Statistical Power Analysis for Single-Cell RNA-Sequencing

Dissertation von Beate Vieth

München 2019

Statistical Power Analysis for Single-Cell RNA-Sequencing

Dissertation an der Fakultät für Biologie

der Ludwig-Maximilians-Universität München

Beate Vieth

München, 2019

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

Erstgutachter:	Professor Dr. Wolfgang Enard
Zweitgutachter:	Professor Dr. Jochen Wolf
Tag der Abgabe:	17. Oktober 2019
Tag der mündlichen Prüfung:	20. Januar 2020

Eidestattliche Versicherung und Erklärung

Eidesstattliche Erklärung

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

München, den 15. Oktober 2019

Beate Vieth

Erklärung

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

München, den 15. Oktober 2019

Beate Vieth

Contents

Abbreviations

Abbreviation	Definition
CPM	Counts Per Million
DE	differential expression
DGE	differential gene expression
DNA	deoxyribonucleic acid
DTW	Dynamic Time Warping
EE	equal expression
ERCC	External RNA Controls Consortium
FACS	Fluorescence-activated cell sorting
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase Million
FWER	Family-Wise Error Rate
IVT	in-vitro transcription
MNN	Mutual Nearest Neighbours
MR	median of ratios
mRNA	messenger RNA
MST	minimal spanning tree
MTP	multiple testing problem
NB	Negative Binomial
P	Poisson
PBMC	peripheral blood mononuclear cells
PCA	principal component analysis
PCR	Polymerase Chain Reaction
RNA	ribonucleic acid
RNA-seq	RNA-sequencing
rRNA	ribosomal RNA
scRNA-seq	Single-cell RNA sequencing
STAMPs	single-cell transcriptomes attached to microparticles
SVM	Support Vector Machines
TMM	weighted trimmed mean of M-values
TPM	Transcripts Per Kilobase Million
TPR	True Positive Rate
UMI	unique molecular identifier
ZINP	Zero-inflated Negative Binomial
ZIP	Zero-inflated Poisson
α	type I error rate
β	type II error rate
1 - β	statistical power

Chronological List of Publications

I. Parekh S, Ziegenhain C, **Vieth B**, Enard W, Hellmann I:
"The impact of amplification on differential expression analyses by RNA-seq." (2016)
Scientific Reports 6 (25533).
doi: 10.1038/srep25533

II. Ziegenhain C, **Vieth B**, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W:
"Comparative Analysis of Single-Cell RNA Sequencing Methods." (2017)
Molecular Cell 65 (4): 631–643.e4.
doi: 10.1016/j.molcel.2017.01.023

III. **Vieth B**, Ziegenhain C, Parekh S, Enard W, Hellmann I:
"powsimR: Power Analysis for Bulk and Single Cell RNA-Seq Experiments." (2017)
Bioinformatics 33 (21): 3486-3488.
doi: 10.1093/bioinformatics/btx435

IV. Parekh S*, Ziegenhain C*, **Vieth B**, Hellmann I, Enard W:
"zUMIs: A fast and flexible pipeline to process RNA sequencing data with UMIs." (2018)
GigaScience 7 (6).giy059.
doi: 10.1093/gigascience/giy059

V. **Vieth B**, Parekh S, Ziegenhain C, Parekh S, Enard W, Hellmann I:
"A Systematic Evaluation of Single Cell RNA-Seq Analysis Pipelines" (2019)
Nature Communications 10 (1):4667-4678.
doi: 10.1038/s41467-019-12266-7

Other Publications

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

VII. Schreck C, Istvánffy R, Ziegenhain C, Sippenauer T, Ruf F, Henkel L, Gärtner F, **Vieth B**, Florian MC, Mende N, Taubenberger A, Prendergast Á, Wagner A, Pagel C, Grziwok S, Götze KS, Guck J, Dean DC, Massberg S, Oostendorp RA:
"Niche WNT5A regulates the actin cytoskeleton during regeneration of hematopoietic stem cells." (2017)
Journal of Experimental Medicine 214 (1): 165-181.

VIII. Ziegenhain C*, **Vieth B***, Parekh S*, Hellmann I, Enard W:
"Quantitative single-cell transcriptomics." (2018)
Briefings in Functional Genomics 17 (4): 220-232.

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

X. Medvedeva VP, Rieger MA, **Vieth B**, Mombereau C, Ziegenhain C, Ghosh T, Cressant A, Enard W, Granon S, Dougherty JD, Groszer M:
"Altered Social Behavior in Mice Carrying a Cortical Foxp2 Deletion." (2019)
Human Molecular Genetics 28 (5): 701–17.

XI. Schreiweis C, Irinopoulou T, **Vieth B**, Laddada L, Oury F, Burguière E, Enard W, Groszer M:
"Mice carrying a humanized Foxp2 knock-in allele show region-specific shifts of striatal Foxp2 expression levels." (2019)
Cortex 118: 212–222.

Declarations of contribution as a co-author

The impact of amplification on differential expression analyses by RNA-seq
This study was conceived by Swati Parekh and Christoph Ziegenhain. I helped with data processing and power simulations. The manuscript was written by Swati Parekh, Ines Hellmann and Wolfgang Enard.

Comparative Analysis of Single-Cell RNA Sequencing Methods
This study was conceived and conducted by Christoph Ziegenhain. I helped with data processing and developed a framework for single cell gene expression simulation and statistical power analysis. The manuscript was written by Christoph Ziegenhain, Wolfgang Enard and Ines Hellmann.

powsimR: Power analysis for bulk and single cell RNA-seq experiments
Ines Hellmann and I conceived the study. The idea for this work emerged after working on power simulations for "The impact of amplification on differential expression analyses by RNA-seq" and "Comparative Analysis of Single-Cell RNA Sequencing Methods". I developed and programmed powsimR. I also tested the program and evaluated its performance relative to empirical scRNA-seq data. Ines Hellmann, Wolfgang Enard and I wrote the manuscript.

zUMIs: A fast and flexible pipeline to process RNA sequencing data with UMIs
Swati Parekh and Christoph Ziegenhain had the idea to this work, designed and implemented the pipeline. I tested the pipeline and performed power simulations to evaluate intron mappings. Swati Parekh, Christoph Ziegenhain, Ines Hellmann, and Wolfgang Enard wrote the manuscript.

