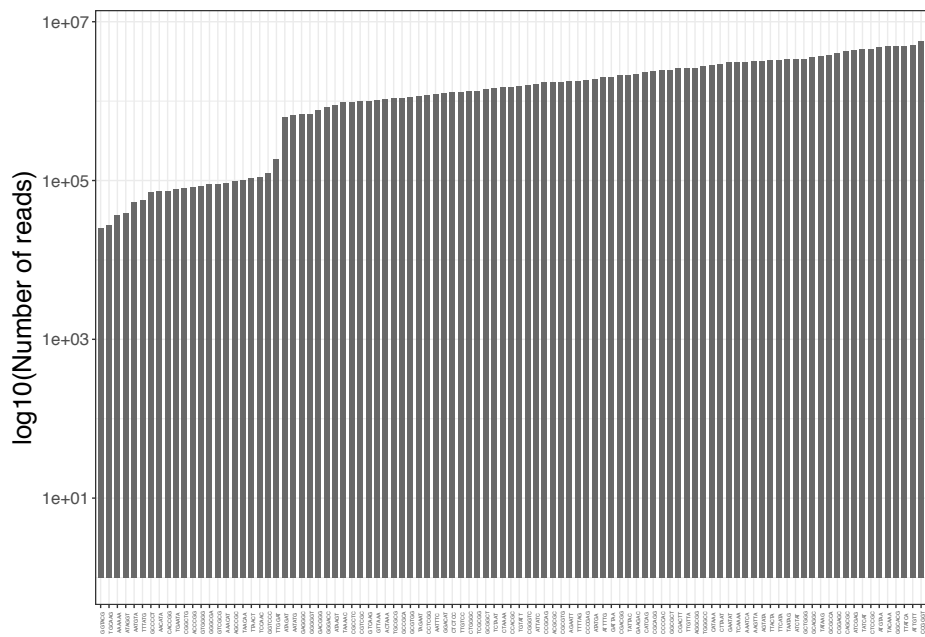


Figure S2: Reads per barcode. Bars show the number of reads assigned to each sample barcode.

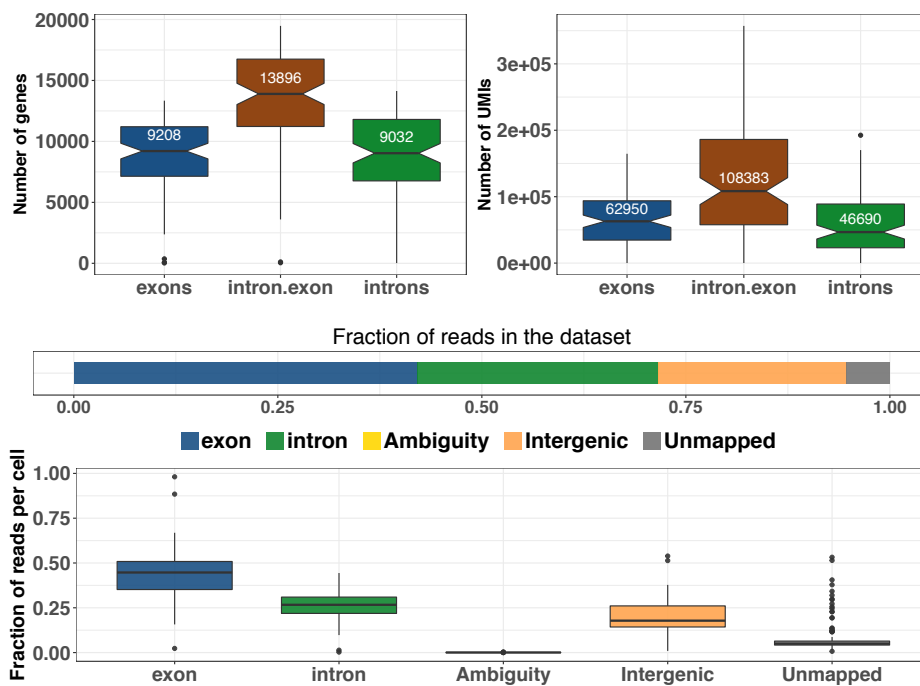


Figure S3: Summary statistics. The boxplot in the left panel shows number of genes (left) and number of UMIs(right) detected per barcode while considering only intronic/exonic counts and intronic+exonic counts. The horizontal relative barplot in the middle indicates total fraction of reads assignment to each feature in the dataset and the boxplot in the lower panel colored by features show fraction of reads assigned in each category where each data point is one cell.

3 Downsampling

zUMIs has inbuilt functionality for downsampling datasets to a user-specified number of reads. When the option "-d" is set, *zUMIs* will attempt to downsample all sample barcodes to the specified number. In case the requested read number is not available for some of the barcodes, only those barcodes will be reported that fulfilled the requirement. In any case, the full data will be output alongside the downsampled data. This basic downsampling is useful to make the often hugely varying library sizes for single cell data more comparable [1]. Another application of the downsampling function is to evaluate whether the current sequencing depth was sufficient to reach saturation of gene and UMI detection. To illustrate the downsampling functionality, we sample several fixed read depths for our exemplary HEK dataset and display the number of detected genes at given depth per cell (Figure S4).

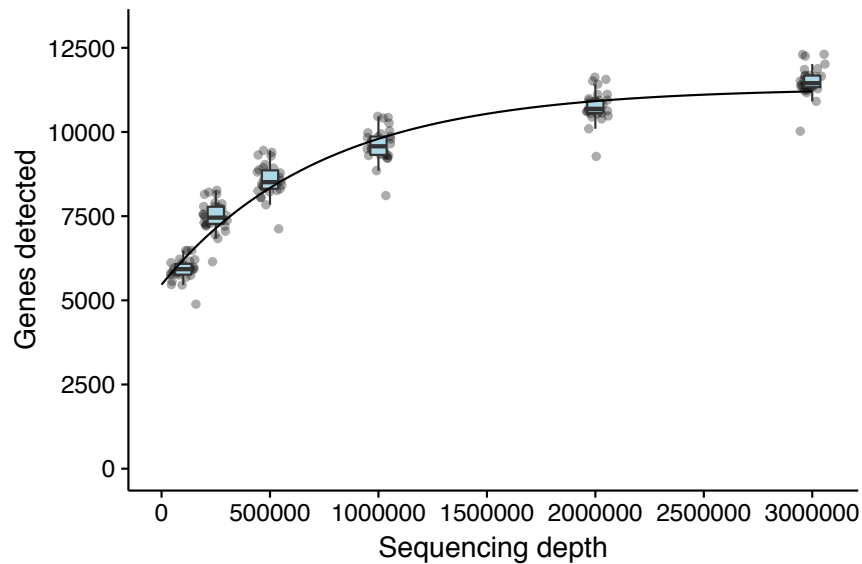


Figure S4: Downsampling. Cells were downsampled to six depths from 100,000 to 3,000,000 reads. For each sequencing depth the reads detected per cell is shown. Here the increase in the number of genes detected using 1 million as compared to 3 million reads is small, suggesting that 1 million reads per sample are sufficient.

References

- [1] Dominic Grün and Alexander van Oudenaarden. Design and analysis of Single-Cell sequencing experiments. *Cell*, 163(4):799–810, 5 November 2015.
- [2] Magali Soumillon, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S Mikkelsen. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*, 5 March 2014.
- [3] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, 65(4):631–643.e4, 16 February 2017.

Optimising cross species differential expression analysis

Strategies for RNA-seq differential expression analysis for closely related species

Strategies for RNA-seq differential expression analysis for closely related species

Swati Parekh¹, Beate Vieth¹, Christoph Ziegenhain¹, Wolfgang Enard¹, Ines Hellmann^{1,2}

¹ **Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians University, 82152 Martinsried, Germany**

² **corresponding author**

Abstract

With the growing appreciation for the role of regulatory differences in evolution, researchers need a quantitative measure of expression level differences between species. However, differences in the reference genomes, due to assembly and annotation quality or simply their inherent mappability can bias the inference of expression changes. Furthermore, for non-model organisms there is often no reference genome available at all. Here, we explore the possibility to map diverged RNA-seq reads to the one high quality reference genome. To this end we produce *in silico* evolved genomes representing a small primate phylogeny ranging from Human to Marmoset (12 % nucleotide divergence). From those genomes we then simulate RNA-seq reads. The mapping of those reads to the genome of origin (self-mapping) as well as to one common reference (cross-mapping) allows us to quantify the effect of sequence divergence alone on differential expression analysis. We also simulate RNA-seq reads based on Ensembl primate genomes to gauge the impact of assembly and annotation quality. We find that for very closely related species such as the great apes ($\leq 4\%$) divergence bias due to cross-mapping is acceptable and given the quality of the current great ape genomes, even preferable to self-mapping.

Introduction

Gene expression is an easily accessible phenotype that can help us bridge the gap between evolutionary changes of the genome and more complex phenotypes such as of brain size [e.g. 1]. Since the evolution of human-specific traits is a favorite issue in biological research, the first high-throughput comparisons of gene expression across species focused on primates and used oligonucleotide arrays that were designed based on human cDNAs [1]. Hence one of the biggest criticisms of this study was that, due to sequence divergence, between species expression differences could be confounded with hybridization differences. When the Chimpanzee genome became available in 2005 [2], it became possible to only

select oligonucleotides without any substitutions between Humans and Chimpanzees [3]. However, for slightly more distant species such as Macaques, this strategy turned out to be less desirable, because the number of probes with identical sequence became too small. To this end efforts to model probe binding affinities also for probes containing species differences came underway [4]. Even though this was a viable approach, the emerging next generation sequencing methods seemed preferable over microarrays. RNA-seq promised to alleviate several problems due to the inherent reference bias of the common one-species arrays. Using RNA-seq, at least the sampling of the transcriptome was unbiased and the sequence information contained within the reads allowed for a more detailed analysis and possibly a correction for divergence. Nevertheless, a quantitative comparison of gene expression levels between species remains far from trivial.

For a quantitative comparison, we must be cautious about systematic technical differences between species. Such systematic differences in RNA-seq studies can easily be introduced by differences in our ability to associate reads with genes, which in the simplest case could be due to the lack of a reference genome in one of the examined species. However, even if a reference genome is available, mappability of RNA-seq reads will depend on the quality of the reference genomes as well as the quality of the gene annotation [5]. Indeed, the differences in genome quality are considerable and no available primate genome comes close to the quality of the human genome. For example, the N50 contig sizes of Chimpanzee, Gorilla, Orangutan, Marmoset and Macaque are between 5 and $50 \times$ smaller than hg38 and only one third as many exons are annotated in Ensembl (Table 1).

However, the most difficult part is to define transcriptional units that are comparable across species. Blekhman et al. [6] identified orthologous exons as exons that were present in all three species of that study, i.e. Human, Chimpanzee and Macaque. Going exon by exon rather than attempting to align whole transcripts helped to avoid truncating genes due to bad genome assemblies. In contrast, Brawand et al. [7] opted to find orthologous transcripts. Because transcript annotation is biased due to the use of Human gene models to annotate primate genomes, Brawand et al. [7] first used their RNA-seq data to improve the annotation of primate transcripts, before intersecting the transcripts to then only count reads from orthologous parts. This reduction to only common exons or transcript parts also serves as a correction for differing gene lengths between species, however this correction is only valid, if reads are evenly sampled across the transcript. Zhu et al. [8] implemented several of the above ideas of identifying and filtering orthologous exons in one pipeline and added an empirical measure of mappability. Finally, Wunderlich et al. [9] used curated whole genome multiple alignments [10] to convert coordinates of reads mapped to non-human primate genomes to Human coordinates and then used the human gene annotation as comparable transcriptional unit.

In summary, efforts so far focused on finding and filtering orthologous genes/exons and then map to the genome of origin, counting only reads mapping to the stringently defined orthologous set. So far, nobody has quantified the expected mapping bias, when mapping to one genome, that is chosen according to assembly and annotation quality.

With simulations of evolution whole genome sequences on a primate phylogeny as well as gene expression levels, we show that mapping to common high quality reference is a viable option for closely related species. Furthermore, our simulations allow us to provide recommendations on sequencing and mapping strategies.

Results

Quality of six primate genomes

We downloaded human genome and the genomes of 5 non-human primates: Chimpanzee, Gorilla, Orangutan, Macaque and Marmoset. The Chimpanzee is with a average nucleotide divergence of 1.3% the most closely related species and with and the Marmoset with 12.3% the most distant of the downloaded genomes [11]. The genomes also vary with respect to assembly quality, whereas the Orangutan has with an N50 of 0.75Mb the shortest scaffolds. Differences in the completeness of the annotation are even bigger, while the human genome has $\sim 43,000$ genes and $\sim 676,000$ exons, Orangutan has only $\sim 27,000$ genes and 213,000 exons annotated (Table 1). Another, for our purposes, more direct measure of genome quality is our ability to unambiguously map reads back to the genome from which they were sampled. Assuming that the repeat content and thus the uniqueness of kmers is expected to be similar, we expect most differences in mappability to be due to ambiguous characters in the genome. We calculate the expected fraction of unique and ambiguous 50mers in the genomes allowing for up to two mismatches [12]. In line with the other genome statistics, also mappability shows that no other primate genome comes close to the quality of hg38: 85.13% of all exonic 50mers are unique (Supplementary Figure S1). The other good genome is the Macaque with 85.45% unique exonic 50mers, the other four primate genomes vary between 69-80%. Longer kmers improve the mappability by average $\sim 11\%$, but mappability still varies systematically among genomes according to their quality. This wide variation in genome and annotation quality, led us to explore the possibility to map all species to the best reference. This strategy would be valuable for species without a genome assembly and might be suitable alternative for very closely related species with large variation in genome and annotation quality.

However, sequence divergence between species is also bound to interfere with mapping and quantification of RNA-seq reads and in reality genome quality and divergence will be convoluted. In order to clearly separate those effects, we decided to simulate genomes corresponding to the primate tree. To this end we used *evolver* [13]. *Evolver* integrates all of our knowledge about genome evolution, one can provide rates for various events such as mutations affecting single nucleotides but also large chromosomal rearrangements thus also allowing for gene duplication, the accept probability of the suggested events is then determined by the amount of constraint on the affected sites. Thus mutations at synonymous sites gene or within the 3'UTR of a coding gene are more likely to remain

than mutations that change amino acids or the loss of an entire exon. Only events that lead to a complete loss of a gene are forbidden, everything else can occur. We used the human genome as the most complete genome with the most comprehensive annotation as a starting point (root) and simulated six derived genomes from corresponding to the primate phylogeny given in Figure 1. The simulated genomes allow us to track genes throughout the tree, so that we know the orthologs and do not need to infer them, which can be tricky [14]. Furthermore, the quality of all simulated genomes is identical, allowing us to assess the effect of sequence divergence alone.

Mapping to a diverged genome

Using Ensemble biomaRt[15] we identified 9,257 human genes that have a one-to-one orthologue in each of the five primate genomes. For those orthologous genes we simulated RNA-seq reads [16]. We did the same for the evolver simulated genes with $\sim 15,000$ orthologous genes (Figure 1).

We also evaluated four different sequencing strategies with respect to their mapping to a diverged genome. We simulated reads with lengths of 50 and 100bp as well as the sequencing of the cDNA fragment from either only one side (single-end, SE) or from both sides (paired-end, PE). We then mapped to the human genome (cross-mapping), using STAR, a splice-aware mapper that also uses existing annotation of splice junctions to improve mapping [17]. In order to evaluate the mapping, we calculated the fraction of reads that remained unmapped and the fraction of reads that were correctly mapped. With only up to 2% the amount of reads that map to the wrong location is small (Supplementary Figure S2), so that the fraction of correctly mapped reads is approximately the inverse of the fraction of unmapped reads. At the first glance, it may seem surprising that divergence in the simulated genomes has a larger effect on mapping than for the Ensembl genomes. This is due to that we simulated expression only for one-to one orthologs, which enriches for highly conserved genes.

Nevertheless, it remains that for both simulated and Ensembl genomes, the fraction of unmapped reads notably increases in the two most diverged genomes (Macaque 2.3% and Marmoset 8.4%), while mapping Chimpanzee (0.6%), Gorilla (0.9%) and Orangutan (1.4%) reads to the human reference results in almost no loss due to divergence (Figure 2). Evaluating the different sequencing strategies, we find that longer reads improve mapping to the more diverged genomes. The unmapped fractions were reduced for 100bp by 1.8% and 4% for Macaque and Marmoset, respectively. Surprisingly, the mapping of PE reads did not improve mapping, on the contrary the fraction of unmapped reads increased. Closer inspection of our simulated data showed that mainly reads from genes with exon gain or loss events were affected. Therefore, we loosened the criteria for reporting proper pairing of PE reads, following an iterative mapping strategy (see Methods) which improved the mapping of PE reads (Figure 2). However, compared SE sequencing the optimally mapped PE sequencing only reduced the unmapped fraction by 0.5% , which is not enough to justify the substantially higher sequencing costs. We therefore focus on 100bp SE reads for the remainder of the study.

Impact of gene-wise divergence on mapping

If the number of reads that is lost due to mapping to a diverged genome was evenly distributed across all genes, this would have little impact on measures for differential expression: The other genome would simply appear to have fewer reads in total and standard normalization procedures would take care of such a discrepancy. Problems will only arise, if due to divergence some genes lost more reads than others, i.e. it is not the total divergence that is of interest, but the variance in divergence - and thus mapping success - across genes.

To this end we use the cross-mapping data from the simulated genomes, whereas we use the mapped reads from all replicates to estimate divergence (see Methods). Note that estimating divergence from the RNA-seq data itself makes it possible to obtain divergence estimates for species without other available genome or transcriptome sequences.

Indeed, we find that the fraction of correctly mapped reads has an inverse correlation with gene-wise divergence and this correlation increases with species divergence (Figure 3). Chimpanzee and Human are so closely related that the little divergence has no discernible effect on the mapping success (Pearson's $r = -0.01$, $p = 0.28$), while for Marmoset and Macaque the fraction correctly mapped reads strongly depends on the sequence divergence of the gene (Cjac $r = -0.50^{***}$, Mmul $r = -0.39^{***}$).

Based on this result we would speculate that mapping Chimpanzee reads to the human genome should not bias differential expression analysis, while there will be an effect in Marmoset and Macaque. However, it might be possible to use the divergence estimates to correct the counts.

DE-analysis across species for simulated genomes

The main goal of this study is to evaluate different strategies for DE-analyses in closely related species. We used the flux simulator [16] to simulate 100bp SE reads for the known orthologs from the six simulated primate genomes, keeping the expression levels fixed across all species. To allow for a meaningful DE-analysis, we simulated 6 replicates/species and we want to evaluate whether it is possible to detect expression changes between the species. We use DESeq2 [18] and compare each of the five *in silico* primate RNA-seq datasets to Human as a common reference.

Reads from non-human primates were mapped to both, the genome of origin (self-mapping) as well as the simulated human genome (cross-mapping). Hence for self-mapping we expect near perfect mappability, but because read counts from different genomes are compared this strategy will be sensitive to genome and annotation quality. Therefore, it is not surprising that the number of false positives for the *in silico* genomes with known orthology and perfect quality is negligible (up to 0.1% Marmoset). With cross-mapping, the false positive rate (FPR) increases with divergence. Chimpanzee and Human are close enough, so that divergence does not increase the FPR. In fact, for both Human and Chimpanzee we did not observe any false positive genes in our simulated data (Figure 4). Also for Gorilla and Orangutan the FPR is with 0.04% and 0.34%,

respectively, acceptable. Moreover, the FPRs are comparable for all three counting strategies, self-mapping, cross-mapping and divergence corrected cross-mapping (see Methods).

For the two more distant genomes Macaque and Marmoset cross-mapping clearly produces more false positives than self-mapping (Macaque FPR cross 2% , self 0.09%; Marmoset FPR cross 14%, self 0.11%). As shown in Figure 3, divergence to Human in those two primates is high enough to effect the mapping probability. For genes with a high divergence fewer reads can be recovered. This is also evident from the observation that the \log_2 -fold changes are biased towards a higher expression in the Human, i.e. the reference species ($\sim 60\%$) (Figure 4A, Supplementary figure S4). In order to correct for unmapped reads due to divergence, we use a log-linear model to predict the number of sampled reads from the number of mapped reads per gene and the divergence estimate (see Methods). This strategy halves reduces the FPR by \sim half (Macaque 1.2%; Marmoset 7.9%), but the FPR still increases with species divergence and is high compared to self-mapping. However, even though divergence to the reference genome introduces detectable false positives, the effect size of those changes is small. Most \log_2 -fold changes are between 1 and -1, and thus smaller than an absolute 2-fold difference (Figure 4B). If we require genes to have a significant absolute \log_2 -fold change of at least 1, the FPR for cross-mapping counts reduces to 0.3% in Macaque and to 1% in Marmoset. Interestingly, divergence correction cannot improve this FPR, suggesting that higher \log_2 -fold changes are not caused by lower mapping rates due to sequence differences. On the contrary, those genes show an enrichment for falsely mapped reads (Supplementary Figure S2). Misplaced primate reads will also increase our divergence estimates and thus over-correct the number of reads letting the expression of the gene appear higher in the nh-primate (Supplementary Figure S3). In summary, our simulations suggest that using some corrections, mapping to a diverged genome for quantitative expression analysis is a viable option.

DE-analysis across species for Ensembl-genomes

Next, we wanted to investigate in how far we can generalize our findings to the more realistic scenario, with varying genome quality and annotation. To this end we simulate RNA-seq reads from the 'real' primate genomes as downloaded from Ensembl (Table 1), otherwise the experimental setup and the DE-analysis were the same as for the *in silico* genomes and should thus be comparable.

To begin with, the self-mapping FPRs are much higher for the Ensembl than for the *in silico* genomes (ensembl: 1.5-2.6% vs. *in silico* $\sim 0.1\%$) . Part of this discrepancy might be due to differences in gene models. Due to annotation problems, differences in gene lengths appear exaggerated in the Ensembl genomes. For example, 3'UTRs are consistently shorter in the non-human primates (Table 1). In order to correct for these gene length differences, we only counted reads mapping to orthologous sequences that were annotated in both the Human and the non-human primate [7, e.g.]. This helped to reduce the FPR to $\sim 1\%$ in all comparisons (Figure 5 and Supplementary Figure S5).

When using counts from cross-mapping, the FPRs for the great ape genomes are low: With 0.6% (Chimpanzee), 0.7% (Gorilla) and 1% Orangutan, cross-mapping FPRs are even lower than counts for self-mapping and counting of orthologous sequences only.

As in the simulated genomes, the FPRs for Macaque (2%) and Marmoset (8.5%) are much higher than in the great apes. The FPR for Macaque agrees with the level estimated for the simulated genomes, while the FPR from the simulated Marmoset genome (14%) is substantially higher. This suggests that the set of one-to-one orthologs is biased towards conserved genes, and the Marmoset as the most diverged species in the comparison is the limiting factor to find one-to-one orthologs across all species. Using divergence estimates to correct cross-mapping counts does not reduce the FPR in Macaque, and provides only a slight improvement by 2.2% for Marmoset (Figure 5, Supplementary Figure S6). Finally, also the effect sizes of the false positives are low, as for the simulated genomes, instating a cut-off for absolute \log_2 -fold changes of at least 1 reduces the FPR for all cross-species comparisons to $\leq 0.5\%$.

Within species DE-analysis while mapping to a diverged reference

Although divergence somewhat impairs DE-analysis between species, it is unclear how mapping to a diverged genome would impact DE-analysis within a species. This would allow to conduct DE-analysis for species for which only a genome of a close relative has been assembled. To test such an analysis strategy, we simulate RNA-seq data for each of the six simulated genomes and two conditions for each species. 10% of the genes are simulated to have a 4-fold change between conditions, whereas equal numbers of genes are up and down regulated (Figure 7A). Again we map all reads to the Human genome and analyze differential expression using DESeq2 to tabulate how many of the simulated DE-genes can be recovered (=sensitivity, true positive rate, TPR) as well as the proportion of false discoveries among all significant genes (=specificity, false discovery rate, FDR). The mapping of human reads to the human genome yielded a TPR of 99% and an FDR of 2.5% (Figure 6). This represents the best case scenario, i.e. self-mapping to a high quality, well annotated genome. Chimpanzee reads yield similarly good sensitivity and specificity. Mapping Gorilla, Orangutan and Macaque reads to the Human genome also yields a sensitivity close to 99%, but the FDR appears slightly increased to 3% in Gorilla and Orangutan and 4.5% for Macaque. Only for Marmoset with an FDR of 6% exceeds the nominal level of 5% and also shows a noticeable drop in TPR to 97% (Figure 6). Because both conditions are affected by divergence to the same extent, divergence correction does not help to reduce the FDR but only decreases sensitivity (Figure 7B). In the absence of an adequate reference, mapping to diverged reference for within species DE-analysis yields good sensitivity and only the FDR for Marmoset starts being problematic.

Detecting relative expression changes between species

In many instances, changes in the regulation of a gene are easier to interpret than expression differences measured under one condition [19]. For example, if one gene is up-regulated during development in one species, but not in the other, the first time-point serves as an internal reference that facilitates the detection of this regulatory difference. This makes the inference of the interspecies difference more robust towards technical bias, including systematic species differences.

For differential expression analyses tools based on (generalized) linear models [18, 20] such a regulatory difference would be formulated as the interaction term between species and condition, i.e. the expression change between conditions depends on the identity of the species. In the previous chapter, we described the simulation of two conditions within each species. Note that we simulated no expression changes between species, in other words condition A, as condition B, should have identical expression profiles for both species. Hence, also the expression changes between conditions occur in both species and thus a significant interaction term species:condition represents a false positive regulatory change.

Again in the comparison between Human and Chimpanzee, there are no false positives, and the FPRs for the other great apes with 0.05% in Gorilla (7 genes) and 0.12% in Orangutan (19 genes) are also very low. Even though Macaque (FPR=0.4%) and Marmoset (FPR=0.6) have higher FPRs than the great apes, they are still low compared to the FPRs for the comparison of absolute expression changes (Figure 4A). Furthermore, divergence correction had no impact on the FPR, confirming that a relative model efficiently corrects for species differences that are due to divergence.

Hence, even though also for relative expression changes, FPRs increase with divergence, it is much better controlled than for absolute expression changes. Thus keeping an cautious eye on the FPR levels, also comparisons for more distantly related species such as Human and Marmoset will yield meaningful results.

Discussion

RNA-seq is a versatile tool that also has many applications in evolution and ecology [21]. However, for many of the organisms of interest, no reference genome exists and even if a reference genome is available, the quality of the genome and the annotation is highly variable. One simple solution to this problem would be to use available reference genomes from closely related species. Here, we investigate the biases that come through mapping RNA-seq to a diverged genome for three different biological questions: 1) Finding expression differences between species, 2) Finding expression differences within the non-reference species and 3) Finding relative expression changes between species. To this end, we use a small primate tree including Chimpanzee, Gorilla, Orangutan, Macaque, Marmoset and Human. Neutral divergence ranges from 1.34 -12.1% (Figure 1). For all those species published genomes exist, however they largely vary in quality, in

particular annotation of the Human genome is much more complete than the annotation of any of the non-human primates. In order to distinguish between problems due to annotation and actual evolutionary differences, we generated *in silico* genomes without assembly or annotation errors. Furthermore, orthologous genes in the *in silico* genomes are known and hence do not show a bias to higher conservation as we saw in the inferred orthologs in the real genomes.

We then simulate RNA-seq reads from real and *in silico* genomes and map them to the Human reference. As expected, the fraction of reads that can not be mapped increases logarithmically with divergence (Figure 2). Sequencing of longer reads improves the mapping a little ($\leq 3.9\%$), so that one might think that also paired-end could bring a further improvement. However, our simulations show the assumed positioning of the pair is not necessarily correct in the diverged species, thus leading to the loss of reads from genes with rearrangements or exon gain or loss events. Hence given our simulations, we think that 100bp single end reads are the best sequencing strategy here. That being said, a genome-wide reduction in the number of mappable reads does not necessarily introduce bias into expression analyses. Bias is only introduced, if the fraction of correctly mapped reads varies systematically among genes and we can indeed show that there is an inverse relationship between the divergence of a gene and the fraction of correctly mapped reads (Figure 3). Therefore, for the first biological question investigating between species expression differences, we expect that more diverged genes will appear to have a lower expression in primates as compared to humans, which is indeed the case (Figure 4), Supplementary Figure S2). Because we can also use mapped RNA-seq reads to estimate gene-wise divergence to the reference genome, we attempted to regress the effect of divergence on the read counts out. The FPR for these divergence corrected read-counts was roughly half of what it was without divergence correction. Furthermore the direction of change became more symmetric, in that false positives could also appear to be down-regulated in Humans. However, especially for the more diverged species (Marmoset and Macaque), some bias towards higher counts in Humans remains. We believe that the reason for the remaining false positives and their directional bias is an underestimation of the true divergence of a gene. Both Marmoset and Macaque have genes that with a divergence of $>10\%$, which is the threshold from which on STAR begins to have problems with mapping. Because we cannot map to regions with high divergence, those cannot be included in the divergence estimation, thus making our correction insufficient. Furthermore, bad divergence estimates can also explain some of the additional false positives with much higher counts in primates. Mismapped reads will inflate our divergence estimates and trigger our regression model to increase the count even more (Supplementary Figure S2 & S3). Fortunately, false mapping rates are low, so that only a few genes will show this rather counter-intuitive pattern. Generally, false positive DE-genes have low \log_2 -fold changes, so that an additional restriction on effect size gets rid of the vast majority of false positives, even for Marmoset (Figure 4).

Most of the results obtained using *in silico* genomes could also be recapitulated with the Ensembl genomes, with the only exception of substantially higher FPRs for

self-mapping, *i.e.* mapping reads to the genome of origin and counting only orthologous regions. Restricting the regions to only the ones that are also annotated in the primate genome should theoretically correct for any gene-length differences, which are probably the cause for the even higher FPRs if this correction is left out (Figure 5). However, this correction assumes that RNA-seq reads are evenly sampled across the entire transcript length, which in practise is rarely the case [22]. Thus, for very closely related species such as Chimpanzee and Gorilla, cross-mapping actually produces fewer false positives. For those two species the benefits of a good annotation and genome quality outweigh the problems introduced by sequence divergence.

The second question that we investigated was in how far cross-mapping can be used to analyze differential expression within the same species. To this end we simulated RNA-seq reads for two conditions in which 10% of the genes were differentially expressed with a \log_2 -fold change = |2|. Mapping the primate reads to the Human genome yields a high sensitivity of over 97% for all species and only the Marmoset had an FDR slightly above the nominal level. This suggest that mapping bias is by in large consistent and thus cancels out when contrasting the two conditions.

Following up on this notion, we investigated in how far relative differences between species can be detected using cross-mapping. We re-used the simulations of two conditions in each species, expression profiles differ between conditions, but not between species. Thus we would not expect to detect any relative changes between species, *i.e.* if a gene is upregulated under one condition in one species, the same change is expected to be seen in the other (Figure 7 and Supplementary Figure S7). Even though the FPR for relative expression changes between species increases with species divergence, it is $\sim 10\times$ lower than the FPR for absolute between species changes.

Hence, in the absence of a good genome assembly or annotation cross-mapping to a closely related species to analyze expression differences between conditions within the same species or to detect relative differences between species is a good alternative. As a general rule of thumb, cross-mapping works as long as the divergence does not exceed the limits of the mapper used. If divergence for all genes remains below this threshold, no further corrections are necessary. However, if some genes exceed the divergence that can be safely handled by the mapper, quantitative comparison between species will produce more false positives, which can be dealt with by introducing an effect size cut-off.

Methods

Generating *in silico* Genomes

We used *evolver* [13] to simulate whole genome sequence evolution. Given an ancestral genome, annotation of gene models, CpG islands, a repeat library as well as rates for nucleotide and amino acid substitution, indels, tandem repeat expansion and contraction *evolver* models sequence evolution of an entire genome. To simulate gene expression of diverged species using RNA-seq it is important to explicitly model the evolution of

coding regions, UTRs, start and stop codons and splice junctions. All possible types of mutations can also affect functional regions, but *evolver* constrains their number based on the set amount of selection on the affected sites. Hence genes can also be rearranged or duplicated and exons can be gained or lost.

To evaluate the effect of mapping and quantification for diverged species on realistic data, we used human genome (hg19) and gene models (GRCh37.75) as ancestral genome because it is the best reference available. Starting with hg19 at the root, we generated six *in-silico* genomes based on the neutral divergence estimates for a primate phylogeny (Figure 1).

We used the hg19 parameter file provided by *evolver* to set up all the rates to model evolution. To set up our simulations, we first replaced all ambiguous bases in hg19 by random bases using the "evo -findns" module from *evolver* and converted them to rev format using "evolver_cvt -fromfasta -torev". We obtained gene models in gff format from the UCSC genome browser [23] and filtered for non-overlapping genes using "evolver_evo -cvtannots". We obtained 15,559 genes after filtering. Rest of the genomic features required for the simulations were set according the default settings in *evolver*.

evolver has two major modules for inter and intra chromosomal evolution simulation. Starting from a common ancestral genome and gene models, we run *evolver* for every internal to leaf nodes on a given 6 primate tree (Figure 1). The inter chromosomal evolution is run by "evolver_evo -interchr" where between chromosomal events like segmental moves, chromosomal fission and fusion take place. On these evolved chromosomes, we simulate intra chromosomal events like substitutions and indels in parallel for each chromosome using "evolver_evo -inseq".

All the simulated genomes are converted from rev to fasta format "evolver_cvt -fromrev -tofasta" and CDS-UTR features into exons using "gff.cdsutr2exons.py" provided in *evolver* script collection. These genomes and gtf files are then used for RNA-seq simulation.

RNA-seq simulation

RNA-seq reads for all 6 primates were simulated using flux-simulator (Griebel et al. 2012) from *in-silico* as well as the Ensembl genomes [24] (Figure 1 and Table 1). The reads were generated as single-end and paired-end with read lengths of 50 and 100 bp each. The error models were built from bulk Smart-seq2 data sequenced on Illumina HiSeq1500 (Supplementary figure S8) using "flux-simulator -t errormodel -tech phred -o mymodel.err -f gemmapping.map". For all 6 primates from simulated and Ensembl genomes, 10 million reads for 6 replicates were simulated totaling to 288 bulk RNA-seq simulations. We simulate equal numbers of expressed molecules for each orthologous gene across species. We first run flux-simulator once with "-x (simulate expression)" option and then modified the .PRO file to simulate the same number of molecules for orthologous genes. Using the modified .PRO files, we proceed library preparation and sequencing steps with "-l" and "-s", respectively.

Mapping and Quantification

The simulated RNA-seq reads were mapped using STAR_2.5.3a [17] in "twopassMode Basic" mode with default parameters, allowing up to 50 multi-mapping reads. Gene models (GTF file) were provided on the fly at the mapping stage whereas the genome index was generated without "-sjdbOverhang" or "-sjdbGTFfile". For paired-end layout, we use an "iterative mapping" strategy. First, the reads were mapped as paired-end and unmapped reads were extracted using "-outReadsUnmapped Fastx" option in STAR. In the second iteration, the unmapped reads were mapped as single-end, ignoring the information of their mate reads. The mapped files from both iterations were merged using "samtools merge".

To address the question if we can use "Human" as a reference for the nh-primate, we employed two different mapping strategies. One being "self-mapping" where reads from different primates were mapped to their own genomes and gene models. Second, "cross-mapping" where reads of all the primates were mapped to human genome. Reads generated from *in-silico* simulated genomes and Ensembl genomes were mapped to the simulated human genome or Ensembl GRCh38.82 genome, respectively.

In order to correct for the annotated gene length differences in the Ensembl genomes, read counting was restricted only to the orthologous regions annotated in human and nh-primates. The orthologous regions were derived using reciprocal liftOver [25] between human and nh-primates and filtered for the genes used for simulations using "bedtools intersect" [26].

The mapped reads were assigned to genes using featureCounts from the bioconductor package Rsubread [27].

Gene-wise Divergence estimation

We empirically estimate divergence of each nh-primate gene to its human ortholog from reads cross-mapped to human. In order to reduce false mismatch calls due to indels, we perform local realignment for indels using GATK "IndelRealigner" with default parameters. Prior to indel-realignment, we pre-processed the BAM files with "AddOrReplaceReadGroups" and "SplitNCigarReads". Pre-processed BAM files are merged using "samtools.1.3.1 merge" and genotype likelihood is generated in VCF format using "samtools.1.3.1 mpileup -Q 0 -uv" parameters. We use a custom R script to estimate gene wise divergence from VCF file. To incorporate sequencing quality in our divergence calculation, back-calculate the probability of having the correct base from the Phred score and sum them up across reads, sites and bases. Substitutions are only counted for bi-allelic sites. To obtain divergence estimates, we use Kimuras two parameter model to correct for multiple hits [28].

Differential Gene Expression

We performed pairwise differential expression analysis of each nh-primate to human from cross mapped counts on simulated data. Genes expressed in less than two samples are filtered out and DESeq2 [18] was used to perform DGE with standard parameters. We used Benjamini Yekuteili (BY) for p-value adjustment under nominal alpha 0.05 for each nh-primate vs Human comparison.