A Systematic Evaluation of Single Cell RNA-Seq Analysis Pipelines
This study was conceived by Ines Hellmann and me. I prepared and analysed the scRNA-seq data. I implemented and conducted the simulation and evaluation framework. 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 Beate Vieth to these publications.

Wolfgang Enard

Aims of the thesis

Single-cell RNA sequencing (scRNA-seq) enables us to profile the gene expression of individual cells. As this technology has rapidly evolved over the last decade, it has provided us with the ability to discover previously unknown cell types and to analyze expression patterns during differentiation, development and cancer with single cell resolution. So while scRNA-seq is becoming an important tool for biology and medicine, it is still a fast-evolving technology and many experimental and computational challenges have not been resolved yet.

The core of my work focuses on the implementation of a realistic simulation framework for scRNA-seq experiments. This enabled us to explore experimental techniques and computational tools for generating and analyzing scRNA-seq data.

As a first study, we examined whether whole-transcriptome amplification introduces unwanted noise or bias in gene expression profiling (Manuscript I). Further investigating RNA-seq library preparation techniques, we then compared six prominent protocols for which I implemented simulations that provided us with the ground truth to assess the protocol's sensitivity and specificity in detecting differential gene expression (DGE) (Manuscript II). To provide the simulation framework as a user-friendly tool, I extended it further and wrapped it up as a user-friendly R-package, powsimR, enabling researchers to evaluate statistical power and sample size requirements for bulk and single-cell RNA-seq experiments (Manuscript III). We also developed zUMIs, a fast and flexible pipeline for RNA-seq with UMIs. Again powsimR simulations were instrumental to evaluate whether the inclusion of intron mapping in gene expression quantification affected the power to detect DGE (Manuscript IV). Lastly, I used powsimR to systematically evaluate $\approx$ 3'000 scRNA-seq analysis pipelines and how choices of library preparation, mapping, count processing and normalisation affect the ability to detect differential expression (Manuscript V).

In conclusion, my thesis covers many aspects of the computational analysis essential for scRNA-seq. I developed a faithful simulation framework that can help in developing and evaluating methods, introduced the first statistical power analysis tool for scRNA-seq and showed how computational choices can affect the validity of scRNA-seq experiments.

Summary

RNA-sequencing (RNA-seq) is an established method to quantify levels of gene expression genome-wide. The recent development of single cell RNA sequencing (scRNA-seq) protocols opens up the possibility to systematically characterize cell transcriptomes and their underlying developmental and regulatory mechanisms. Since the first publication on single-cell transcriptomics a decade ago, hundreds of scRNA-seq datasets from a variety of sources have been released, profiling gene expression of sorted cells, tumors, whole dissociated organs and even complete organisms. Currently, it is also the main tool to systematically characterize human cells within the Human Cell Atlas Project.

Given its wide applicability and increasing popularity, many experimental protocols and computational analysis approaches exist for scRNA-seq. However, the technology remains experimentally and computationally challenging. Firstly, single cells contain only minute mRNA amounts that need to be reliably captured and amplified for accurate quantification by sequencing. Importantly, the Polymerase Chain Reaction (PCR) is commonly used for amplification which might introduce biases and increase technical variation. Secondly, once the sequencing results are obtained, finding the best computational processing pipeline can be a struggle. A number of comparison studies have already been conducted - esp. for bulk RNA-seq - but usually they deal only with one aspect of the workflow. Furthermore, in how far the conclusions and recommendations of these studies can be transferred to scRNA-seq is unknown.

Related to the processing of RNA-sequencing, we investigate the effect of PCR amplification on differential expression analysis. We find that computational removal of duplicates has either a negligible or a negative impact on specificity and sensitivity of differential expression

analysis, and we therefore recommend not to remove read duplicates by mapping position. In contrast, if duplicates are identified using unique molecular identifiers (UMIs) tagging RNA molecules, both specificity and sensitivity improve.

The first integral step of any scRNA-seq experiment is the preparation of sequencing libraries from the cells. We conducted an independent benchmarking study of popular library preparation protocols in terms of detection sensitivity, accuracy and precision using the same mouse embryonic stem cells and exogenous mRNA spike-ins. We recapitulate our previous finding that technical variance is markedly decreased when using UMIs to remove duplicates. In order to assign a monetary value to the detected amounts of technical variance, we developed a simulation framework, that enabled us to compare the power to detect differentially expressed genes across the scRNA-seq library preparation protocols. Our experiences during this comparison study led to the development of the sequencing data processing in zUMIs and the simulation framework and power analysis in powsimR. zUMIs is a pipeline for processing scRNA-seq data with flexible choices regarding UMI and cell barcode design. In addition, we showed with powsimR simulations that the inclusion of intronic reads for gene expression quantification increases the power to detect DE genes and added it as a unique feature to zUMIs. In powsimR, we present our simulation framework extending choices concerning data analysis, enabling researchers to assess experimental design and analysis plans of RNA-seq in terms of statistical power.

Lastly, we conducted a systematic evaluation of scRNA-seq experimental and analytical pipelines. We found that choices made concerning normalisation and library preparation protocols have the biggest impact on the validity of scRNA-seq DE analysis. Choosing a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the cell sample size.

Taken together, we have established and applied a simulation framework that allowed us to benchmark experimental and computational scRNA-seq protocols and hence inform the experimental design and method choices of this important technology.

1 | Introduction

In the following sections, I provide the background information necessary for understanding my work in this thesis, including an introduction to basic cell biology and gene expression, the technologies we use to measure expression and computational methods for analyzing the resulting data. The section **Gene Expression** describes how information flows and is controlled within a cell and the molecules, especially messenger RNA, involved in these processes. In the following section, I introduce **RNA-Sequencing** as a method to study gene expression, starting with the basic principles of library preparation and data generation by high-throughput sequencing to covering the established computational approaches that are used to extract meaning from this kind of sequencing data.