To adjust the abundance level of genes for sequence divergence to equivalent human loci, we used empirically estimated divergence to correct expression levels of each gene. We used a log-linear fit between simulated counts (*sim*) and counts from cross-mapping (*cross*) to human with estimated divergence(*div*) as an interaction term to cross mapped counts (Equation 3).

$$\log_{10}(sim + 1) \sim \log_{10}(cross + 1) + \log_{10}(cross + 1) : div \quad (1)$$

We performed the same DGE analysis as explained above with the fitted counts from this model and compared the rate of false positive calls.

Relative expression differences

We used unique molecule counts from 12 bulk UHRR samples from previously published dataset [22] to estimate mean and dispersion and simulated a 2-group gene expression profile with 6 replicates in each group with 10% differentially expressed genes at \log_2 fold-change of 2 and -2 symmetrically using powsimR [29]. These molecule counts are further used to adapt the number of expressed molecules in .PRO files in flux-simulator for each primate to simulate RNA-seq reads. In this simulation framework, we keep no expression difference between species while within species between conditions the same 10% genes are differentially expressed (Figure 7A). RNA-seq reads are generated from each species but all the reads are cross mapped to human reference and DGE is performed using DESeq2 with the design formula " \sim species + condition + species:condition". We perform differential expression testing between condition in each species and the condition effect between species (interaction term). We used Benjamini Yekuteili (BY) [30] for p-value adjustment under nominal alpha 0.05 for all the comparisons.

Supplementary Material

Supplementary figures S1-S8 and Supplementary table S1 are available at Molecular Biology and Evolution online (<http://www.mbe.oxfordjournals.org/>).

Species	Assembly	Ensembl	Genome 10 ⁹ bp	N50 Scaf- folds Mbp	Genes/ 1000	Exons/ 1000	avg 3.UTR length (bp)
Human	hg38	84	3.21	67.79	43	676	269
Chimpanzee	Ptro2.1.4	85	3.31	9.14	27	203	83
Gorilla	Ggor3.1	85	3.04	0.91	28	237	73
Orangutan	Ppyg2	85	3.11	0.75	27	213	86
Macaca	Mmul8.0.1	86	3.24	4.19	32	276	100
Marmoset	Cjac3.2.1	85	2.91	5.17	30	313	84

Table 1. Genome assembly and annotation properties of 6 primates.

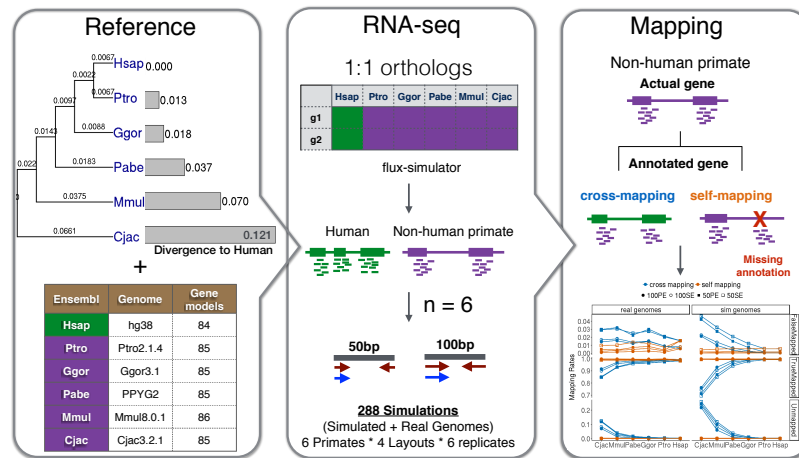


Figure 1. Schematic overview of the study design. The first panel starting from left shows two set of genome references used in this study 1) evolver genomes: in-silico genome sequence evolution simulation using hg19 as a common ancestor and a 6 primate tree. The primates are named as abbreviation of their scientific names (Cjac = Marmoset, Mmul = rhesus Macaque, Pabe = Orangutan, Ggor = Gorilla, Ptro = Chimpanzee, Hsap = Human). 2) Ensembl genomes: the table lists assembly and annotation versions of genomes and gene models downloaded from Ensembl for 6 primates. The middle panel shows RNA-seq simulation workflow. We simulate RNA-seq reads using flux-simulator under no biological variance across 1:1 orthologous genes across primates. Reads are simulated with 6 replicates and 4 sequencing layouts totalling to 288 RNA-seq datasets. The last panel shows two mapping strategies employed in this study. 1) cross-mapping: reads from all the primates are mapped to human reference and 2) self-mapping: reads are mapped to species' own genomic reference. The schematic here shows the under representation of read counts in a gene with missing annotation in non-human primates. "Actual gene" refers to a biologically present gene in the species and "Annotated gene" refers to available information annotated in a GTF (gene models) file in the databases.

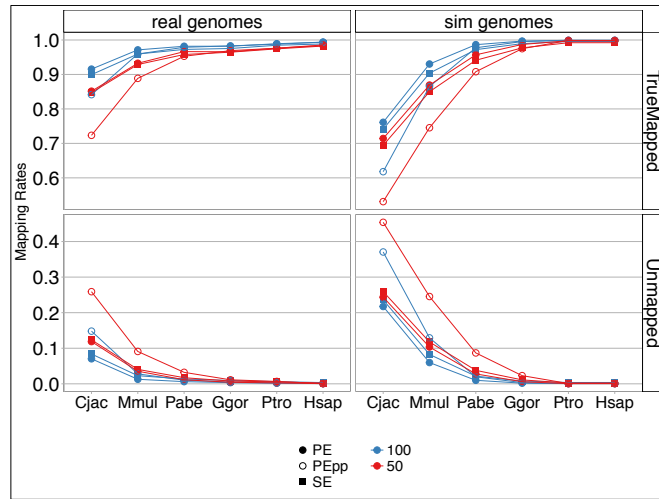


Figure 2. Cross-mapping statistics for *in silico* and Ensembl genomes. Here we plot the fraction of true mapped and unmapped reads (horizontal panels) with species on x-axis arranged by the evolutionary distance from Human (Hsap). The left panels show mapping statistics for reads sampled from Ensembl genomes, the right from evolver simulated genomes. Squares are single-end and circles are paired-end layouts. Empty circles represent the statistics if only proper pairs were kept, the filled circles represent an iterative approach.

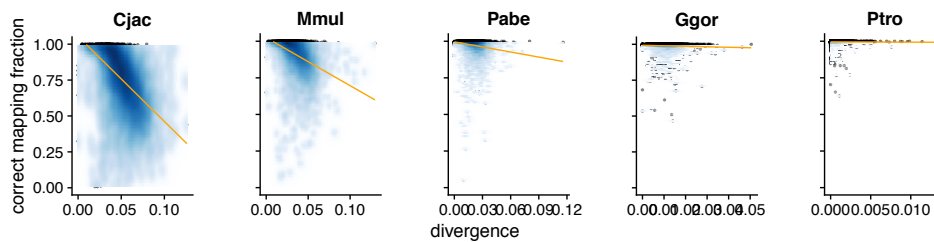


Figure 3. Gene-wise divergence impacts the fraction of correctly mapped reads. Divergence was estimated for genes with > 100 simulated reads. The fraction of correctly mapped reads, is the fraction of reads that could recovered from the ones that were sampled from that gene. The grey line is the linear regression of the divergence on the mapping fraction.

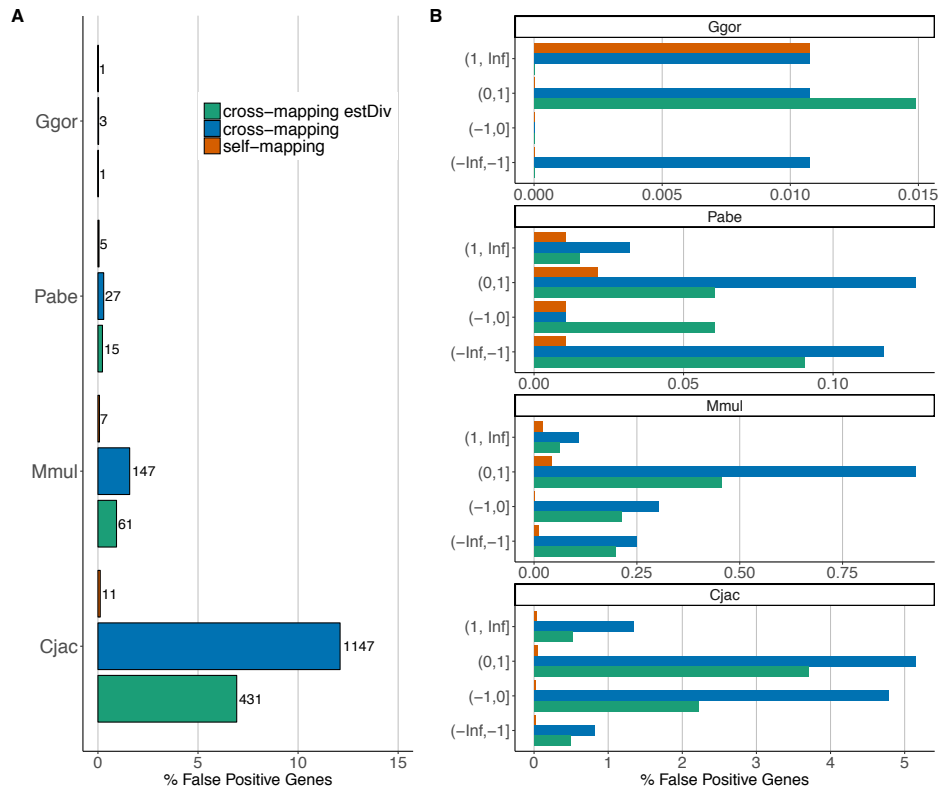


Figure 4. False Positive DE genes for different analysis strategies for *in-silico* genomes. We use simulated genomes with equal quality and known orthologous genes to simulate 100bp SE RNA-seq reads for six *in-silico* genomes, whereas the expression levels for orthologous genes were kept constant. Reads were either mapped to the genome of origin (self-mapping) or to the *in silico* Human genome (cross-mapping). In the cross-mapping scenario, we either took the read-counts or we corrected the read counts for the divergence of the gene (see Methods). We then used DESeq2 [18] to find DE-genes between Human and the non-human primate ($FDR \leq 0.05$). Panel **A** shows the overall % FPR for all nh-primate-Human comparisons for all three counting strategies. The numbers give the numbers of false positives genes in the comparison. In panel **B**, we stratify the FPR according to the estimated \log_2 -fold change. A positive \log_2 -fold change indicates a higher and a negative lower expression in the Human (reference) as compared to the nh-primate.

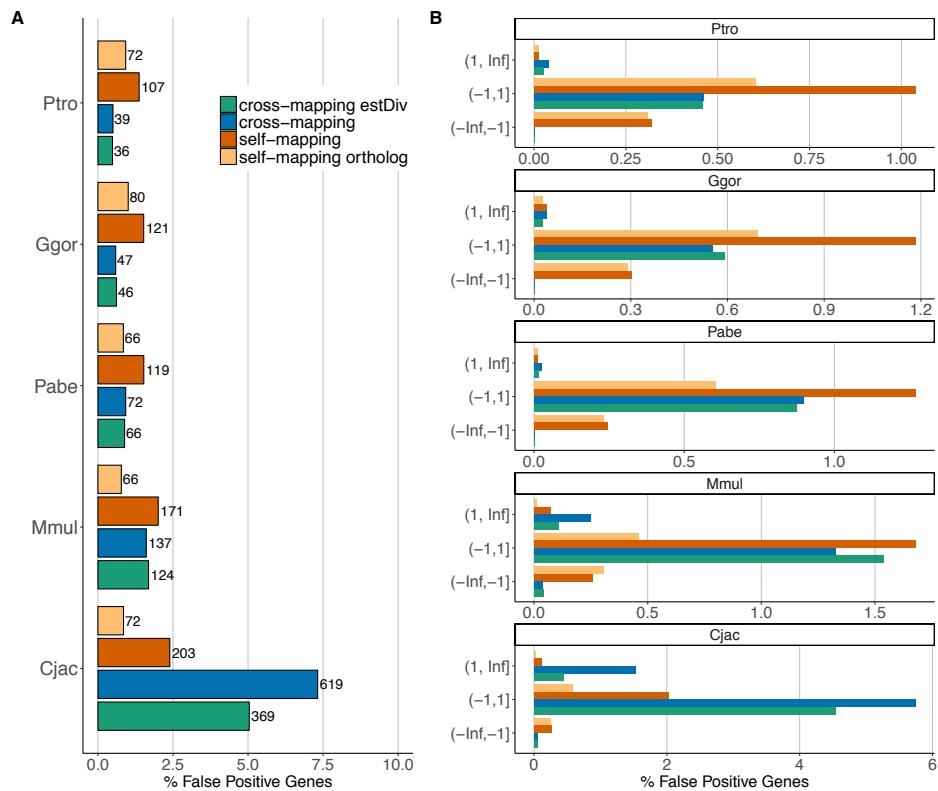


Figure 5. False Positive DE genes for different analysis strategies for Ensembl genomes. This plot is the same as Figure 5, except that here we simulate RNA-seq reads from Ensembl genomes (hg38, Ptro2.1, Ggor3.1, Ppyg2, Mmul8.0.1, Cjac3.2). Thus in contrast to Figure 5, the genomes differ in quality and orthology of genes is inferred.

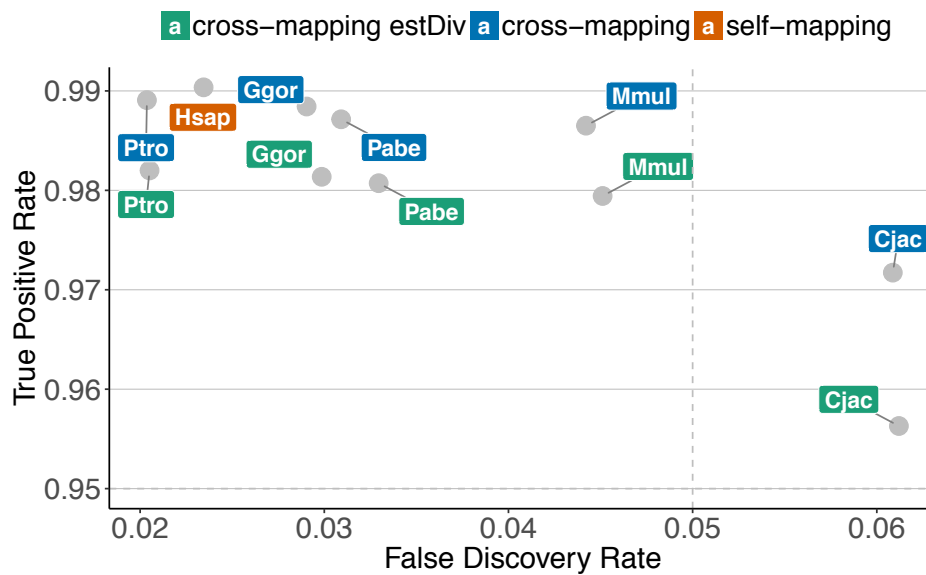


Figure 6. Within-species DE-analysis mapping to a diverged Genome. We plot here true positive rate (TPR) and false discovery rate (FDR) to find condition differences in nh-primates. Reads for the nh-primates primates are mapped to Human genome (cross-mapping). The colors indicate if the read counts per gene are corrected for the divergence to Human (green) or not (blue).

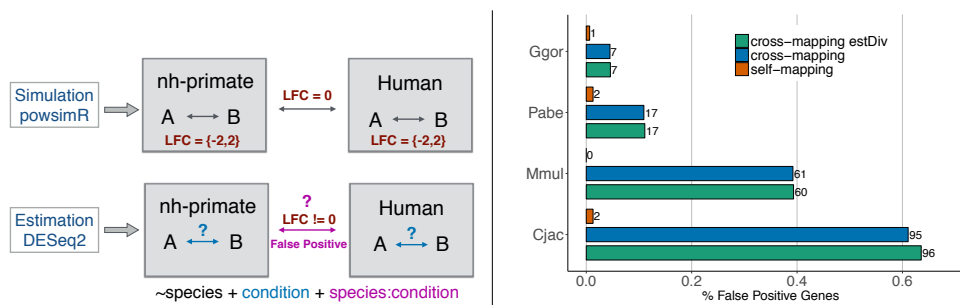


Figure 7. Relative expression measures. A) Schematic of relative expression study design. Upper panel shows simulation design and the lower panel shows estimation. For every pair of nh-primate (5 primates) with human we simulated two group RNA-seq with symmetrically differentially expressed genes with log2 fold-change of -2 and 2 each for 5% genes. There is no differential expression simulated between species. We then used DESeq2 to find DE-genes between conditions in each primate and condition difference between species at 5% FDR. B) The barplot here shows % of false positive genes detected in condition differences between species. The colors indicate counting strategies (self-mapping - Orange, cross-mapping - blue and cross-mapping with divergence correction - green).

References

1. Wolfgang Enard, Philipp Khaitovich, Joachim Klose, Sebastian Zöllner, Florian Heissig, Patrick Givalisco, Kay Nieselt-Struwe, Elaine Muchmore, Ajit Varki, Rivka Ravid, Gaby M Doxiadis, Ronald E Bontrop, and Svante Pääbo. Intra- and interspecific variation in primate gene expression patterns. *Science*, 296(5566):340–343, April 2002.
2. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, September 2005.
3. Philipp Khaitovich, Ines Hellmann, Wolfgang Enard, Katja Nowick, Marcus Leinweber, Henriette Franz, Gunter Weiss, Michael Lachmann, and Svante Pääbo. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, 309(5742):1850–1854, September 2005.
4. Michael Dannemann, Anna Lorenc, Ines Hellmann, Philipp Khaitovich, and Michael Lachmann. The effects of probe binding affinity differences on gene expression measurements and how to deal with them. *Bioinformatics*, 25(21):2772–2779, November 2009.
5. Ashlee M Benjamin, Marshall Nichols, Thomas W Burke, Geoffrey S Ginsburg, and Joseph E Lucas. Comparing reference-based RNA-Seq mapping methods for non-human primate data. *BMC Genomics*, 15(1):570, December 2014.
6. Ran Blekhman, John C Marioni, Paul Zumbo, Matthew Stephens, and Yoav Gilad. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.*, 20(2):180–189, February 2010.
7. David Brawand, Magali Soumillon, Anamaria Necșulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, October 2011.
8. Ying Zhu, Mingfeng Li, André Mm Sousa, and Nenad Sestan. XSanno: a framework for building ortholog models in cross-species transcriptome comparisons. *BMC Genomics*, 15:343, May 2014.
9. Stephanie Wunderlich, Martin Kircher, Beate Vieth, Alexandra Haase, Sylvia Merkert, Jennifer Beier, Gudrun Göhring, Silke Glage, Axel Schambach, Eliza C Curnow, Svante Pääbo, Ulrich Martin, and Wolfgang Enard. Primate iPSC cells as tools for evolutionary analyses. *Stem Cell Res.*, 12(3):622–629, May 2014.
10. Benedict Paten, Javier Herrero, Kathryn Beal, Stephen Fitzgerald, and Ewan Birney. Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, 18(11):1814–1828, November 2008.
11. Albert J Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19(2):327–335, February 2009.
12. Thomas Derrien, Jordi Estellé, Santiago Marco Sola, David G Knowles, Emanuele Raineri, Roderic Guigó, and Paolo Ribeca. Fast computation and applications of genome mappability. *PLoS One*, 7(1):e30377, January 2012.
13. Robert C Edgar, George Asimenos, Serafim Batzoglou, and Arend Sidow. Evolver. *Website* <http://www.drive5.com/evolver>, 2009.
14. Hervé Philippe, Henner Brinkmann, Dennis V Lavrov, D Timothy J Littlewood, Michael Manuel, Gert Wörheide, and Denis Baurain. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.*, 9(3):e1000602, March 2011.