In section **Single-cell RNA-Sequencing**, I present and discuss new technologies that have enabled the measurements of gene expression levels of individual cells, detailing the necessary steps of isolating and capturing of single cells and finally preparing sequencing libraries. I outline in depth the following computational analysis pipeline, highlighting the many possible methods to answer research questions utilizing scRNA-seq.

Given the focus of my thesis, namely the development of a simulation framework for scRNA-seq, I cover in my last introductory section **Experimental Design and Power Analysis**. In particular, I outline how simulations are a useful tool not only for statistical power analysis and sample size calculations, but how simulations are an ideal framework for evaluating and comparing all aspects of scRNA-seq, including library preparation as well as computational analysis tools.

1.1 Gene Expression

The central dogma of molecular biology describes the flow of information within a biological system[1,2]. Deoxyribonucleic acid (DNA) is the essential molecular basis for carrying genetic information within the cell[3]. The information flow starts with the transcription of DNA regions, called genes, into ribonucleic acid (RNA). After this process of transcription, the copies of RNA get translated into proteins. Genes together with the DNA sequences that regulate when and how much RNA of a gene is transcribed, is defined as the functional part of the DNA. Together with the nonfunctional DNA, the haploid DNA content of an organism is called its genome[4].

In eukaryotes, when an RNA molecule is transcribed from genes, it initially contains sequence regions that encode information (exons) that alternate with much larger non-coding sequence regions (introns). The intronic sequences are removed through a process known as RNA splicing and a sequence of adenine nucleotides - so called poly(A) tail - is added where transcription ends at the 3' end. This splicing process allows multiple forms of a transcript (isoforms) to be produced from a single gene by selecting which coding sequences are retained or removed. The expression level of a gene is given by the total number of RNA copies present within a cell. Ultimately, the mature mRNA transcript is converted to a protein made up of amino acids in the ribosome. This process is called translation. Proteins are the building blocks of cells and essential for proper functioning. These functions include tasks such as metabolism, nutrient transportation, sensing environmental cues and gene expression regulation.

The regulation of gene expression is the basis to determine the cell's development and function within a multi-cellular organism. There are multiple mechanisms contributing to this regulation. For instance, the chromatin state is given by the set of chromatin-associated proteins and histone modifications that determine the accessibility of a gene for transcription [5]. DNA methylation of cytosine residues of CpG dense promoters is a major driver of gene expression silencing[6], whereas highly transcribed genes have an enriched methylation of the gene body[7]. The binding of transcription factors to specific regulatory DNA sequences such as promoters and enhancers can drive or repress the transcriptional process[8]. The

aforementioned maturation process, alternative splicing as well as the lifetime of mRNAs are examples of mechanisms controlling RNA processing and stability [4].

There are many possibilities beyond the regulation of mRNA transcript abundances that affect the location and efficiency of translation as well as the location and function of proteins. Nevertheless, the types and amounts of mRNA transcripts set the basis for all of these processes. Hence, quantifying the mRNA make-up of a cell is highly informative.

1.2 RNA-sequencing

RNA sequencing (RNA-seq) provides a reliable method for measuring RNA expression levels with high throughput [9] (Figure 1.1). In essence, RNA is isolated from a biological sample, converted to complementary DNA (cDNA) and after the addition of adapters, the resulting library can be sequenced on a machine as manufactured by the company Illumina. The sequencing output consists of millions of short nucleotide sequences, so called reads. Compared to previous assays like probe-based micro-arrays [10], RNA-seq is able to cover a broader range of gene expression levels and in addition, no prior knowledge of coding sequences is required to measure gene expression [11]. Furthermore, RNA-seq enables a genome-wide survey of the transcriptome in contrast to quantitative PCR techniques of single mRNA species [12]. Coupled with the rapid decrease in sequencing costs [13], RNA-seq has become the most prevalent method to quantify gene expression [14].

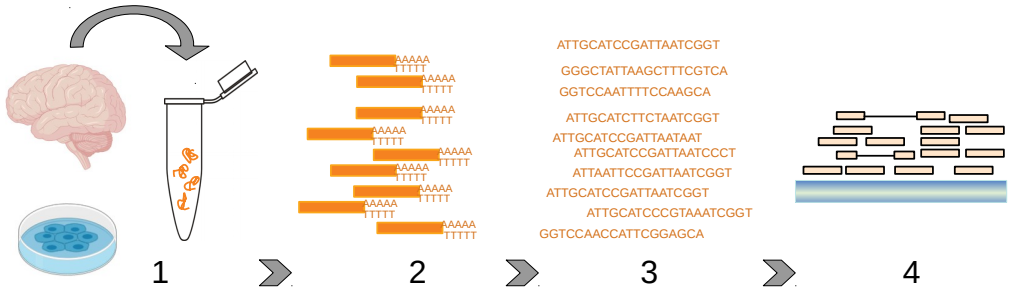


Figure 1.1. RNA-sequencing workflow.
1) Isolation of RNA from cells or tissues. 2) mRNA fraction is purified and cDNA libraries are prepared. 3) Sequencing of short nucleotide reads. 4) Quantification of gene expression.

1.2.1 Library Preparation

The lysis of living cells is the first step in preparing a sample for RNA-seq, whereby the membrane disintegrates and the cellular content is released into the buffer solution. RNA molecules can then be isolated, by physical separation (e.g. silica column) or chemical extraction (e.g. phenol)[15]. In mammalian cells, the vast majority of RNA molecules are ribosomal, contributing more than 80 percent of the total RNA content[16]. On the other hand, the mRNA fraction only ranges between 2 to 7 percent[17]. Thus, sequencing the total RNA content of a sample would reduce our ability to detect this rarer RNA species given that the total amount of sequencing is a limiting factor in most RNA-seq experiments. Hence, selection methods are applied[18]. polyA enrichment is widely used which is achieved by oligonucleotide probes that bind to the poly(A) tail of mRNA molecules and thereby rRNA is passively depleted (Figure 1.2). An alternative method is active ribosomal RNA depletion by Ribonuclease H Enzyme or rRNA-specific probes. In either case, each selection method has its limitations and the choice of selection method has been shown to result in biases[19,20].

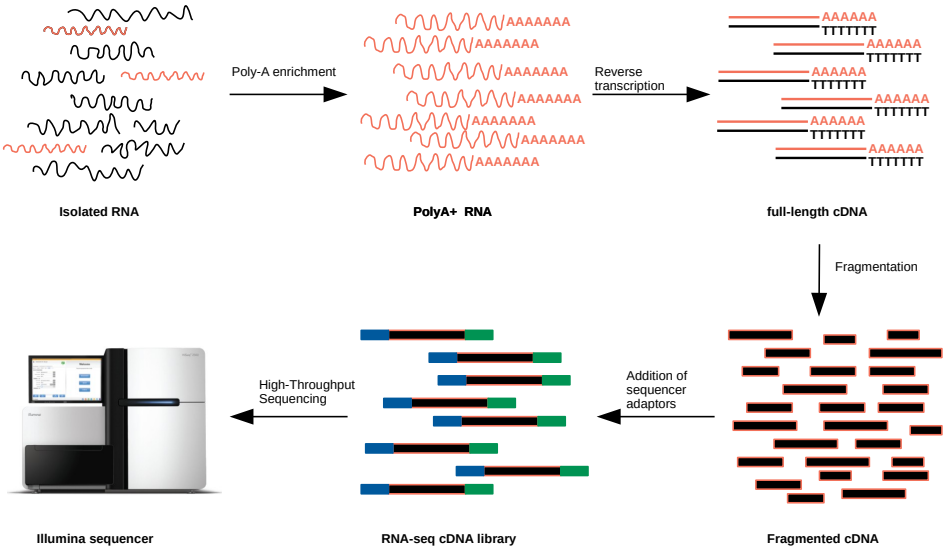


Figure 1.2. RNA-seq library preparation.
After RNA extraction, mRNA with polyA-tails are enriched prior to reverse transcription. The resulting cDNA is fragmented and tagged with sequencing adapters. This final library is then sequenced on a high-throughput machine, e.g. Illumina Hiseq sequencer (adapted from Illumina, Inc.[21]).

Most sequencing machines only work with DNA, so the sample must first be reverse transcribed into complementary DNA (cDNA) using a retroviral enzyme[18]. Furthermore, mRNA transcripts are usually much longer than the sequencing machines maximal efficient read length, so that the molecules need to be fragmented before hand[22]. In some protocols, fragmentation is performed after conversion to cDNA, e.g. by enzymatic processes[23], rather than at the RNA stage, e.g. by heat fragmentation[24]. To complete the workflow, it is necessary to attach adapter sequences to the cDNA libraries that are used to bind the molecules on the flowcells of the sequencing machines[25]. Usually these sequences also contain indexes that enables the multiplexing of samples in an individual sequencing run.

1.2.2 High-throughput sequencing

The Illumina machines with their trademarked Sequence by Synthesis technology are the most prevalent sequencing platform[26] (Figure 1.3). In a first step, the double stranded cDNA fragments are separated into single stranded DNA so that the sequencing adapters can bind to complementary sequences on the flow cell (Figure 1.3A). This hybridization happens at both ends of the fragment, so that a bridge structure is formed along which the complementary DNA strand is synthesized by enzymes. After repeated rounds of this amplification, clonal clusters consisting of approximately a thousand copies of each fragment are created[27]. The bridge is denatured by adapter cleavage and the amplified single-stranded DNA fragments are now ready for sequencing[28]. For that, proprietary modified nucleotides with fluorescent labels are incorporated, the flow cell is washed to remove unbound fluorophores and the bound fluorophores are detected by laser excitation and direct imaging[27] (Figure 1.3B). The nucleotides also act as terminators of synthesis for each reaction. After multiple rounds of these sequencing reactions, the resulting images are processed by e.g. bcl2fastq to produce millions of nucleotide sequences with associated quality scores (PHRED)[28,29]

1.2.3 Processing of RNA-sequencing data

Usually, multiple RNA-seq libraries are sequenced on one flow cell together and the libraries therefore need to be demultiplexed[30] (Figure 1.4). This is followed by aligning the reads to a reference genome. There are specific aligners such as STAR[31] that take the splicing

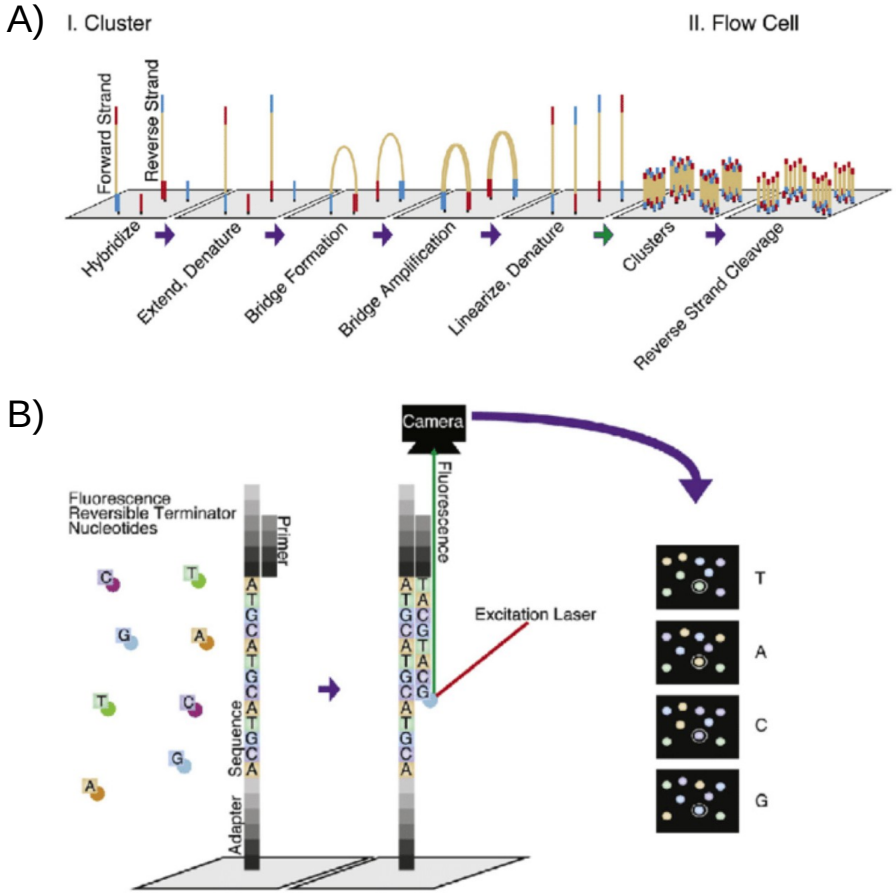


Figure 1.3. Illumina sequencing workflow.
A) Library fragments are flowed across a flow cell and hybridize with complementary Illumina adapter oligos. Complementary fragments are extended, amplified via bridge amplification PCR, and denatured, resulting in clusters of identical single-stranded library fragments. B) Fragments are primed and sequenced utilizing reversible terminator nucleotides. Base pairs are identified after laser excitation and fluorescence detection. (taken from Chaitankar et al. 2016[28], CC BY-NC-ND 4.0 license)