-
15. Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomart. *Nat. Protoc.*, 4(8):1184–1191, July 2009.
 16. Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.*, 40(20):10073–10083, November 2012.
 17. Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.
 18. Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, 2014.
 19. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol.*, 32(9):903–914, September 2014.
 20. Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47, April 2015.
 21. Erica V Todd, Michael A Black, and Neil J Gemmell. The power and promise of RNA-seq in ecology and evolution. *Mol. Ecol.*, 25(6):1224–1241, March 2016.
 22. Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.*, 6:25533, May 2016.
 23. W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, June 2002.
 24. Bronwen L Aken, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E Hunt, Sophie H Janacek, Thomas Juettemann, Stephen Keenan, Matthew R Laird, Ilias Lavidas, Thomas Maurel, William McLaren, Benjamin Moore, Daniel N Murphy, Rishi Nag, Victoria Newman, Michael Nuhn, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Daniel Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Steven P Wilder, Amonida Zadissa, Myrto Kostadima, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Daniel M Staines, Stephen J Trevanion, Fiona Cunningham, Andrew Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2017. *Nucleic Acids Res.*, 45(D1):D635–D642, January 2017.
 25. Robert M Kuhn, David Haussler, and W James Kent. The UCSC genome browser and associated tools. *Brief. Bioinform.*, 14(2):144–161, March 2013.
 26. Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.
 27. Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, April 2014.
 28. M Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16(2):111–120, December 1980.
 29. Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: Power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, July 2017.
 30. Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29(4):1165–1188, 2001.

Strategies for RNA-seq differential expression analysis for closely
related species

SUPPLEMENTARY INFORMATION

by

Swati Parekh¹, Beate Vieth¹, Christoph Ziegenhain¹, Wolfgang Enard¹ and Ines
Hellmann^{1,*}

¹Anthropology & Human Genomics, Department of Biology II,
Ludwig-Maximilians University, Munich, Germany

*Corresponding author

1 Supplementary Figures

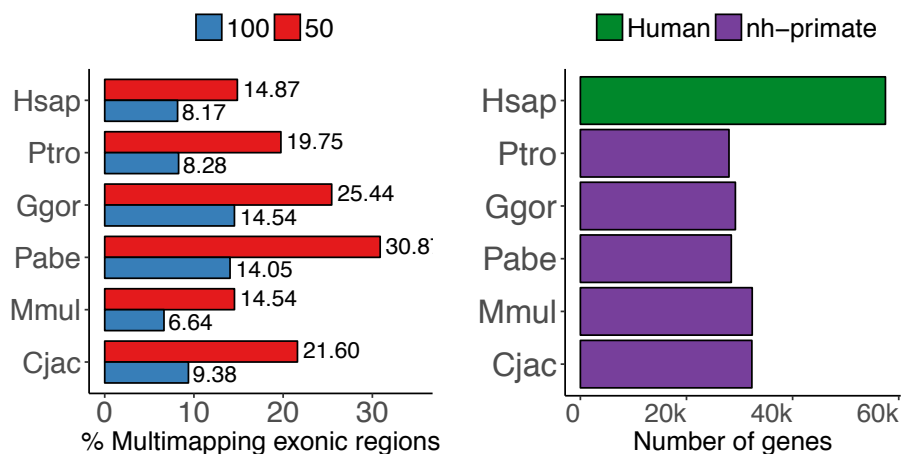


Figure S1: Varying quality of genomic resources. A) For 6 primates ensemble genomes, we plot fraction of exonic regions having potential to map at more than one loci on the genome. The mappability score is calculated for the whole genome using GEM-mappability for 50bp (light blue) and 100bp (dark blue) read length with 4% mismatches allowed. B) We plot total number of genes annotated in each primate derived from Ensembl database. The assembly and annotation versions are given in the first panel of Figure 1. The primate names are abbreviated from their scientific names (Cjac = Marmoset, Mmul = Rhesus macaque, Pabe = Orangutan, Ggor = Gorilla, Ptro = Chimpanzee, Hsap = Human).

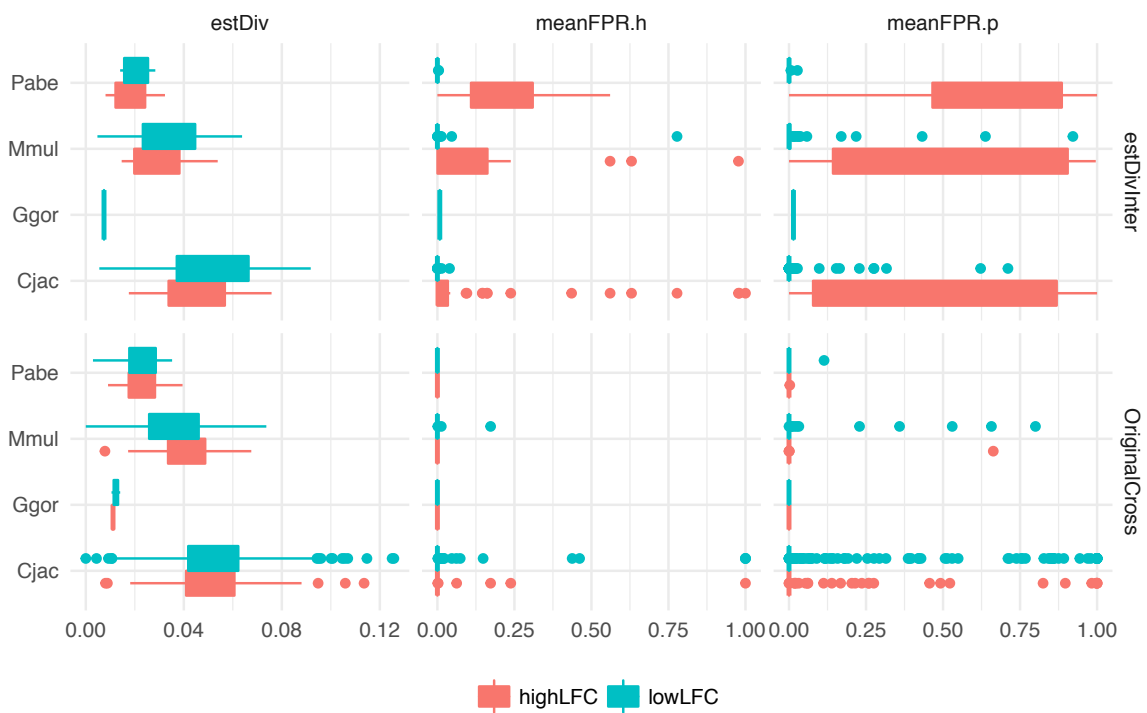


Figure S2: Estimated divergence and the fraction of falsely mapped reads for false positive DE-genes. Panels from left to right represent estimated divergence, mean FPR of mapping in Human and mean FPR of mapping in nh-primates. The data points are falsely detected DE-genes. Lower panel shows DGE results from the model with only cross-mapped read counts while the upper panel shows DGE results where cross-mapped read counts are corrected for the divergence.

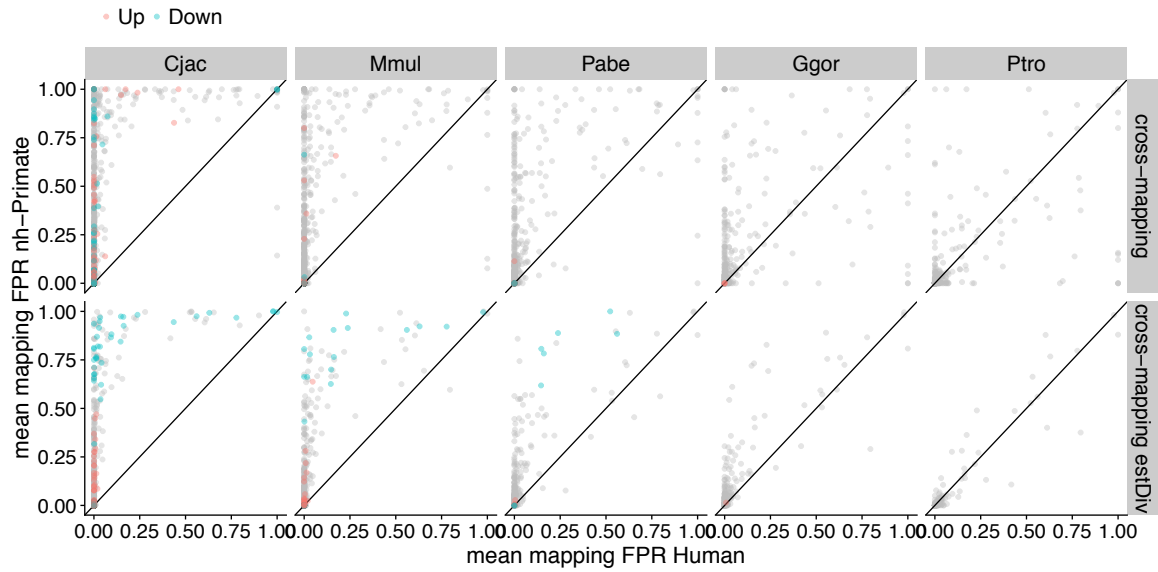


Figure S3: Average False Positive Mapping Rates. We plot the average fraction of falsely mapped genes for cross-mapping of primate and human reads to the *in silico* Human genome. Dots in grey represent genes that are not detected as DE. Blue are detected as down-regulated in the Human, red are up-regulated in the Human relative to the primate expression. False mapping fraction for human and primate genes does not correlate and it is higher for the primate reads. Upper panel shows DGE results from the model with only cross-mapped read counts while the lower panel shows DGE results where cross-mapped read counts are corrected for the divergence.

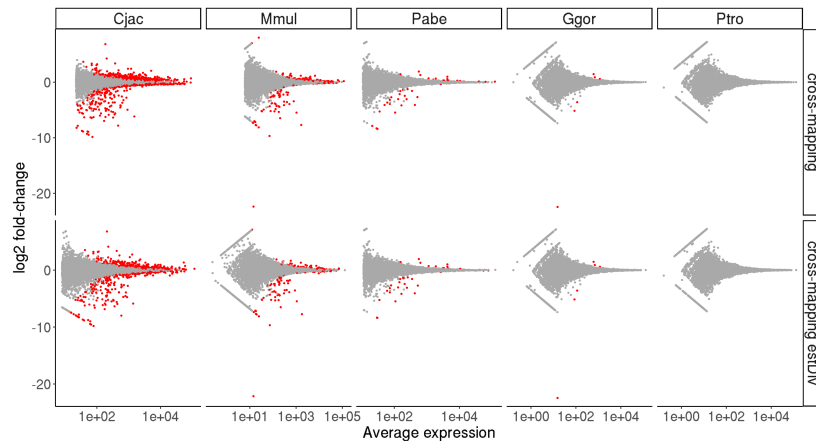


Figure S4: MA plots of differential gene expression analysis from simulated genomes. We perform DGE using DESeq2 and plot here \log_2 fold-change vs Average expression for each pairwise comparison between nh-primate and human. Red dots are significantly differentially expressed genes. Panels from left to right are arranged by evolutionary distance of nh-primate to human. Upper panel shows DGE results from the model with only cross mapped read counts while the lower panel shows DGE results with fitted counts derived from the model with estimated divergence added as an interaction term with cross mapped counts.

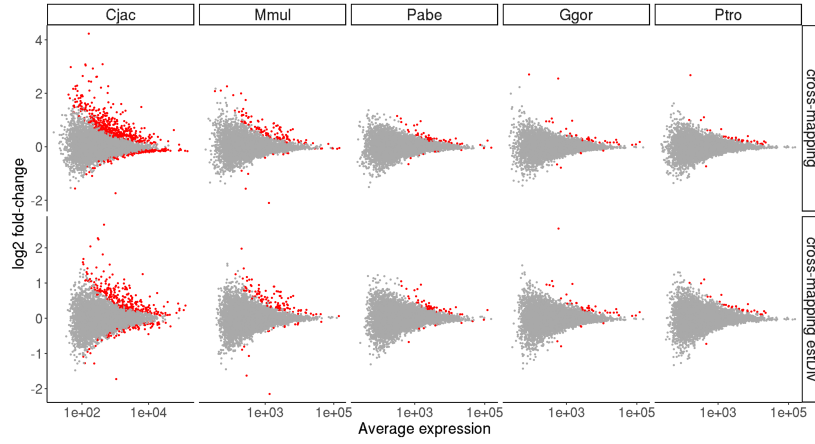


Figure S5: MA plots of differential gene expression analysis from real genomes. This is the same plot as "Supplementary Figure S4" except that the read counts are simulated from Ensembl genomes.

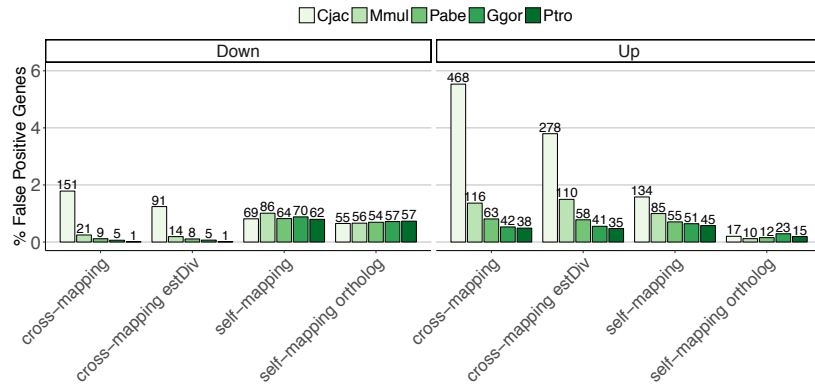


Figure S6: False Positive DE genes for different analysis strategies classified by the directionality of expression differences. Bar plots show the number of False Positive differentially expressed genes between nh-primate and Human. We classify the FPR according to the estimated \log_2 -fold change in different analysis strategies. The \log_2 -fold changes in the left panel indicate lower expression and the right panel higher expression in the Human (reference) as compared to the nh-primate.

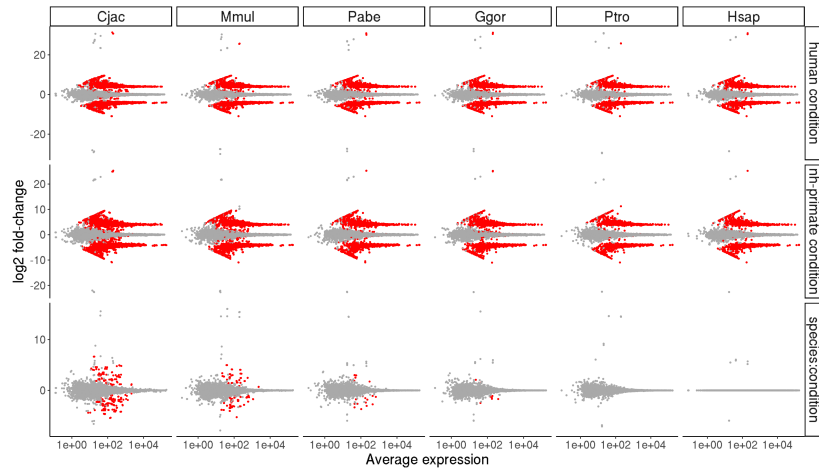


Figure S7: MA plot of relative expression differences. We perform DGE using DESeq2 as described in Figure 7 and plot here \log_2 fold-change vs Average expression for three DE tests in each pairwise comparison between nh-primate and human. Red dots are significantly differentially expressed genes. Panels from left to right are arranged by evolutionary distance of nh-primate to human. Upper panel shows DGE results for between condition differences in Human, middle panel for condition differences in nh-primate and the lower panel for condition differences relative to species.

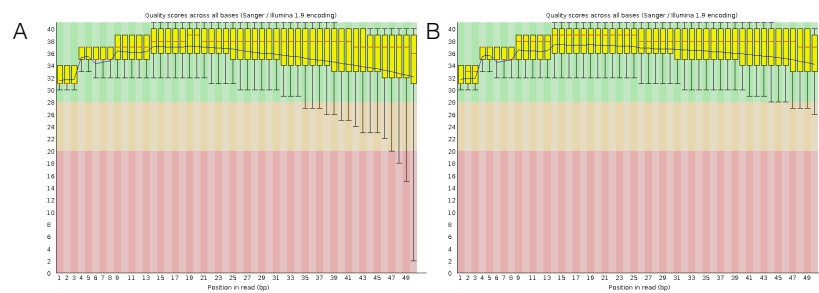


Figure S8: Per-base sequence quality of real and simulated data. The error models for RNA-seq reads simulation in flux-simulator were built using Smart-seq2 data generated previously. We show here the per-base sequence quality plots generated by fastqc of A) UHRR Smart-seq2 data and B) Human 50SE RNA-seq data simulated by flux-simulator.