structure of mRNA transcripts into account in contrast to aligners designed for genomic DNA sequencing, e.g. bwa[32]. The expression is then quantified by counting the reads that overlap annotated genetic features[33,34]. Quality control of all steps from wet lab to computational processing are essential[35,36]: Starting from the library creation, e.g. fragment size distribution, over sequencing results, e.g. base-calling score distribution and over-representation of polyA sequence reads, to expression quantification, e.g. read alignments and gene length coverage. Synthetic spike-in standards such as the RNA controls developed by the External RNA Controls Consortium (ERCC)[37] can be a useful tool to measure sensitivity, accuracy and possible biases and limitations of RNA-seq experiments[38,39].

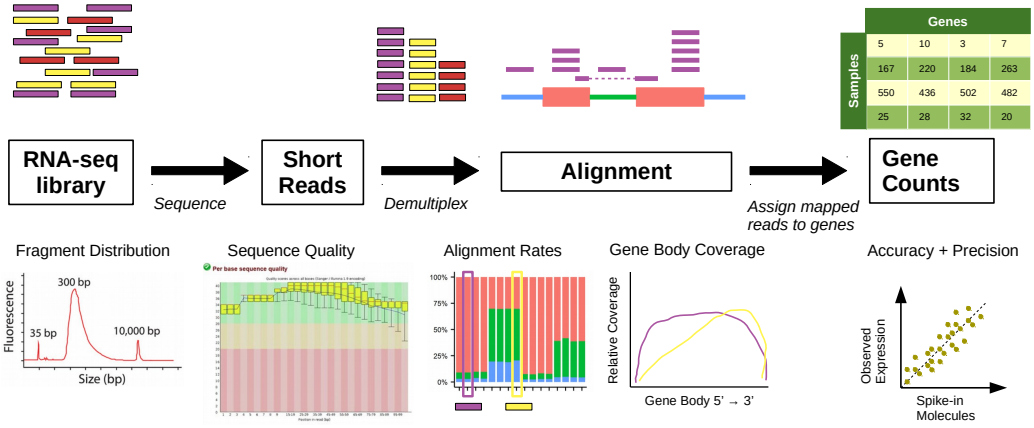


Figure 1.4. Computational pipeline for RNA-seq data.
After sequencing the library, the short reads are demultiplexed according to multiplexing barcodes, aligned to a reference genome, and mappings are assigned to genetic features. The lower panel illustrates a selection of quality measures and filters for each step.

Once the expression levels are quantified, there is a wide range of possible downstream analyses that are well established: Genetic sequence variants can be directly identified from RNA-seq reads that might affect gene expression levels e.g. for eQTL mapping studies[40]. RNA-seq can also be used to identify alternative splicing, transcription start sites and isoform switching as well as the differential abundances over time[41]. Nevertheless the majority of RNA-seq studies focus on the comparison of gene expression levels across sample conditions [29]. This is usually extended with further downstream analysis such as gene set enrichment analysis[42] and network analysis[43].

1.3 Single-Cell RNA-sequencing

In contrast to bulk experiments, single-cell RNA sequencing (scRNA-seq) enables the investigation of the transcriptome with single-cell resolution[44]. This resolution is particularly important for identifying cell type-specific developments or reactions to perturbations[45]. In bulk RNA-seq, this is often hindered by averaging expression profiles over all cells present in a biological sample, thereby also masking the cellular composition. Hence, studies previously selected or enriched specific cell types prior to library preparation. However, this separation might limit the investigation of dynamic interactions in complex systems such as multi-cellular tissues. With scRNA-seq technologies it is in principle possible to look at the transcriptome of all the cell types in a tissue simultaneously, allowing a fine-grained look at individual cell types and enabling the discovery of previously unknown cell types[46]. The first scRNA-seq protocol was published in 2009[44] (Figure 1.5A). While this approach allowed measurements of the transcriptome in individual cells, it required labor-intensive manual isolation and library preparation so that the transcriptome of only a few blastomere cells were profiled. Since then, many scRNA-seq protocols have been developed and the number of cells in scRNA-seq experiments has scaled exponentially[47] (Figure 1.5B).

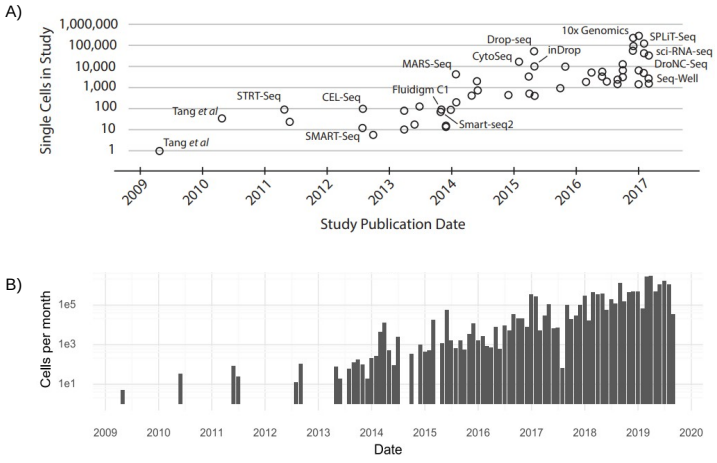


Figure 1.5. Scale of scRNA-seq experiments.
A) Cell numbers reported in representative publications by publication date. Key technologies and protocols are marked (figure taken from Svensson et al. 2017[47], CC BY 4.0 license). B) The aggregate number of cells measured per month since 2009 (figure taken from Svensson & da Veiga Beltrame 2019[48], CC BY-NC-ND 4.0 license).

1.3.1 Isolating Single Cells

In order to capture cells they must first be dissociated into single cell suspensions. This is an essential, sometimes overlooked non-trivial task (Figure 1.6). The isolation of already suspended, cultured cells is rather straightforward, whereas dissociation of cells from solid tissues like the brain or tumors can be challenging [49,50]. Furthermore, dissociation treatments may affect the well-being of the cells as well as their transcriptome. Cells might need different dissociation times which could lead to depletion of certain cell types and/or clumping of other cell types due to incomplete dissociation [51]. Nevertheless, once these obstacles are overcome, the next major step is the capture of single cells (Figure 1.6). Many studies have relied on a preselection of cell types using known molecular markers compatible with Fluorescence-activated cell sorting (FACS) so that individual cells can be directly sorted into individual wells on a plate, e.g. peripheral blood mononuclear cells (PBMC) [52].

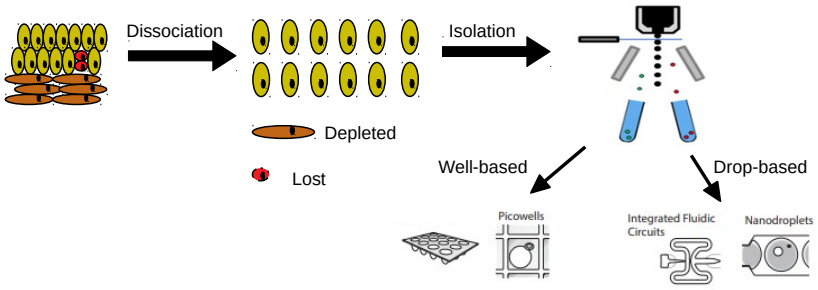


Figure 1.6. Isolating and capturing single cells for sequencing.
Dissociation can lead to depletion of certain cell types by damage and cell death. The dissociated cells are isolated and captured using well-based or microfluidic / droplet-based technologies (single cell capture technologies figures taken from Svensson et al. 2017 [47], CC BY 4.0 license; schematic FACS figure taken from Dholakia et al. 2007 [53], CC BY 4.0 license).

1.3.2 Capturing Single Cells

Early Single-Cell Capture Technologies

The first commercially available cell capture platform is the Fluidigm C1 instrument [54,55]. This system uses microfluidics to separate cells into individual wells on a chip. The captured cells are then lyzed, the mRNA is reverse transcribed to yield cDNA which is then amplified

by polymerase chain reaction (PCR). This platform provides opportunities for a range of experiments but also has a number of known disadvantages and limitations: The chips used have a fixed size range, meaning that only cells of a particular size can be captured in a single run. Furthermore, chips have only a limited number of captures, where the 96 well plate chip is the most commonly used. On top of that, capturing multiple cells is a known issue[56].

Droplet Technologies

Other microfluidic devices have also been developed that rely on droplet chemistry for the encapsulation of cells in combination with barcoding beads, thereby dramatically increasing cellular throughput as well as reducing the costs compared to individual barcoding and amplification of well-based methods[57] (Figure 1.5A). On the other hand, given that so many cells are captured and prepared, single cell transcriptomes are sequenced at a much lower depth with these devices. Drop-seq[58], InDrop[59] and InDrops[60] were the first representatives of this approach: The cell suspension are piped into those devices where they form aqueous droplets, together with the lysis buffer and beads, within mineral oil. Inside the droplets, the cells are then lyzed and the mRNA molecules hybridize with the primers on the beads. After this initial capture of single-cell transcriptomes attached to microparticles (STAMPs), the droplets are broken and pooled for reverse transcription and PCR amplification, resulting in an individual cDNA library for each cell. Although they differ in some aspects, they can be set up on a lab bench quite easily, requiring only syringes, automatic plungers and a microscope[58].

A commercially available droplet device is the 10X Genomics Chromium device[61]. On this platform a range of applications can be performed, including scRNA-seq for gene expression profiling as well as scATAC-seq for profiling of open chromatin in single cells. Furthermore, 10X provides additional support for sequencing analysis with the CellRanger software, an automated preprocessing pipeline. While droplet-based approaches feature similar throughput, Drop-seq has the lowest cell capture efficiency (3-4% of cells[58]) while inDrops and 10X Genomics have far higher efficiencies (65-70% of cells[61,60]), making these methods preferable if the number of suspended cells is limited. Furthermore, the use of

droplets increases throughput by at least an order of magnitude compared to protocols based on well plates or conventional microfluidics like Fluidigm C1, which is appealing for large-scale projects such as the Human Cell Atlas [46].

1.3.3 Preparing single-cell RNA-seq libraries

After successful cell capture, RNA is obtained and ready to be processed for library preparation and subsequent sequencing. Protocols consists of these three major steps (Figure 1.7A): reverse transcription of mRNA into cDNA, followed by amplification and subsequent final library preparation for sequencing, mostly on Illumina. Individual cells contain very small amounts of RNA. In order to obtain enough cDNA for sequencing, an amplification step by polymerase chain reaction (PCR) or in-vitro transcription (IVT) is necessary [62,63] (Figure 1.7B). Transcripts may be amplified at different rates by PCR which can distort their relative proportions within a library. In contrast, IVT is a linear amplification technique and therefore exhibits less amplification bias. In any case, many methods incorporate short random nucleotide sequences known as Unique Molecular Identifiers (UMIs) in the oligo-dT primers needed for the reverse transcription reaction. Furthermore, these primers usually also contain cell-specific barcodes to increase throughput [64,58,61]. This early barcoding also allows the pooling of reactions, saving reagent costs and labor time [65].