2 Supplementary Tables

Node	term	estimate	std.error	statistic	p.value	Genome
Cjac	Intercept	1.07E-01	4.06E-03	2.64E+01	1.65E-140	Real
Cjac	log10(crossmapping+1)	9.43E-01	1.67E-03	5.66E+02	0.00E+00	Real
Cjac	log10(crossmapping+1):estDiv	8.39E-01	1.77E-02	4.75E+01	0.00E+00	Real
Mmul	Intercept	4.58E-02	2.93E-03	1.56E+01	1.23E-51	Real
Mmul	log10(crossmapping+1)	9.83E-01	1.18E-03	8.33E+02	0.00E+00	Real
Mmul	log10(crossmapping+1):estDiv	2.89E-01	1.98E-02	1.46E+01	1.20E-44	Real
Pabe	Intercept	2.38E-02	2.41E-03	9.86E+00	2.16E-19	Real
Pabe	log10(crossmapping+1)	9.95E-01	9.43E-04	1.06E+03	0.00E+00	Real
Pabe	log10(crossmapping+1):estDiv	1.06E-01	2.70E-02	3.95E+00	1.34E-04	Real
Ggor	Intercept	2.23E-02	2.35E-03	9.51E+00	3.81E-20	Real
Ggor	log10(crossmapping+1)	9.96E-01	9.02E-04	1.10E+03	0.00E+00	Real
Ggor	log10(crossmapping+1):estDiv	-3.70E-05	4.71E-02	-1.11E-03	8.53E-01	Real
Ptro	Intercept	1.59E-02	2.00E-03	7.96E+00	6.16E-13	Real
Ptro	log10(crossmapping+1)	9.96E-01	7.58E-04	1.31E+03	0.00E+00	Real
Ptro	log10(crossmapping+1):estDiv	1.01E-01	4.57E-02	2.22E+00	4.16E-02	Real
Hsap	Intercept	2.15E-02	2.42E-03	8.81E+00	2.09E-14	Real
Hsap	log10(crossmapping+1)	9.95E-01	8.12E-04	1.23E+03	0.00E+00	Real
Hsap	log10(crossmapping+1):estDiv	-3.13E+00	5.84E-01	-5.31E+00	1.16E-06	Real
Cjac	Intercept	1.24E-01	5.77E-03	2.15E+01	8.44E-85	Simulated
Cjac	log10(crossmapping+1)	8.74E-01	5.17E-03	1.69E+02	0.00E+00	Simulated
Cjac	log10(crossmapping+1):estDiv	2.20E+00	8.71E-02	2.53E+01	5.06E-130	Simulated
Mmul	Intercept	1.82E-02	3.64E-03	4.99E+00	1.73E-06	Simulated
Mmul	log10(crossmapping+1)	9.68E-01	3.07E-03	3.16E+02	0.00E+00	Simulated
Mmul	log10(crossmapping+1):estDiv	9.37E-01	7.09E-02	1.32E+01	2.59E-37	Simulated
Pabe	Intercept	-4.67E-03	2.52E-03	-1.86E+00	1.14E-01	Simulated
Pabe	log10(crossmapping+1)	9.93E-01	1.99E-03	5.01E+02	0.00E+00	Simulated
Pabe	log10(crossmapping+1):estDiv	4.95E-01	7.80E-02	6.39E+00	1.09E-08	Simulated
Ggor	Intercept	-5.76E-03	1.74E-03	-3.31E+00	2.87E-03	Simulated
Ggor	log10(crossmapping+1)	9.98E-01	1.35E-03	7.45E+02	0.00E+00	Simulated
Ggor	log10(crossmapping+1):estDiv	3.60E-01	1.03E-01	3.51E+00	8.87E-04	Simulated
Ptro	Intercept	-2.09E-03	9.81E-04	-2.12E+00	9.15E-02	Simulated
Ptro	log10(crossmapping+1)	1.00E+00	5.24E-04	1.93E+03	0.00E+00	Simulated
Ptro	log10(crossmapping+1):estDiv	3.03E-02	4.76E-01	3.42E-02	6.74E-01	Simulated
Hsap	Intercept	-2.82E-03	1.07E-03	-2.61E+00	2.88E-02	Simulated
Hsap	log10(crossmapping+1)	1.00E+00	5.76E-04	1.76E+03	0.00E+00	Simulated
Hsap	log10(crossmapping+1):estDiv	4.76E-01	5.16E-01	9.04E-01	3.88E-01	Simulated

Table S1: Model estimates of log-linear fit between simulated counts (sim) and cross-mapping (cross) counts with estimated divergence(div) as an interaction term.

Discussion

Impact of amplification noise in quantitative RNA-seq

Recent advances in sequencing technology have created incredible potential for biology and biomedicine. RNA-seq has become a standard method to profile gene expression in various applications (GTEx Consortium et al. 2017; Hrdlickova, Toloue, and Tian 2017). This emerging technology also generates new challenges for computational analyses. For example, to quantify gene expression levels by RNA-seq one needs to estimate the number of transcripts per gene in a sample from the number of sequencing reads per gene. This requires read numbers to be proportional to transcript numbers. However, PCR amplification is a necessary step for essentially all RNA-seq protocols and hence many reads can be generated from a single transcript. The noise introduced by PCR amplification can then reduce the accuracy and precision of transcript quantification by RNA-seq. This is especially relevant as when amplifying the small amounts of RNA from single cells (Stegle, Teichmann, and Marioni 2015).

It is still unclear how to treat read duplicates in RNA-seq data and how much do read duplicates impact the quantification. Here I perform such an analysis using available benchmark datasets as well as data that were specifically generated for this study generated by three major library preparation methods TruSeq (SEQC/MAQC-III Consortium 2014), Smart-seq2 (Picelli et al. 2013) and UMI-seq (Soumillon et al. 2014). These protocols mainly differ by the fragmentation method, amount of starting material, number PCR cycles performed and sample pooling before or after amplification.

Computationally, duplicates are identified based on their 5' mapping position, read orientation and sequence identity. On the contrary, incorporating unique molecular

identifiers (UMIs) (Kivioja et al. 2011; Shiroguchi et al. 2012) to the cDNA molecules before amplification, as in UMI-seq, makes it possible to track original molecules.

Various methods are available to flag the duplicate reads (<http://broadinstitute.github.io/picard>, (Li et al. 2009) to account for bias in SNP calling or peak calling (DePristo et al. 2011; Y. Chen et al. 2012; Li et al. 2009). However, it is difficult to computationally distinguish between natural duplicates and PCR duplicates in RNA-seq for various reasons: 1) fragmentation can have a preference for cutting sites that produce reads that look alike 2) some methods have a full length cDNA amplification step where reads from duplicated cDNA molecules can have different 5' mapping positions 3) highly transcribed genes have higher chances of sampling fragments starting at the same site and 4) with increasing sequencing depth it is more likely to get the same fragments by chance and hence it increases the probability of reads having the same 5' mapping.

Previous studies have also observed the high fraction of natural duplicates in RNA-seq data (Lappalainen et al. 2013; Bansal 2017). Some tools have attempted to tackle such problems by smoothing the read coverage such as eXpress (Roberts and Pachter 2013) or by modelling probability of natural duplicates and adjust the observed number of PCR duplicates (Mezlini et al. 2013 and Baumann et. al. 2013). However, this approach is not applicable to situations in which systematic over-estimation of read counts on a large fraction of genes exists. Bansal 2017 (Bansal 2017) proposed a model for estimating the rate of PCR duplicates accounting for natural duplicates derived from heterozygous variant sites. This approach becomes impractical for haploid genomes and genomes with fewer heterozygous variants. Furthermore, we observed that the duplicates stemming from the full length cDNA amplification step in Smart-seq2 can not be identified by their mapping positions. This notion is corroborated by the observation that the standard curve between our measure of expression, calculated to include duplicates, fits the initial concentrations of the ERCC spike-ins better than those excluding duplicates. Clearly, computational methods

of duplicate removal cannot accurately identify PCR duplicates, but the next question is how much do duplicates affect relative quantification.

We simulated differential expression between two groups using empirically estimating mean and dispersion parameters for each dataset with and without duplicate removal using an adapted simulation framework as described in PROPER (Wu, Wang, and Wu 2015), later developed into an independent package, *powsimR* (Vieth et al. 2017). Confirming the finding that computational removal of duplicate reads is inefficient, our results show that for the UMI-seq dataset where samples are pooled before amplification, the TPR is much higher even with more PCR cycles and the FDR is well controlled when duplicates are removed by UMIs.

Building on the fact that the TPR to detect differential expression is negatively correlated with the number of PCR cycles used, amplification noise can be more severe in scRNA-seq experiments. In our scRNA-seq methods benchmarking paper (Ziegenhain et al. 2017), we identify upto ~95% of duplicated reads per cell from the UMI-based protocol. As the variance of sampling reads for gene expression measurements is dependent on the mean following the Poisson distribution in RNA-seq (Pachter 2011), additional amplification noise can be expressed as Extra-Poisson Variability (Ziegenhain et al. 2017). Our results indicate higher amplification noise for full length methods or when UMIs are not collapsed to count unique molecule per gene. This data certainly adds to the growing evidence that confirm UMIs are indeed capable of removing amplification noise as discussed in a technology feature on PCR duplicates (Marx 2017).

Compared to bulk RNA-seq, scRNA-seq data is sparser, and apart from mean and dispersion, it is necessary to take into account dropout rates(p_0) as a parameter when simulating differential expression for scRNA-seq data. To fulfill this, *powsimR*, a power simulation framework that implements most of the widely used scRNA-seq filtering, normalisation and differential testing methods that capture specific characteristics of

single-cell RNA-seq data generated by different methods, was developed (Vieth et al. 2017). Using *powsimR*, we show that for scRNA-seq data, the gain in power to detect differential expression is much higher when UMI-based de-duplication is performed compared to that of in bulk RNA-seq (Parekh et al. 2016; Ziegenhain et al. 2017).

In summary, we clearly find that computational removal of read duplicates is not recommended, because many of the read duplicates are due to sampling of independent molecules (natural duplicates) and not PCR-duplicates. Moreover, the amplification rate is variable across samples, and UMI-based early pooling methods lead to an appreciable increase in power to detect differential expression especially for low starting material.

Identifying and addressing computational challenges in single-cell RNA-seq data analysis

Single-cell RNA sequencing (scRNA-seq) is a transformative technology that is rapidly deepening our understanding of biology (Wagner, Regev, and Yosef 2016). Since the first attempt of sequencing a whole transcriptome from a single cell (Tang et al. 2009), there has been a boom in the development of scRNA-seq protocols for various applications (Kolodziejczyk, Kim, Svensson, et al. 2015; Shapiro, Biezuner, and Linnarsson 2013; Wagner, Regev, and Yosef 2016). During the handling of small starting amounts, it is very important to minimize the loss of material and optimize the procedure to reduce technical variation in the data. It is necessary to assess the performance of these protocols and point out the advantages and disadvantages of every method to make an informed choice of a protocol to use for an experiment. We analysed data from mouse embryonic stem cells (mESCs) generated with 6 major scRNA-seq protocols (Smart-seq/C1 (Pollen et al. 2014), Smart-seq2 (Picelli et al. 2013), SCRB-seq (Soumillon et al. 2014), CEL-seq2/C1

(Hashimshony et al. 2016), MARS-seq (Jaitin et al. 2014) and Drop-seq (Macosko et al. 2015) to evaluate the performance of each protocol.

The first measure of comparison used is the sensitivity of each protocol, i.e. the capacity to capture transcriptome as fully as possible. Here, one of the limiting step is thought to be reverse transcription (Picelli et al. 2013), with an estimated efficiency to capture mRNA molecules of 10-50% (Islam et al. 2014; Grün, Kester, and van Oudenaarden 2014). We find that out of the two full length methods, Smart-seq2 is much more sensitive compared to Smart-seq/C1, whereas among the four UMI-based methods, SCR-seq and CEL-seq2/C1 are significantly more sensitive compared to Drop-seq and MARS-seq. This finding is also largely confirmed by the other recently publish scRNA-seq protocol comparison study (Svensson et al. 2017). In Svensson et al. study, the comparison across 15 different protocols is performed mainly based on the detection probability of ERCC spike-in molecules. ERCC spike-in molecules are known to have different physical properties compared to those of endogenous transcripts (Risso et al. 2014). Moreover, the datasets in comparison exhibit different cell types generated in different labs at hugely varying library sizes. Despite having inherent variance coming from different sources, the data generated from their own laboratory show that Smart-seq2 is the most sensitive method. Moreover, for some applications of RNA-seq like discovery of splice variants, fusion genes or variant calling, full transcript length coverage is important. We observe that the Smart-seq2 protocol currently sequences full-length transcripts with the most even coverage. The second metric for comparison is accuracy, which describes the capacity of a protocol to represent the mRNA abundance levels in the expression estimates. We used ERCC spike-ins (External RNA Controls Consortium 2005; Lichun Jiang et al. 2011; Paul et al. 2016) to compare the accuracy of six protocols. We observe that although Smart-seq2 demonstrates the highest accuracy, all the other protocols also achieve nearly the same accuracy level at least for highly expressed genes. Smart-seq2 was also found to be the

most sensitive protocol thus it has an advantage of measuring better correlation at broader range of expression level. The next measure is precision, which is defined as the variability of measured gene expression estimates around its mean expression. As discussed in the previous chapter, we have shown that incorporation of UMIs makes it possible to distinguish duplicates from original molecules and thus remove amplification noise. Obviously, the UMI-based methods show better precision compared to full length methods, followed closely by Smart-seq2 and other UMI-based methods. Our last measure of comparison is the power to detect differential expression in a two-group comparison at different sample sizes using *powsimR* (Vieth et al. 2017). We observe that SCRB-seq shows the highest power at a sequencing depth of 1 million reads, most likely due to high sensitivity and low amplification noise with the use of UMIs. We also calculate the cost efficiency in terms of power. As expected, the introduction of cell barcodes as early in the library preparation process as possible reduces cost and labour time by pooling many cells. Nevertheless, when applying this high level of multiplexing, care must be taken to avoid the assignment of reads to the wrong barcode, such that transcriptomes remain at single-cell purity.

Practically, it is not economical to sequence all the cells to saturation but the protocol that captures the most transcripts at a certain depth is considered to be the most cost efficient. We performed saturation analysis by downsampling all the libraries at different depths to estimate the optimal depth to capture the most information. Apart from saturation analysis, downsampling is suggested especially in scRNA-seq data to reduce complications at normalisation (Vallejos et al. 2017; Evans, Hardin, and Stoebel 2017). Having said this, we have implemented adaptive downsampling of over represented libraries within three median absolute deviations suggested elsewhere (Grün and van Oudenaarden 2015) in *zUMIs* (Parekh et al. 2017). In summary, while no single best scRNA-seq method exists, different tradeoffs between sensitivity, accuracy, precision, throughput and costs should be

considered when choosing the most appropriate method for the research question at hand and new developments should be independently benchmarked for these parameters.

When considering higher throughput experiments, library preparation methods using pooled samples with integrated cell barcodes in the sequencing construct are on the rise. For the well-based methods the barcodes are known and designed to have maximal error distances, e.g. Illumina indices i5 & i7. Such barcodes are fairly straightforward to demultiplex and some methods even provide a probabilistic assignment also considering sequence quality thus allowing for an unbiased and rigorous quality assessment (Renaud et al. 2015; Galanti, Shasha, and Gunsalus 2017). It is known that the sequencing quality deteriorates with more cycles (Buermans and den Dunnen 2014; Laehnemann, Borkhardt, and McHardy 2016). Hence, it is recommended to sequence the barcode read first since barcode assignment is shorter and more sensitive to base call errors compared to mapping of cDNA reads to a reference. In droplet-based methods, both the barcode sequences and the total number of cells are unknown. This makes cell identification more difficult because the identification of one cell is no longer independent from the identification of other cells in the mix. Removing barcodes with low sequencing quality will reduce spurious associations and is thus our recommended first step. Evidently, barcodes associated with intact cells have significantly more reads compared to dead or broken cells or debris. Consequently, as long as the sequenceable RNA-content does not vary by orders of magnitude, a read count cut-off should be sufficient to distinguish cells from debris. We expect a bimodal distribution, where the first peak probably consists of all the spurious barcodes and the second peak has the barcodes of the viable cells (Macosko et al. 2015). In zUMIs (Parekh et al. 2017), we have implemented the possibility to choose barcodes based on a given list of sequences, an expected number of cells and an automatic barcode selection method. We fit a k-dimensional multivariate normal distribution to the number of reads per barcode and

choose only the last peak with the largest mean to automatically select the barcodes with the most number of associated reads.

Apart from cellular barcodes, every transcript molecule is also tagged with unique molecular barcodes (UMI). In principle, every UMI of a gene represents a transcript molecule but due to PCR/sequencing errors some barcodes can be assigned as a different UMI molecule resulting into over estimation of the transcript molecules. An evaluation of how best to account for such errors in the barcodes is currently lacking. We have observed that normally a sequence quality based cutoff can be applied to filter molecular barcodes as well (Parekh et al. 2017). We compared our phred score based filtering method with directional adjacency based UMI collapsing method implemented in UMI-tools (Smith, Heger, and Sudbery 2017). Indeed, the power to detect differential expression increases after filtering but there is no significant difference in power between the two UMI collapsing methods. The theory behind this observation is that the spurious UMIs stem from sequencing errors and a *a priori* low base call quality filter of such reads reduces the chances of forming such nodes with one edit distance.

To overcome the issue of utilizing frozen or preserved tissue dissociations, plate-based and droplet-based methods are being developed to sequence from single nuclei (Habib et al. 2017, 2016; Lake et al. 2016; Lacar et al. 2016). The data from single nuclei contain a significant fraction of nascent mRNAs, which means a lot of unspliced introns are present. Moreover, recently, a new method has been developed that measures RNA velocity by measuring the abundance of unspliced and spliced RNA from scRNA-seq data (La Manno et al. 2017). This method also relies on the use of intronic reads in the data. We show that using intronic reads in addition to exonic reads achieves an increased resolution of clusters and this is probably helpful to increase the sensitivity and precision of scRNA-seq quantification. *zUMIs* was the first pipeline that produced count matrices where UMIs are collapsed from exonic or intronic specific reads as well as exonic plus intronic reads.

For the ease of analysis, several processing tools are available to carry out one or several of the processing steps to generate an expression profile from the raw data (Macosko et al. 2015; Guo et al. 2015; Ilicic et al. 2016; Tian et al. 2017; Smith, Heger, and Sudbery 2017; Mangul et al. 2017; Petukhov et al. 2017; Alonso et al. 2017). However, current pipelines that process UMI-based RNA-seq data are normally designed for a specific protocol (Macosko et al. 2015), not open source (Macosko et al. 2015), published as individual modules for each processing step (Smith, Heger, and Sudbery 2017; Petukhov et al. 2017), uses only transcriptome mapping (Svensson et al. 2017; Hashimshony et al. 2016). zUMIs is the only pipeline that generates gene expression profiles from raw data in a single run and at the same time has features like adaptive downsampling, intronic reads counting, automatic cell barcode selection and it is compatible with all the UMI-based protocols.

Optimising cross species differential expression analysis

With dropping sequencing costs, RNA-seq is now becoming a common method to study gene expression dynamics in various applications (Z. Wang, Gerstein, and Snyder 2009). Today, quantitative transcriptomic technologies provide a global snapshot of transcription under certain conditions which can be directly related to phenotypic information relevant for understanding evolutionary dynamics across diverged species (Brawand et al. 2011; Warnefors and Kaessmann 2013; Romero, Ruvinsky, and Gilad 2012). For the precise estimates of expression levels, accurate placement of reads on the genome and assignment to gene models is important. However, for non-model organisms, the major drawback is the lack of good quality genomic resources. For certain species no genome reference and annotations exist, while for other species they are poorly resolved. For inter-species comparison of gene expression profiles, variation in mappability and annotations could affect the power to detect differential expression between species.

Here, we investigate on the impact of underlying differences in the quality and availability of the genomic resources on differential gene expression within and between species. Using a well resolved closely related species as a reference could circumvent this issue. However, available RNA-seq mappers (Baruzzo et al. 2016) assume low sequence divergence between the subject and query sequences and can not handle sequence divergence above ~10%. Therefore, this mapping strategy brings another question of deconvoluting real biological signal from mapping bias due to sequence divergence. Since every read can be independently assessed for its mapping efficiency, mapping is one of the most amenable process to study by computational simulations.

In order to assess the impact of mapping bias between diverged species independent of assembly and annotation errors, we simulated whole genome sequence evolution across a 6 primate phylogenetic tree (Marmoset, Macaque, Orangutan, Gorilla, Chimpanzee and Human) without any annotation errors. Using these, we simulated RNA-seq expression profiles under no biological variance in orthologous genes across species to quantify False Discovery Rate (FDR) to detect expression differences between species due to mapping bias. Such RNA-seq profiles were also simulated from the actual available primate genomes and gene models obtained from the Ensembl database to directly compare biases introduced by variable quality in the reference. Together, this allows us to assess the “cross-mapping” strategy, where a common high quality genome sequence and gene models are used for mapping diverged species.

By simulating RNA-seq profiles with different read lengths (50bp and 100bp) and sequencing layouts (single- or paired-end), we could show that while longer reads increase the sensitivity, sequencing layout does not have high impact on mapping. Unsurprisingly, our results show that the mapping efficiency is negatively correlated with the sequence divergence between the species. These results indicate that if mapping bias due to sequence divergence can be corrected at gene level quantification, the issue of uneven

genomic reference quality can be resolved by cross-mapping. We developed a method to empirically calculate gene-wise divergence to adjust the abundance levels of genes. We carry out pairwise differential gene expression analysis between each nh-primate and human. In this simulation scenario, any significant difference detected between the orthologous genes of two species is a False Positive call. For highly diverged species, FPR reduced by ~40% when sequence divergence correction was used to adjust the abundance levels of each gene. As mentioned above, available RNA-seq mappers can not handle divergence levels above 10% which is believed to be the reason for the remaining false positive genes. We also show that RNA-seq simulations from Ensembl genomes produce lower FPR for cross-mapping compared to the estimates derived from mapping to their own genomes (self-mapping). This shows that the noise introduced by poor quality annotations in closely related species is bigger than mapping bias due to sequence divergence. These biases can be avoided by restricting the transcriptional units to only orthologous regions annotated in both the species in comparison. Thus, correcting for gene length differences as proposed previously can be used with caution for distantly related species (Zhu et al. 2014; Wunderlich et al. 2014; Brawand et al. 2011).

Generally, biologically interesting questions are not restricted to the direct comparison of species but also to detect the relative differences between conditions across species (Enard et al. 2002; Brawand et al. 2011; Khaitovich et al. 2006). To evaluate the impact of cross-mapping in detecting relative changes of certain condition between species, we simulated 10% symmetrically differentially expressed genes between two conditions in each primate assuming no variance in conditions between species. We observed that the False Positive Rate (FPR) to estimate relative differences between diverged species is correlated with the sequence divergence. Nevertheless, the FPR is as low as 0.6% for Marmoset, 0.4% for Macaque and 0.1% for Orangutan whereas for closely related species (Gorilla and Chimpanzee) the cross-mapping strategy can be used without discernible loss

of information. This simulation framework also enables us to investigate if a high quality genomic reference can be used to detect expression changes within species whose reference genome is not available. Our results show that the differential expression within non-reference species can be detected using the cross-mapping strategy with a very high sensitivity for most of the species with only Marmoset having slightly higher than nominal FDR. This notion is corroborated with the finding that within-species differences could be detected using microarrays from a high quality reference from a closely related species (Oshlack et al. 2007).

To summarize, the evaluated cross-mapping strategy is safe to use for the detection of relative expression changes between conditions across species. Moreover, we can also say that if genomic resources are not available, the closest most resolved species can be used to perform differential expression analysis using RNA-seq within species.

Conclusions and Outlook

In this work I focused on optimising quantitative RNA-seq data analysis for different applications. I addressed the most pressing issues of how to handle read duplicates, identified and addressed challenges pertaining to single cell RNA-seq data processing, provided solutions for cross species differential gene expression by developing and optimising computational strategies. RNA-seq techniques and their applications are evolving at an accelerating rate giving rise to new computational challenges. Here, I contributed to the field by carefully investigating the impact of amplification noise on quantification. I have developed *zUMIs*, an all in one pipeline to address single cell RNA-seq data processing issues, and contributed in the development of *powsimR*, a comprehensive power simulation framework. I also extended our work to optimising RNA-seq for evolutionary studies by systematically identifying the best computational strategy to handle relative quantification of diverged species in an unbiased manner.

The exciting possibilities of quantitative RNA-seq has led to international initiatives like the Genotype-Tissue Expression (GTEx) consortium and the Human Cell Atlas (HCA). These resources set out to provide a comprehensive reference framework for the whole human at tissue and single cell resolution. Additionally, the substantial reduction in the cost of sequencing and the increased development in computational methods has opened doors to improvise genomic resources for many new species. To overcome the issue of loss of RNA molecules at reverse transcription, methods are being developed by Nanopore Sequencing to directly sequence RNA (Garalde et al. 2018); additionally, these methods will also eliminate amplification bias. Such developments will bring a paradigm shift in our understanding of cellular and molecular heterogeneity in a biological system. However, as these methods are still in the early phases of development and are not yet widely used, current RNA-seq quantification methods will be used for some time.

References

- Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merrill, A. Wu, B. Olde, and R. F. Moreno. 1991. "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project." *Science* 252 (5013):1651–56.
- Adey, Andrew, Hilary G. Morrison, Asan, Xu Xun, Jacob O. Kitzman, Emily H. Turner, Bethany Stackhouse, et al. 2010. "Rapid, Low-Input, Low-Bias Construction of Shotgun Fragment Libraries by High-Density in Vitro Transposition." *Genome Biology* 11 (12):R119.
- Alonso, Arnald, Brittany N. Lasseigne, Kelly Williams, Josh Nielsen, Ryne C. Ramaker, Andrew A. Hardigan, Bobbi Johnston, et al. 2017. "aRNApipe: A Balanced, Efficient and Distributed Pipeline for Processing RNA-Seq Data in High-Performance Computing Environments." *Bioinformatics* 33 (11):1727–29.
- Alwine, J. C., D. J. Kemp, and G. R. Stark. 1977. "Method for Detection of Specific RNAs in Agarose Gels by Transfer to Diazobenzyloxymethyl-Paper and Hybridization with DNA Probes." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12):5350–54.
- Anders, S., P. T. Pyl, and W. Huber. 2014. "HTSeq—a Python Framework to Work with High-Throughput Sequencing Data." *Bioinformatics*. Oxford Univ Press. <http://bioinformatics.oxfordjournals.org/content/early/2014/09/25/bioinformatics.btu638.short>.
- Anson, Wilhelm J. 2009. "Next-Generation DNA Sequencing Techniques." *New Biotechnology* 25 (4):195–203.
- Auer, Paul L., and R. W. Doerge. 2010. "Statistical Design and Analysis of RNA Sequencing Data." *Genetics* 185 (2):405–16.
- Azrolan, N., and J. L. Breslow. 1990. "A Solution hybridization/RNase Protection Assay with Riboprobes to Determine Absolute Levels of apoB, A-I, and E mRNA in Human Hepatoma Cell Lines." *Journal of Lipid Research* 31 (6):1141–46.
- Bacher, Rhonda, and Christina Kendziorski. 2016. "Design and Computational Analysis of Single-Cell RNA-Sequencing Experiments." *Genome Biology* 17 (1):63.
- Bagnoli, Johannes W., Christoph Ziegenhain, Aleksandar Janjic, Lucas E. Wange, Beate Vieth, Swati Parekh, Johanna Geuder, Ines Hellmann, and Wolfgang Enard. 2017. "mcSCRIB-Seq: Sensitive and Powerful Single-Cell RNA Sequencing." *bioRxiv*. <https://doi.org/10.1101/188367>.
- Bakken, Trygve E., Jeremy A. Miller, Song-Lin Ding, Susan M. Sunkin, Kimberly A. Smith, Lydia Ng, Aaron Szafer, et al. 2016. "A Comprehensive Transcriptional Map of Primate

- Brain Development.” *Nature* 535 (7612):367–75.
- Bansal, Vikas. 2017. “A Computational Method for Estimating the PCR Duplication Rate in DNA and RNA-Seq Experiments.” *BMC Bioinformatics* 18 (Suppl 3):43.
- Baruzzo, Giacomo, Katharina E. Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A. FitzGerald, and Gregory R. Grant. 2016. “Simulation-Based Comprehensive Benchmarking of RNA-Seq Aligners.” *Nature Methods*, December. Nature Research. <https://doi.org/10.1038/nmeth.4106>.
- Baumann, Douglas D., and Rebecca W. Doerge. 2014. “Robust Adjustment of Sequence Tag Abundance.” *Bioinformatics* 30 (5):601–5.
- Becker-André, M., and K. Hahlbrock. 1989. “Absolute mRNA Quantification Using the Polymerase Chain Reaction (PCR). A Novel Approach by a PCR Aided Transcript Titration Assay (PATTY).” *Nucleic Acids Research* 17 (22):9437–46.
- Benjamin, Ashlee M., Marshall Nichols, Thomas W. Burke, Geoffrey S. Ginsburg, and Joseph E. Lucas. 2014. “Comparing Reference-Based RNA-Seq Mapping Methods for Non-Human Primate Data.” *BMC Genomics* 15 (1). BioMed Central:570.
- Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, et al. 2008. “Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry.” *Nature* 456 (7218):53–59.
- Bioinformatics, Babraham. 2011. “FastQC A Quality Control Tool for High Throughput Sequence Data.” <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. 2011. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Björklund, Åsa K., Marianne Forkel, Simone Picelli, Viktoria Konya, Jakob Theorell, Danielle Friberg, Rickard Sandberg, and Jenny Mjösberg. 2016. “The Heterogeneity of Human CD127+ Innate Lymphoid Cells Revealed by Single-Cell RNA Sequencing.” *Nature Immunology*, February. Nature Publishing Group. <https://doi.org/10.1038/ni.3368>.
- Blekhman, Ran, John C. Marioni, Paul Zumbo, Matthew Stephens, and Yoav Gilad. 2010. “Sex-Specific and Lineage-Specific Alternative Splicing in Primates.” *Genome Research* 20 (2):180–89.
- Brawand, David, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, et al. 2011. “The Evolution of Gene Expression Levels in Mammalian Organs.” *Nature* 478 (7369):343–48.
- Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5):525–27.
- Buermans, H. P. J., and J. T. den Dunnen. 2014. “Next Generation Sequencing Technology: Advances and Applications.” *Biochimica et Biophysica Acta* 1842 (10):1932–41.
- Bullard, James H., Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. 2010. “Evaluation of Statistical Methods for Normalization and Differential Expression in

- mRNA-Seq Experiments.” *BMC Bioinformatics* 11 (February):94.
- Cáceres, Mario, Joel Lachuer, Matthew A. Zapala, John C. Redmond, Lili Kudo, Daniel H. Geschwind, David J. Lockhart, Todd M. Preuss, and Carolee Barlow. 2003. “Elevated Gene Expression Levels Distinguish Human from Non-Human Primate Brains.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (22):13030–35.
- Carroll, Sean B. 2005. “Evolution at Two Levels: On Genes and Form.” *PLoS Biology* 3 (7). Public Library of Science:e245.
- Chelly, Jamel, Jean-Claude Kaplan, Pascal Maire, Sophie Gautron, and Axel Kahn. 1988. “Transcription of the Dystrophin Gene in Human Muscle and Non-Muscle Tissues.” *Nature* 333 (6176). Springer:858–60.
- Chen, Cheng-Yao. 2014. “DNA Polymerases Drive DNA Sequencing-by-Synthesis Technologies: Both Past and Present.” *Frontiers in Microbiology* 5 (June):305.
- Chen, Yiwen, Nicolas Negre, Qunhua Li, Joanna O. Mieczkowska, Matthew Slattery, Tao Liu, Yong Zhang, et al. 2012. “Systematic Evaluation of Factors Influencing CHIP-Seq Fidelity.” *Nature Methods* 9 (6):609–14.
- Choy, Jocelyn Y. H., Priscilla L. S. Boon, Nicolas Bertin, and Melissa J. Fullwood. 2015. “A Resource of Ribosomal RNA-Depleted RNA-Seq Data from Different Normal Adult and Fetal Human Tissues.” *Scientific Data* 2 (November):150063.
- Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szczesniak, et al. 2016. “A Survey of Best Practices for RNA-Seq Data Analysis.” *Genome Biology* 17 (January):13.
- Craig, David W., John V. Pearson, Szabolcs Szelinger, Aswin Sekar, Margot Redman, Jason J. Corneveaux, Traci L. Pawlowski, et al. 2008. “Identification of Genetic Variants Using Bar-Coded Multiplexed Sequencing.” *Nature Methods* 5 (10):887–93.
- Crick, F. 1970. “Central Dogma of Molecular Biology.” *Nature* 227 (5258):561–63.
- Dannemann, Michael, Anna Lorenc, Ines Hellmann, Philipp Khaitovich, and Michael Lachmann. 2009. “The Effects of Probe Binding Affinity Differences on Gene Expression Measurements and How to Deal with Them.” *Bioinformatics* 25 (21):2772–79.
- Davis, Matthew P. A., Stijn van Dongen, Cei Abreu-Goodger, Nenad Bartonicek, and Anton J. Enright. 2013. “Kraken: A Set of Tools for Quality Control and Analysis of High-Throughput Sequence Data.” *Methods* 63 (1):41–49.
- Delgado, M. Dolore, and Javier León. 2006. “Gene Expression Regulation and Cancer.” *Clinical & Translational Oncology: Official Publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico* 8 (11):780–87.
- DeLuca, David S., Joshua Z. Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie

- Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. 2012. "RNA-SeQC: RNA-Seq Metrics for Quality Control and Process Optimization." *Bioinformatics* 28 (11):1530–32.
- Deng, Qiaolin, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. 2014. "Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells." *Science* 343 (6167):193–96.
- DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5):491–98.
- Derr, Alan, Chaoxing Yang, Rapolas Zilionis, Alexey Sergushichev, David M. Blodgett, Sambra Redick, Rita Bortell, et al. 2016. "End Sequence Analysis Toolkit (ESAT) Expands the Extractable Information from Single-Cell RNA-Seq Data." *Genome Research* 26 (10):1397–1410.
- Dijk, Erwin L. van, Yan Jaszczyszyn, and Claude Thermes. 2014. "Library Preparation Methods for next-Generation Sequencing: Tone down the Bias." *Experimental Cell Research* 322 (1):12–20.
- 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.
- Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent. 1999. "Expression Profiling Using cDNA Microarrays." *Nature Genetics* 21 (1 Suppl):10–14.
- Ebbert, Mark T. W., Mark E. Wadsworth, Lyndsay A. Staley, Kaitlyn L. Hoyt, Brandon Pickett, Justin Miller, John Duce, Alzheimer's Disease Neuroimaging Initiative, John S. K. Kauwe, and Perry G. Ridge. 2016. "Evaluating the Necessity of PCR Duplicate Removal from next-Generation Sequencing Data and a Comparison of Approaches." *BMC Bioinformatics* 17 Suppl 7 (July):239.
- Ellis, Richard J., Kenneth D. Bruce, Claire Jenkins, J. Russell Stothard, Lilly Ajarova, Lawrence Mugisha, and Mark E. Viney. 2013. "Comparison of the Distal Gut Microbiota from People and Animals in Africa." *PloS One* 8 (1):e54783.
- Emilsson, Valur, Gudmar Thorleifsson, Bin Zhang, Amy S. Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, et al. 2008. "Genetics of Gene Expression and Its Effect on Disease." *Nature* 452 (7186):423–28.
- Enard, Wolfgang, Philipp Khaitovich, Joachim Klose, Sebastian Zöllner, Florian Heissig, Patrick Giavalisco, Kay Nieselt-Struwe, et al. 2002. "Intra- and Interspecific Variation in Primate Gene Expression Patterns." *Science* 296 (5566):340–43.
- Engström, Pär G., Tamara Steijger, Botond Sipos, Gregory R. Grant, André Kahles, Gunnar Rättsch, Nick Goldman, et al. 2013. "Systematic Evaluation of Spliced Alignment