The addition of UMIs enables the removal of PCR duplicates introduced by library amplification which improves quantification of gene expression considerably [66,67] (Figure 1.7C), especially given that the random UMI sequences nowadays are long enough so that it is nearly impossible to capture two different transcript copies with the exact same UMI. On the other hand, because only the ends of each transcript can be tagged, library preparation methods with UMIs cannot achieve the full length coverage of protocols like Smart-seq2, which is for example needed for de-novo transcriptome assemblies [51]. Even so, sequences originating from transcript sections relatively far from the ends have been observed [68]. These could point to the presence of unannotated transcription start sites (TSS) or alternative polyadenylation. Given that genomic DNA is usually not depleted prior to reverse transcription, oligo-dT primers could also capture these sequences that contain enough adenine nucleotides. In any case, studies utilizing cellular barcoding and/or UMI tagging need extra careful processing:

For once, sequencing errors also occur in the UMI sequences and UMI library composition can be biased due to preferential amplification of certain barcode and UMI sequences [69,70].

As described for bulk RNA-sequencing, the amplified cDNA library is prepared for sequencing and the Illumina platform is again a popular choice, in combination with the Nextera kit for fragmentation and adapter incorporation [71].

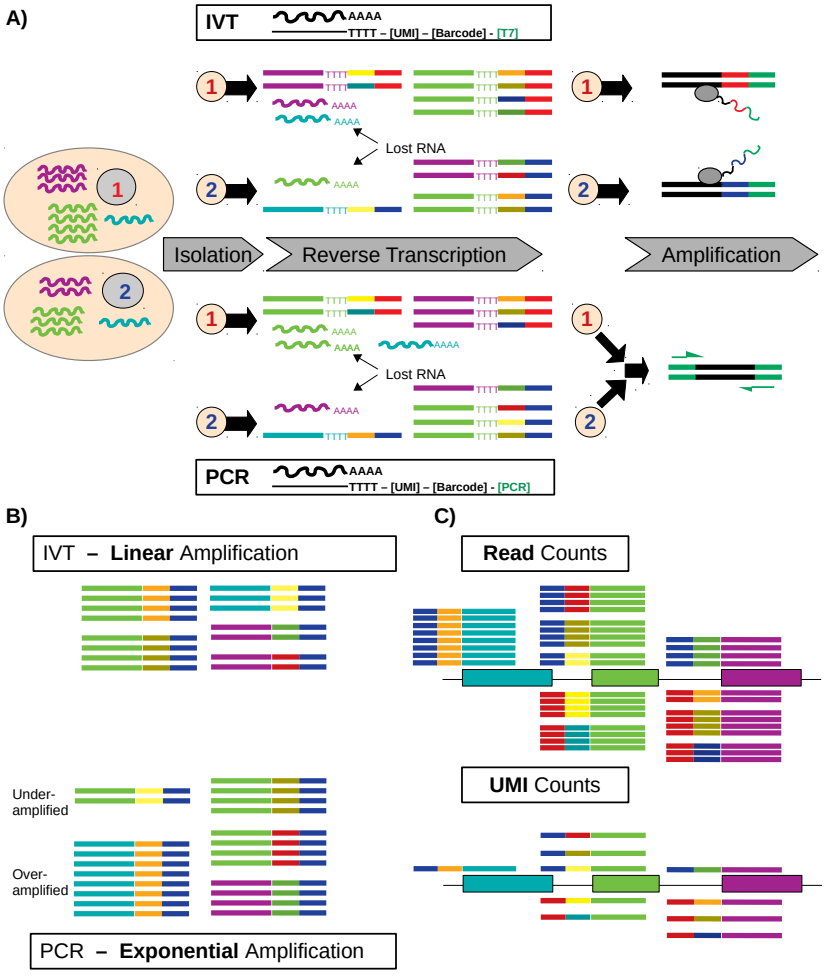


Figure 1.7. Preparation of scRNA-seq libraries.
A) Typical whole transcriptome amplification strategies are illustrated for two cells (top: IVT, bottom: PCR with early pooling). B) Amplified molecules for the blue cell (top: IVT, bottom: PCR). C) Alignment results (above: blue cell, below: red cell) for three genes (teal, green, violet). Upper panel: using PCR amplified sequence reads. Lower Panel: uniquely retained reads based on UMI sequences.

1.3.4 Analyzing single-cell RNA-seq data

The rapid pace of scRNA-seq protocol development has propagated to computational analysis methods of the resulting sequencing data. However, both parts have not converged to an optimum yet. Even though it is possible to define a rough standard processing workflow [72], the possibilities and choices of the computational analysis heavily depend on the research question as well as experimental setup. Here, I will particularly focus on the data acquisition and initial preprocessing of samples including data cleaning, quality control and normalisation to produce a cell by gene expression matrix that can be used to compare transcriptional profiles across cells in downstream analyses, which range from differential expression analysis (DEA), cell type identification to trajectory reconstruction.

Quantification of gene expression

After sequencing, the basic data processing for any scRNA-seq experiment involves 1) demultiplexing of cDNA reads using the Illumina indices and/or cell barcode nucleotide sequences, 2) alignment and 3) summarising expression by annotated feature, usually genes. These steps are similar to bulk analysis detailed above. Because cell barcode sequences are usually unknown or too many, probabilistic demultiplexer are unsuited for the task [30]. Therefore, demultiplexing usually relies on automatic detection and subsequent filtering of barcode lists by either providing a whitelist of expected barcodes or keeping top number of barcodes, e.g. Cell Ranger [61].

Due to the rapid increase in generated cDNA libraries as well as sequencing technologies, the classical alignment step by genome mapping can be computationally intensive and can take a significant amount of time [73]. Quite recently, pseudo-alignment has been implemented in a number of tools, e.g. kallisto [74] and Salmon [75], where the raw sequence reads are directly compared to the transcriptome de Bruijn graph containing k-mer transcript compatibility classes and then the transcript's abundance is quantified. These approaches are orders of magnitude faster than genome alignments while giving similar abundance estimates for full-length RNA-seq data [12,76]. However, the power to detect differential expression is not improved compared to STAR by e.g. a higher accuracy in gene expression estimates

[77]. Furthermore, it is unclear in how far pseudo-alignments are affected by the choice of annotation to build the k-mer index and whether short single end reads mainly covering the 3' or 5' end of the transcript are sufficient for this type of alignment, which is the standard for scRNA-seq protocols with UMIs. Another drawback is the lack of barcode demultiplexing capabilities of current tools.