- Programs for RNA-Seq Data." *Nature Methods* 10 (12):1185–91.
- Evans, Ciaran, Johanna Hardin, and Daniel M. Stoebel. 2017. "Selecting between-Sample RNA-Seq Normalization Methods from the Perspective of Their Assumptions." *Briefings in Bioinformatics*, February. <https://doi.org/10.1093/bib/bbx008>.
- External RNA Controls Consortium. 2005. "Proposed Methods for Testing and Selecting the ERCC External RNA Controls." *BMC Genomics* 6 (November):150.
- Fedurco, Milan, Anthony Romieu, Scott Williams, Isabelle Lawrence, and Gerardo Turcatti. 2006. "BTA, a Novel Reagent for DNA Attachment on Glass and Efficient Generation of Solid-Phase Amplified DNA Colonies." *Nucleic Acids Research* 34 (3):e22.
- Finak, Greg, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, et al. 2015. "MAST: A Flexible Statistical Framework for Assessing Transcriptional Changes and Characterizing Heterogeneity in Single-Cell RNA Sequencing Data." *Genome Biology* 16 (1):1–13.
- Fonseca, Nuno A., Johan Rung, Alvis Brazma, and John C. Marioni. 2012. "Tools for Mapping High-Throughput Sequencing Data." *Bioinformatics* 28 (24):3169–77.
- Fortna, Andrew, Young Kim, Erik MacLaren, Kriste Marshall, Gretchen Hahn, Lynne Meltesen, Matthew Brenton, et al. 2004. "Lineage-Specific Gene Duplication and Loss in Human and Great Ape Evolution." *PLoS Biology* 2 (7):E207.
- Galanti, Lior, Dennis Shasha, and Kristin Gunsalus. 2017. "Pheniqs: Fast and Flexible Quality-Aware Sequence Demultiplexing." *bioRxiv*. <https://doi.org/10.1101/128512>.
- Galalde, Daniel R., Elizabeth A. Snell, Daniel Jachimowicz, Botond Sipos, Joseph H. Lloyd, Mark Bruce, Nadia Pantic, et al. 2018. "Highly Parallel Direct RNA Sequencing on an Array of Nanopores." *Nature Methods*, January. <https://doi.org/10.1038/nmeth.4577>.
- Garber, Manuel, Manfred G. Grabherr, Mitchell Guttman, and Cole Trapnell. 2011. "Computational Methods for Transcriptome Annotation and Quantification Using RNA-Seq." *Nature Methods* 8 (6). Nature Publishing Group:469–77.
- García-Alcalde, Fernando, Konstantin Okonechnikov, José Carbonell, Luis M. Cruz, Stefan Götz, Sonia Tarazona, Joaquín Dopazo, Thomas F. Meyer, and Ana Conesa. 2012. "Qualimap: Evaluating next-Generation Sequencing Alignment Data." *Bioinformatics* 28 (20):2678–79.
- Gilad, Yoav, Alicia Oshlack, Gordon K. Smyth, Terence P. Speed, and Kevin P. White. 2006. "Expression Profiling in Primates Reveals a Rapid Evolution of Human Transcription Factors." *Nature* 440 (7081):242–45.
- Gilad, Yoav, Scott A. Rifkin, Paul Bertone, Mark Gerstein, and Kevin P. White. 2005. "Multi-Species Microarrays Reveal the Effect of Sequence Divergence on Gene Expression Profiles." *Genome Research* 15 (5):674–80.
- Gokce, Ozgun, Geoffrey M. Stanley, Barbara Treutlein, Norma F. Neff, J. Gray Camp, Robert

- C. Malenka, Patrick E. Rothwell, Marc V. Fuccillo, Thomas C. Südhof, and Stephen R. Quake. 2016. "Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq." *Cell Reports* 16 (4). Elsevier:1126–37.
- Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." *Nature Reviews. Genetics* 17 (6):333–51.
- Gowda, Malali, Chatchawan Jantasuriyarat, Ralph A. Dean, and Guo-Liang Wang. 2004. "Robust-LongSAGE (RL-SAGE): A Substantially Improved LongSAGE Method for Gene Discovery and Transcriptome Analysis." *Plant Physiology* 134 (3):890–97.
- Griebel, Thasso, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. 2012. "Modelling and Simulating Generic RNA-Seq Experiments with the Flux Simulator." *Nucleic Acids Research* 40 (20):10073–83.
- Grün, Dominic, Lennart Kester, and Alexander van Oudenaarden. 2014. "Validation of Noise Models for Single-Cell Transcriptomics." *Nature Methods* 11 (6):637–40.
- Grün, Dominic, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. 2015. "Single-Cell Messenger RNA Sequencing Reveals Rare Intestinal Cell Types." *Nature* 525 (7568):251–55.
- Grün, Dominic, and Alexander van Oudenaarden. 2015. "Design and Analysis of Single-Cell Sequencing Experiments." *Cell* 163 (4):799–810.
- GTEX Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, et al. 2017. "Genetic Effects on Gene Expression across Human Tissues." *Nature* 550 (7675):204–13.
- Guo, Minzhe, Hui Wang, S. Steven Potter, Jeffrey A. Whitsett, and Yan Xu. 2015. "SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis." *PLoS Computational Biology* 11 (11):e1004575.
- Habib, Naomi, Inbal Avraham-Davidi, Anindita Basu, Tyler Burks, Karthik Shekhar, Matan Hofree, Sourav R. Choudhury, et al. 2017. "Massively Parallel Single-Nucleus RNA-Seq with DroNc-Seq." *Nature Methods* 14 (10):955–58.
- Habib, Naomi, Yinqing Li, Matthias Heidenreich, Lukasz Swiech, Inbal Avraham-Davidi, John J. Trombetta, Cynthia Hession, Feng Zhang, and Aviv Regev. 2016. "Div-Seq: Single-Nucleus RNA-Seq Reveals Dynamics of Rare Adult Newborn Neurons." *Science* 353 (6302):925–28.
- Hancock, John M., John M. Hancock, and Marketa J. Zvelebil. 2004. "HAVANA (Human and Vertebrate Analysis and Annotation)." In *Dictionary of Bioinformatics and Computational Biology*. John Wiley & Sons, Ltd.
- Hashimshony, Tamar, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, et al. 2016. "CEL-Seq2: Sensitive Highly-Multiplexed

- Single-Cell RNA-Seq." *Genome Biology* 17 (1):77.
- Hashimshony, Tamar, Florian Wagner, Noa Sher, and Itai Yanai. 2012. "CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification." *Cell Reports* 2 (3):666–73.
- Hrdlickova, Radmila, Masoud Toloue, and Bin Tian. 2017. "RNA-Seq Methods for Transcriptome Analysis." *Wiley Interdisciplinary Reviews. RNA* 8 (1). <https://doi.org/10.1002/wrna.1364>.
- Hunkapiller, T., R. J. Kaiser, B. F. Koop, and L. Hood. 1991. "Large-Scale and Automated DNA Sequence Determination." *Science* 254 (5028):59–67.
- Ilicic, Tomislav, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C. Marioni, and Sarah A. Teichmann. 2016. "Classification of Low Quality Cells from Single-Cell RNA-Seq Data." *Genome Biology* 17 (1):29.
- Islam, Saiful, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. 2014. "Quantitative Single-Cell RNA-Seq with Unique Molecular Identifiers." *Nature Methods* 11 (2):163–66.
- Jaitin, Diego Adhemar, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, et al. 2014. "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types." *Science* 343 (6172):776–79.
- Jiang, Lichun, Felix Schlesinger, Carrie A. Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R. Gingeras, and Brian Oliver. 2011. "Synthetic Spike-in Standards for RNA-Seq Experiments." *Genome Research* 21 (9):1543–51.
- Katz, Yarden, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. 2010. "Analysis and Design of RNA Sequencing Experiments for Identifying Isoform Regulation." *Nature Methods* 7 (November). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.:1009.
- Kerr, Kathleen F. 2007. "Extended Analysis of Benchmark Datasets for Agilent Two-Color Microarrays." *BMC Bioinformatics* 8 (October):371.
- Khaitovich, Philipp, Wolfgang Enard, Michael Lachmann, and Svante Pääbo. 2006. "Evolution of Primate Gene Expression." *Nature Reviews. Genetics* 7 (9):693–702.
- Khaitovich, Philipp, Ines Hellmann, Wolfgang Enard, Katja Nowick, Marcus Leinweber, Henriette Franz, Gunter Weiss, Michael Lachmann, and Svante Pääbo. 2005. "Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees." *Science* 309 (5742):1850–54.
- Khaitovich, Philipp, Bjoern Muetzel, Xinwei She, Michael Lachmann, Ines Hellmann, Janko Dietzsch, Stephan Steigele, et al. 2004. "Regional Patterns of Gene Expression in Human and Chimpanzee Brains." *Genome Research* 14 (8):1462–73.
- Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. "HISAT: A Fast Spliced

- Aligner with Low Memory Requirements.” *Nature Methods* 12 (4):357–60.
- King, M. C., and A. C. Wilson. 1975. “Evolution at Two Levels in Humans and Chimpanzees.” *Science* 188 (4184):107–16.
- Kircher, Martin, and Janet Kelso. 2010. “High-Throughput DNA Sequencing--Concepts and Limitations.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 32 (6):524–36.
- Kircher, Martin, Susanna Sawyer, and Matthias Meyer. 2012. “Double Indexing Overcomes Inaccuracies in Multiplex Sequencing on the Illumina Platform.” *Nucleic Acids Research* 40 (1):e3.
- Kircher, Martin, Udo Stenzel, and Janet Kelso. 2009. “Improved Base Calling for the Illumina Genome Analyzer Using Machine Learning Strategies.” *Genome Biology* 10 (8):R83.
- Kivioja, Teemu, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. 2011. “Counting Absolute Numbers of Molecules Using Unique Molecular Identifiers.” *Nature Methods* 9 (1):72–74.
- Klein, Allon M., Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. 2015. “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells.” *Cell* 161 (5):1187–1201.
- Kolodziejczyk, Aleksandra A., Jong Kyoung Kim, Valentine Svensson, John C. Marioni, and Sarah A. Teichmann. 2015. “The Technology and Biology of Single-Cell RNA Sequencing.” *Molecular Cell* 58 (4):610–20.
- Kolodziejczyk, Aleksandra A., Jong Kyoung Kim, Jason C. H. Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N. Natarajan, Alex C. Tuck, et al. 2015. “Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation.” *Cell Stem Cell* 17 (4):471–85.
- Kozarewa, Iwanka, Zemin Ning, Michael A. Quail, Mandy J. Sanders, Matthew Berriman, and Daniel J. Turner. 2009. “Amplification-Free Illumina Sequencing-Library Preparation Facilitates Improved Mapping and Assembly of (G+C)-Biased Genomes.” *Nature Methods* 6 (4). Nature Publishing Group:291–95.
- Łabaj, Paweł P., Germán G. Leperc, Bryan E. Linggi, Lye Meng Markillie, H. Steven Wiley, and David P. Kreil. 2011. “Characterization and Improvement of RNA-Seq Precision in Quantitative Transcript Expression Profiling.” *Bioinformatics* 27 (13):i383–91.
- Lacar, Benjamin, Sara B. Linker, Baptiste N. Jaeger, Suguna Krishnaswami, Jerika Barron, Martijn Kelder, Sarah Parylak, et al. 2016. “Nuclear RNA-Seq of Single Neurons Reveals Molecular Signatures of Activation.” *Nature Communications* 7 (April):11022.
- Laehnemann, David, Arndt Borkhardt, and Alice Carolyn McHardy. 2016. “Denoising DNA Deep Sequencing Data-High-Throughput Sequencing Errors and Their Correction.” *Briefings in Bioinformatics* 17 (1):154–79.

- Lake, Blue B., Rizi Ai, Gwendolyn E. Kaeser, Neeraj S. Salathia, Yun C. Yung, Rui Liu, Andre Wildberg, et al. 2016. "Neuronal Subtypes and Diversity Revealed by Single-Nucleus RNA Sequencing of the Human Brain." *Science* 352 (6293):1586–90.
- La Manno, Gioele, Daniel Gyllborg, Simone Codeluppi, Kaneyasu Nishimura, Carmen Salto, Amit Zeisel, Lars E. Borm, et al. 2016. "Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells." *Cell* 167 (2):566–80.e19.
- La Manno, Gioele, Ruslan Soldatov, Hannah Hochgerner, Amit Zeisel, Viktor Petukhov, Maria Kastriiti, Peter Lonnerberg, et al. 2017. "RNA Velocity in Single Cells." *bioRxiv*. <https://doi.org/10.1101/206052>.
- Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822):860–921.
- Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar González-Porta, et al. 2013. "Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans." *Nature* 501 (7468):506–11.
- Levin, Joshua Z., Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. 2010. "Comprehensive Comparative Analysis of Strand-Specific RNA Sequencing Methods." *Nature Methods* 7 (9):709–15.
- Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7):923–30.
- Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16):2078–79.
- Liu, Hongfang, Ionut Bebu, and Xin Li. 2010. "Microarray Probes and Probe Sets." *Frontiers in Bioscience* 2 (January):325–38.
- Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. "The Genotype-Tissue Expression (GTEx) Project." *Nature Genetics* 45 (May). The Author(s):580.
- Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161 (5):1202–14.
- Magalhães, João Pedro de, João Curado, and George M. Church. 2009. "Meta-Analysis of Age-Related Gene Expression Profiles Identifies Common Signatures of Aging." *Bioinformatics* 25 (7):875–81.

- Mamanova, Lira, Robert M. Andrews, Keith D. James, Elizabeth M. Sheridan, Peter D. Ellis, Cordelia F. Langford, Tobias W. B. Ost, John E. Collins, and Daniel J. Turner. 2010. "FRT-Seq: Amplification-Free, Strand-Specific Transcriptome Sequencing." *Nature Methods* 7 (2):130–32.
- Mangul, Serghei, Sarah Van Driesche, Lana S. Martin, Kelsey C. Martin, and Eleazar Eskin. 2017. "UMI-Reducer: Collapsing Duplicate Sequencing Reads via Unique Molecular Identifiers." *bioRxiv*. <https://doi.org/10.1101/103267>.
- MAQC Consortium, Leming Shi, Laura H. Reid, Wendell D. Jones, Richard Shippy, Janet A. Warrington, Shawn C. Baker, et al. 2006. "The MicroArray Quality Control (MAQC) Project Shows Inter- and Intraplatform Reproducibility of Gene Expression Measurements." *Nature Biotechnology* 24 (9):1151–61.
- Margulies, Marcel, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bembien, Jan Berka, et al. 2005. "Genome Sequencing in Microfabricated High-Density Picolitre Reactors." *Nature* 437 (7057):376–80.
- Marioni, John C., Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. 2008. "RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays." *Genome Research* 18 (9):1509–17.
- Martinez-Jimenez, Celia Pilar, Nils Eling, Hung-Chang Chen, Catalina A. Vallejos, Aleksandra A. Kolodziejczyk, Frances Connor, Lovorka Stojic, et al. 2017. "Aging Increases Cell-to-Cell Transcriptional Variability upon Immune Stimulation." *Science* 355 (6332):1433–36.
- Marx, Vivien. 2017. "How to Deduplicate PCR." *Nature Methods* 14 (5):473–76.
- Matsumura, Hideo, Akiko Ito, Hiromasa Saitoh, Peter Winter, Günter Kahl, Monika Reuter, Detlev H. Krüger, and Ryohei Terauchi. 2005. "SuperSAGE." *Cellular Microbiology* 7 (1):11–18.
- McCarroll, Steven A., Coleen T. Murphy, Sige Zou, Scott D. Pletcher, Chen-Shan Chin, Yuh Nung Jan, Cynthia Kenyon, Cornelia I. Bargmann, and Hao Li. 2004. "Comparing Genomic Expression Patterns across Species Identifies Shared Transcriptional Profile in Aging." *Nature Genetics* 36 (2):197–204.
- McCarthy, Davis J., Kieran R. Campbell, Aaron T. L. Lun, and Quin F. Wills. 2017. "Scater: Pre-Processing, Quality Control, Normalization and Visualization of Single-Cell RNA-Seq Data in R." *Bioinformatics*, January. <https://doi.org/10.1093/bioinformatics/btw777>.
- Metzker, Michael L. 2010. "Sequencing Technologies—the next Generation." *Nature Reviews. Genetics* 11 (1). Nature Publishing Group:31–46.
- Meyer, Matthias, Udo Stenzel, Sean Myles, Kay Prüfer, and Michael Hofreiter. 2007. "Targeted High-Throughput Sequencing of Tagged Nucleic Acid Samples." *Nucleic Acids Research* 35 (15):e97.

- Mezlini, Aziz M., Eric J. M. Smith, Marc Fiume, Orion Buske, Gleb L. Savich, Sohrab Shah, Sam Aparicio, Derek Y. Chiang, Anna Goldenberg, and Michael Brudno. 2013. "iReckon: Simultaneous Isoform Discovery and Abundance Estimation from RNA-Seq Data." *Genome Research* 23 (3):519–29.
- Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7):621–28.
- Muraro, Mauro J., Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon van Gurp, et al. 2016. "A Single-Cell Transcriptome Atlas of the Human Pancreas." *Cell Systems* 3 (4):385–94.e3.
- Nellore, Abhinav, Andrew E. Jaffe, Jean-Philippe Fortin, José Alquicira-Hernández, Leonardo Collado-Torres, Siruo Wang, Robert A. Phillips lii, et al. 2016. "Human Splicing Diversity and the Extent of Unannotated Splice Junctions across Human RNA-Seq Samples on the Sequence Read Archive." *Genome Biology* 17 (1):266.
- Okonechnikov, Konstantin, Ana Conesa, and Fernando García-Alcalde. 2016. "Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data." *Bioinformatics* 32 (2):292–94.
- O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciuffo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. "Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation." *Nucleic Acids Research* 44 (D1):D733–45.
- O'Neil, Dominic, Heike Glowatz, and Martin Schlumpberger. 2013. "Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity." *Current Protocols in Molecular Biology* / Edited by Frederick M. Ausubel ... [et Al.] Chapter 4 (July):Unit 4.19.
- Oshlack, Alicia, Adrien E. Chabot, Gordon K. Smyth, and Yoav Gilad. 2007. "Using DNA Microarrays to Study Gene Expression in Closely Related Species." *Bioinformatics* 23 (10). Oxford University Press:1235–42.
- Ozsolak, Fatih, and Patrice M. Milos. 2011. "RNA Sequencing: Advances, Challenges and Opportunities." *Nature Reviews. Genetics* 12 (2):87–98.
- Pachter, Lior. 2011. "Models for Transcript Quantification from RNA-Seq." *arXiv [q-bio.GN]*. arXiv. <http://arxiv.org/abs/1104.3889>.
- Parekh, Swati, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. 2016. "The Impact of Amplification on Differential Expression Analyses by RNA-Seq." *Scientific Reports* 6 (May):25533.
- Parekh, Swati*, Christoph Ziegenhain*, Beate Vieth, Wolfgang Enard, and Ines Hellmann. 2017. "zUMIs: A Fast and Flexible Pipeline to Process RNA Sequencing Data with UMIs." *bioRxiv*. <https://doi.org/10.1101/153940>.

- Patel, Ravi K., and Mukesh Jain. 2012. "NGS QC Toolkit: A Toolkit for Quality Control of next Generation Sequencing Data." *PloS One* 7 (2):e30619.
- Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4):417–19.
- Petropoulos, Sophie, Daniel Edsgård, Björn Reinius, Qiaolin Deng, Sarita Pauliina Panula, Simone Codeluppi, Alvaro Plaza Reyes, Sten Linnarsson, Rickard Sandberg, and Fredrik Lanner. 2016. "Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos." *Cell* 167 (1):285.
- Petukhov, Viktor, Jimin Guo, Ninib Baryawno, Nicolas Severe, David Scadden, and Peter V. Kharchenko. 2017. "Accurate Estimation of Molecular Counts in Droplet-Based Single-Cell RNA-Seq Experiments." *bioRxiv*. <https://doi.org/10.1101/171496>.
- Picelli, Simone, Åsa K. Björklund, Omid R. Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. 2013. "Smart-seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells." *Nature Methods* 10 (11):1096–98.
- Picelli, Simone, Omid R. Faridani, Asa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. 2014. "Full-Length RNA-Seq from Single Cells Using Smart-seq2." *Nature Protocols* 9 (1):171–81.
- Pipes, Lenore, Sheng Li, Marjan Bozinoski, Robert Palermo, Xinxia Peng, Phillip Blood, Sara Kelly, et al. 2013. "The Non-Human Primate Reference Transcriptome Resource (NHPRT) for Comparative Functional Genomics." *Nucleic Acids Research* 41 (Database issue):D906–14.
- Pollen, Alex A., Tomasz J. Nowakowski, Joe Shuga, Xiaohui Wang, Anne A. Leyrat, Jan H. Lui, Nianzhen Li, et al. 2014. "Low-Coverage Single-Cell mRNA Sequencing Reveals Cellular Heterogeneity and Activated Signaling Pathways in Developing Cerebral Cortex." *Nature Biotechnology*, August. <https://doi.org/10.1038/nbt.2967>.
- Poplawski, Alicia, and Harald Binder. 2017. "Feasibility of Sample Size Calculation for RNA-Seq Studies." *Briefings in Bioinformatics*, January. <https://doi.org/10.1093/bib/bbw144>.
- Poulin, Jean-Francois, Bosiljka Tasic, Jens Hjerling-Leffler, Jeffrey M. Trimarchi, and Rajeshwar Awatramani. 2016. "Disentangling Neural Cell Diversity Using Single-Cell Transcriptomics." *Nature Neuroscience* 19 (9):1131–41.
- Putney, S. D., W. C. Herlihy, and P. Schimmel. 1983. "A New Troponin T and cDNA Clones for 13 Different Muscle Proteins, Found by Shotgun Sequencing." *Nature* 302 (5910):718–21.
- Rappolee, D. A., D. Mark, M. J. Banda, and Z. Werb. 1988. "Wound Macrophages Express TGF-Alpha and Other Growth Factors in Vivo: Analysis by mRNA Phenotyping." *Science* 241 (4866). JSTOR:708–12.

- Rasche, Axel, Hadi Al-Hasani, and Ralf Herwig. 2008. "Meta-Analysis Approach Identifies Candidate Genes and Associated Molecular Networks for Type-2 Diabetes Mellitus." *BMC Genomics* 9 (June):310.
- Regev, Aviv, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, et al. 2017. "The Human Cell Atlas." *eLife* 6 (December). <https://doi.org/10.7554/eLife.27041>.
- Reinius, Björn, and Rickard Sandberg. 2015. "Random Monoallelic Expression of Autosomal Genes: Stochastic Transcription and Allele-Level Regulation." *Nature Reviews. Genetics* 16 (11):653–64.
- Renaud, Gabriel, Udo Stenzel, Tomislav Maricic, Victor Wiebe, and Janet Kelso. 2015. "deML: Robust Demultiplexing of Illumina Sequences Using a Likelihood-Based Approach." *Bioinformatics* 31 (5):770–72.
- Reuter, Jason A., Damek V. Spacek, and Michael P. Snyder. 2015. "High-Throughput Sequencing Technologies." *Molecular Cell* 58 (4):586–97.
- Risso, Davide, John Ngai, Terence P. Speed, and Sandrine Dudoit. 2014. "Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples." *Nature Biotechnology* 32 (9):896–902.
- Roberts, Adam, and Lior Pachter. 2013. "Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments." *Nature Methods* 10 (1):71–73.
- Romero, Irene Gallego, Ilya Ruvinsky, and Yoav Gilad. 2012. "Comparative Studies of Gene Expression and the Evolution of Gene Regulation." *Nature Reviews. Genetics* 13 (7):505–16.
- Saha, Saurabh, Andrew B. Sparks, Carlo Rago, Viatcheslav Akmaev, Clarence J. Wang, Bert Vogelstein, Kenneth W. Kinzler, and Victor E. Velculescu. 2002. "Using the Transcriptome to Annotate the Genome." *Nature Biotechnology* 20 (May). Nature Publishing Group:508.
- Sambrook, Joseph, and David William Russell. 2001. *Molecular Cloning: A Laboratory Manual*. CSHL Press.
- Sanger, F., S. Nicklen, and A. R. Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12):5463–67.
- Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." *Science* 270 (5235):467–70.
- Schloss, Jeffery A. 2008. "How to Get Genomes at One Ten-Thousandth the Cost." *Nature Biotechnology* 26 (10):1113–15.

- Schulze, A., and J. Downward. 2001. "Navigating Gene Expression Using Microarrays--a Technology Review." *Nature Cell Biology* 3 (8):E190–95.
- Segal, Eran, Nir Friedman, Naftali Kaminski, Aviv Regev, and Daphne Koller. 2005. "From Signatures to Models: Understanding Cancer Using Microarrays." *Nature Genetics* 37 Suppl (June):S38–45.
- SEQC/MAQC-III Consortium. 2014. "A Comprehensive Assessment of RNA-Seq Accuracy, Reproducibility and Information Content by the Sequencing Quality Control Consortium." *Nature Biotechnology* 32 (9):903–14.
- Shapiro, Ehud, Tamir Biezuner, and Sten Linnarsson. 2013. "Single-Cell Sequencing-Based Technologies Will Revolutionize Whole-Organism Science." *Nature Reviews. Genetics* 14 (9):618–30.
- Shendure, Jay, and Erez Lieberman Aiden. 2012. "The Expanding Scope of DNA Sequencing." *Nature Biotechnology* 30 (11):1084–94.
- Shi, Leming, Gregory Campbell, Wendell D. Jones, Fabien Campagne, Zhining Wen, Stephen J. Walker, Zhenqiang Su, et al. 2010. "The MicroArray Quality Control (MAQC)-II Study of Common Practices for the Development and Validation of Microarray-Based Predictive Models." *Nature Biotechnology* 28 (8):827–38.
- Shippy, Richard, Stephanie Fulmer-Smentek, Roderick V. Jensen, Wendell D. Jones, Paul K. Wolber, Charles D. Johnson, P. Scott Pine, et al. 2006. "Using RNA Sample Titrations to Assess Microarray Platform Performance and Normalization Techniques." *Nature Biotechnology* 24 (9):1123–31.
- Shiroguchi, Katsuyuki, Tony Z. Jia, Peter A. Sims, and X. Sunney Xie. 2012. "Digital RNA Sequencing Minimizes Sequence-Dependent Bias and Amplification Noise with Optimized Single-Molecule Barcodes." *Proceedings of the National Academy of Sciences of the United States of America* 109 (4):1347–52.
- Smith, Tom Sean, Andreas Heger, and Ian Sudbery. 2017. "UMI-Tools: Modelling Sequencing Errors in Unique Molecular Identifiers to Improve Quantification Accuracy." *Genome Research*, January. <https://doi.org/10.1101/gr.209601.116>.
- Soumillon, Magali, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S. Mikkelsen. 2014. "Characterization of Directed Differentiation by High-Throughput Single-Cell RNA-Seq." *bioRxiv*, March. <https://doi.org/10.1101/003236>.
- Sousa, André M. M., Ying Zhu, Mary Ann Raghanti, Robert R. Kitchen, Marco Onorati, Andrew T. N. Tebbenkamp, Bernardo Stutz, et al. 2017. "Molecular and Cellular Reorganization of Neural Circuits in the Human Lineage." *Science* 358 (6366):1027–32.
- Srivastava, Avi, Hirak Sarkar, Nitish Gupta, and Rob Patro. 2016. "RapMap: A Rapid, Sensitive and Accurate Tool for Mapping RNA-Seq Reads to Transcriptomes." *Bioinformatics* 32 (12):i192–200.

- Stegle, Oliver, Sarah A. Teichmann, and John C. Marioni. 2015. "Computational and Analytical Challenges in Single-Cell Transcriptomics." *Nature Reviews. Genetics* 16 (3):133–45.
- Stern, David L., and Virginie Orgogozo. 2008. "The Loci of Evolution: How Predictable Is Genetic Evolution?" *Evolution; International Journal of Organic Evolution* 62 (9):2155–77.
- Stranger, Barbara E., Alexandra C. Nica, Matthew S. Forrest, Antigone Dimas, Christine P. Bird, Claude Beazley, Catherine E. Ingle, et al. 2007. "Population Genomics of Human Gene Expression." *Nature Genetics* 39 (10):1217–24.
- Strasser, Bruno J. 2006. "A World in One Dimension: Linus Pauling, Francis Crick and the Central Dogma of Molecular Biology." *History and Philosophy of the Life Sciences* 28 (4):491–512.
- Sutcliffe, J. G., R. J. Milner, F. E. Bloom, and R. A. Lerner. 1982. "Common 82-Nucleotide Sequence Unique to Brain RNA." *Proceedings of the National Academy of Sciences of the United States of America* 79 (16):4942–46.
- Svensson, Valentine, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J. Miragaia, Charlotte Labalette, Iain C. Macaulay, Ana Cvejic, and Sarah A. Teichmann. 2017. "Power Analysis of Single-Cell RNA-Sequencing Experiments." *Nature Methods* 14 (4):381–87.
- Sweet-Cordero, Alejandro, Sayan Mukherjee, Aravind Subramanian, Han You, Jeffrey J. Roix, Christine Ladd-Acosta, Jill Mesirov, Todd R. Golub, and Tyler Jacks. 2005. "An Oncogenic KRAS2 Expression Signature Identified by Cross-Species Gene-Expression Analysis." *Nature Genetics* 37 (1):48–55.
- Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, et al. 2009. "mRNA-Seq Whole-Transcriptome Analysis of a Single Cell." *Nature Methods* 6 (5):377–82.
- Tasic, Bosiljka, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, et al. 2016. "Adult Mouse Cortical Cell Taxonomy Revealed by Single Cell Transcriptomics." *Nature Neuroscience*, January. <https://doi.org/10.1038/nn.4216>.
- Teng, Mingxiang, Michael I. Love, Carrie A. Davis, Sarah Djebali, Alexander Dobin, Brenton R. Graveley, Sheng Li, et al. 2016. "A Benchmark for RNA-Seq Quantification Pipelines." *Genome Biology* 17 (April):74.
- Tian, Luyi, Shian Su, Daniela Amann-Zalcenstein, Christine Biben, Shalin H. Naik, and Matthew E. Ritchie. 2017. "scPipe: A Flexible Data Preprocessing Pipeline for Single-Cell RNA-Sequencing Data." *bioRxiv*. <https://doi.org/10.1101/175927>.
- Trapnell, Cole. 2015. "Defining Cell Types and States with Single-Cell Genomics." *Genome Research* 25 (10):1491–98.
- Tung, Po-Yuan, John D. Blischak, Chiaowen Joyce Hsiao, David A. Knowles, Jonathan E.

- Burnett, Jonathan K. Pritchard, and Yoav Gilad. 2017. "Batch Effects and the Effective Design of Single-Cell Gene Expression Studies." *Scientific Reports* 7 (January). Nature Publishing Group:39921.
- Uddin, Monica, Derek E. Wildman, Guozhen Liu, Wenbo Xu, Robert M. Johnson, Patrick R. Hof, Gregory Kapatos, Lawrence I. Grossman, and Morris Goodman. 2004. "Sister Grouping of Chimpanzees and Humans as Revealed by Genome-Wide Phylogenetic Analysis of Brain Gene Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 101 (9):2957–62.
- Vallejos, Catalina A., Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C. Marioni. 2017. "Normalizing Single-Cell RNA Sequencing Data: Challenges and Opportunities." *Nature Methods* 14 (6):565–71.
- Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler. 1995. "Serial Analysis of Gene Expression." *Science* 270 (5235):484–87.
- Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507):1304–51.
- Vieth, Beate, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. 2017. "powsimR: Power Analysis for Bulk and Single Cell RNA-Seq Experiments." *Bioinformatics*, July. <https://doi.org/10.1093/bioinformatics/btx435>.
- Villani, Alexandra-Chloé, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, et al. 2017. "Single-Cell RNA-Seq Reveals New Types of Human Blood Dendritic Cells, Monocytes, and Progenitors." *Science* 356 (6335). <https://doi.org/10.1126/science.aah4573>.
- Villar, Diego, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, Miguel Pignatelli, Thomas J. Park, et al. 2015. "Enhancer Evolution across 20 Mammalian Species." *Cell* 160 (3):554–66.
- Voelkerding, Karl V., Shale A. Dames, and Jacob D. Durtschi. 2009. "Next-Generation Sequencing: From Basic Research to Diagnostics." *Clinical Chemistry* 55 (4):641–58.
- Wagner, Allon, Aviv Regev, and Nir Yosef. 2016. "Revealing the Vectors of Cellular Identity with Single-Cell Genomics." *Nature Biotechnology* 34 (11):1145–60.
- Wang, A. M., M. V. Doyle, and D. F. Mark. 1989. "Quantitation of mRNA by the Polymerase Chain Reaction." *Proceedings of the National Academy of Sciences of the United States of America* 86 (24). National Acad Sciences:9717–21.
- Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews. Genetics* 10 (1):57–63.
- Warnefors, Maria, and Henrik Kaessmann. 2013. "Evolution of the Correlation between Expression Divergence and Protein Divergence in Mammals." *Genome Biology and Evolution* 5 (7):1324–35.

- Weis, J. H., S. S. Tan, B. K. Martin, and C. T. Wittwer. 1992. "Detection of Rare mRNAs via Quantitative RT-PCR." *Trends in Genetics: TIG* 8 (8):263–64.
- Wills, Quin F., Kenneth J. Livak, Alex J. Tipping, Tariq Enver, Andrew J. Goldson, Darren W. Sexton, and Chris Holmes. 2013. "Single-Cell Gene Expression Analysis Reveals Genetic Associations Masked in Whole-Tissue Experiments." *Nature Biotechnology* 31 (8):748–52.
- Wu, Hao, Chi Wang, and Zhijin Wu. 2015. "PROPER: Comprehensive Power Evaluation for Differential Expression Using RNA-Seq." *Bioinformatics* 31 (2):233–41.
- Wunderlich, Stephanie, Martin Kircher, Beate Vieth, Alexandra Haase, Sylvia Merkert, Jennifer Beier, Gudrun Göhring, et al. 2014. "Primate iPS Cells as Tools for Evolutionary Analyses." *Stem Cell Research* 12 (3):622–29.
- Zeisel, Amit, Ana B. Muñoz Machado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, et al. 2015. "Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-Cell RNA-Seq." *Science*, February. <https://doi.org/10.1126/science.aaa1934>.
- Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2017. "Ensembl 2018." *Nucleic Acids Research*, November. <https://doi.org/10.1093/nar/gkx1098>.
- Zhao, Shanrong. 2014. "Assessment of the Impact of Using a Reference Transcriptome in Mapping Short RNA-Seq Reads." *PloS One* 9 (7):e101374.
- Zhao, Shanrong, and Baohong Zhang. 2015. "A Comprehensive Evaluation of Ensembl, RefSeq, and UCSC Annotations in the Context of RNA-Seq Read Mapping and Gene Quantification." *BMC Genomics* 16 (February):97.
- Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. "Massively Parallel Digital Transcriptional Profiling of Single Cells." *Nature Communications* 8 (January):14049.
- Zhou, Wanding, Tenghui Chen, Hao Zhao, Agda Karina Eterovic, Funda Meric-Bernstam, Gordon B. Mills, and Ken Chen. 2014. "Bias from Removing Read Duplication in Ultra-Deep Sequencing Experiments." *Bioinformatics*, January. <https://doi.org/10.1093/bioinformatics/btt771>.
- Zhou, Xiaofan, and Antonis Rokas. 2014. "Prevention, Diagnosis and Treatment of High-Throughput Sequencing Data Pathologies." *Molecular Ecology* 23 (7):1679–1700.
- Zhu, Ying, Mingfeng Li, André M. M. Sousa, and Nenad Sestan. 2014. "XSAnno: A Framework for Building Ortholog Models in Cross-Species Transcriptome Comparisons." *BMC Genomics* 15 (May):343.
- Ziegenhain, Christoph, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard.

2017. "Comparative Analysis of Single-Cell RNA Sequencing Methods." *Molecular Cell* 65 (4). Elsevier:631–43.e4.

Zilionis, Rapolas, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M. Klein, and Linas Mazutis. 2017. "Single-Cell Barcoding and Sequencing Using Droplet Microfluidics." *Nature Protocols* 12 (1):44–73.

Zimmerman, E. 2014. "Smartest Companies: Illumina." MIT Technology Review.

List of Figures

Figure 1: Basic steps of RNA-sequencing.	p. 21
Figure 2: RNA-seq data processing.	p. 23
Figure 3: Schematic of variable PCR amplification rate.	p. 29
Figure 4: Unique Molecular Identifiers.	p. 30
Figure 5: Annotation difference between species.	p. 36

List of Tables

Table 1: Utilities of commonly used RNA-seq mapping and quantification tools.	p. 26
Table 2: Method specific barcode read information.	p. 33

Acknowledgments

First and foremost, I want to thank Ines Hellmann for believing in me and introducing me to Wolfgang Enard. Thank you Wolfgang for providing me an opportunity to work with you. Wolfgang and Ines, you have created this incredibly positive atmosphere in the group that helped me grow as a scientist by realising my own potential. Thank you Wolfgang for providing timely advice anytime just with a knock on the door and always helping me get through any problems.

Ines, you have not only been a PhD supervisor but also a mentor who made me what I am today as a scientist. Thank you for a warm welcome to Germany, for introducing me to the culture, for being very clear in how things should be done and last but not least the way of pursuing science.

A big vote of thanks to the other two members of the “holy trinity”, Beate and Christoph. Riding the rollercoaster of PhD life with you two made it much easier and filled with cherishable memories. I don't think I can list everything that I am thankful for with you two. Drinking coffees, raging, sharing ideas, hand holding, debugging, developing ... everything.

A special thanks to Sabrina for being my first but shared student with Christoph. You came to me with your issues and I rather learned many more things while solving them. Thank you for introducing me to cool sports like skiing, climbing, longboarding and being so diligent and patient while me being at sloth pace in learning to ski.

Thank you Ines Bliesener for your moral support, eating breakfast together and sharing cute gifts. I want to thank Mari and Johanna for iPSCs and a lot of help with my stupid biology questions. A special thanks to Aleks & Ilse for helping in grammar correction in the thesis.

I would also like to thank every past and present members of AG Enard for their wonderful company and creating awesome work culture.

A hearty thanks to Lola, Freya, Elly and Daisy for creating positive energy in the office. You puppies always calm me down when something was not right.

My sincere thanks to Sylvia, my landlady for being such a sweet host for 4 years.

I owe a big thanks to my parents and my little niece Heerva for their unconditional love and faith in me. My special thanks to Dost for his open ended support and the best friendship I could ever imagine.

Last but not least, my husband Neeraj, I could have not done this without his love and support. Thank you for always being there for me.