An essential part of gene expression quantification of UMI data is deduplication. We define reads as duplicates when reads with identical UMI sequences align to the same position so that it is very likely that these reads are the product of PCR duplication instead of being true copies of a transcript [78]. Thus, deduplication methods need to separate out distinct UMIs. A number of tools have been developed when using conventional alignment. For example, UMI-tools [79] 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. A much simpler and hence faster approach is to apply a sequence quality threshold keeping only high quality UMI sequences [80,58]. Pseudo-alignment methods have also implemented add-ons to deal with UMI-based data sets, e.g. alevin implemented in Salmon [81]. However, these pipelines are limited in their flexibility, because they are usually expecting a particular read design (barcode and UMI sequence lengths) and/or are restricted to one alignment method.

As for bulk analysis, sequencing quality control can be performed on various levels including the quality scores of the reads themselves and how or where they align to features of the expression matrix [82].

Characteristics of scRNA-seq expression profiles

The result of the alignment is a matrix of counts: Rows are annotated features (genes) and columns are samples (single cells). The count values therefore show the expression level of a particular gene in a cell. This is per se not different than the data generated by bulk RNA-seq, but single-cell expression profiles have a number of unique characteristics that sets them apart from bulk RNA-seq and therefore attention must be paid in processing. Single-cell RNA-seq protocols have developed rapidly but the data they produce still presents a number of challenges. A major obstacle is the small amounts of starting RNA material. In

a mammalian cell, the total RNA molecules comes down to 10 to 20 pg, where only 1 to 5% of the total cellular RNA are mRNAs[83]. Existing approaches have greatly improved the conversion efficiency of mRNA into cDNA from initially 10 to 25 percent[63,80,84,51] to nearly 50 percent of transcripts[85]. Several studies have reported that small reaction volumes can increase the efficiency of this crucial step[64,55]. Nevertheless, low conversion efficiency is still the major bottleneck in terms of gene detection, particularly for lowly expressed transcripts [86] (Figure 1.7A). Increasing the sequencing depth can only alleviate this low sensitivity to a certain extend[82]. The small amount of starting material also contributes to high levels of technical noise as considerable amplification is needed for sequencing, complicating downstream analysis and making it difficult to detect genuine biological differences between cells[87,88,89] (Figure 1.7B).

In addition, scRNA-seq data sets are very sparse due to these limiting efficiencies usually coupled with a shallow sequencing depth, i.e. there are many instances where no gene expression has been measured. Naturally, this could of course be the true biological state but it could also be the result of confounders distorting the true expression profile of a cell: Differences in cell cycle stages, random transcriptional bursting or even "unwanted" environmental events can be seen as nuisance factors[90]. Technical factors also contribute to sampling noise by introducing so-called gene expression dropouts[86,91]. That said, dropouts and large variability in expression measurements are common phenomena in scRNA-seq studies that must accounted for in downstream analysis, as otherwise underlying assumptions of existing methods developed for bulk RNA-seq are violated[82].

Preprocessing

Quality control of cells is important as scRNA-seq experiments will contain poor-quality cells that can be uninformative or lead to misleading results. Particular types of cells that are commonly removed include damaged cells, doublets where multiple cells have been captured together[92,93] and empty droplets or wells that have been sequenced but do not contain a cell[94]. The identification of cell outliers is usually done manually by visualizing various of the data set, e.g. total number of detected genes per cell, percentage of reads allocated to spike-in transcripts or mitochondrial genes. Given the distribution of metrics, one can then

apply thresholds to filter outlying cells out, either as hard cutoff or using median absolute deviations like the scater package[95]. Besides the quality control of single cells, there is also the issue that a large number of genes or transcripts are lowly expressed resulting in very sparse counts. These features are typically removed. A number of downstream analysis steps, e.g. dimension reduction, are applied to a count matrix considering only highly variable genes. However this filtering might actually exaggerate unwanted technical noise rather than true biological differences. Selecting genes using the Fano Factor as a measure of variability can also remove marker genes of cell types with low abundances in complex cell mixtures[96]. Given the sparsity of gene expression measurements, an alternative has been implemented in M3Drop where biologically relevant features are identified as outlying measurements of averaged expression level in relation to dropout rate across heterogeneous cell populations[97].

Normalisation

Normalisation is a very important step in any RNA-seq experiments, for bulk as well as single cells[98,99,90,100]. First and foremost, it is necessary to correct for variation in the sequencing depth per library. Classical normalisation methods achieves this by a simple division by total read counts yielding Counts Per Million (CPM). Methods incorporating gene length correction such as Fragments Per Kilobase Million (FPKM) or Transcripts Per Kilobase Million (TPM) transformations can be used. On the other hand, libraries prepared with UMIs do not require this gene length correction since the transcript ends are mainly sequenced[68]. However, these methods assume equal amounts of RNA per sample and a balanced up- and downregulation, so that the total mRNA content is comparable among samples[98]. These assumptions are almost always violated in single-cell data. Firstly, RNA amounts vary considerably from cell to cell[101], especially in complex tissues[102]. Secondly, technical variance in combination with biological variation (e.g. transcriptional bursting) contributes to the high frequency of zeroes and strong intercellular variability in scRNA-seq data[103,99,101]. Therefore, cell-wise size factors such as weighted trimmed mean of M-values (TMM)[104] or median of ratios (MR)[105] are biased[100].

Awareness of the above issues has led to the development of normalisation methods that are geared towards single cells. scran solves the zero inflation issue by pooling cells and

then deconvoluting to obtain cell-wise size factors[99]. SCNorm applies a quantile regression for bins of genes with similar mean expression to estimate gene-wise size factors[106]. Both appear to be able to handle the zero inflation as well as large differences in mean expression between groups. CENSUS attempts to estimate absolute RNA levels from relative expression measurements (TPM, FPKM)[84]. The underlying model has certain assumptions concerning amplification bias and capture efficiency which have been derived from a small set of experiments. Therefore, the derived parameters may or may not be applicable to one's own data.

Methods operating in a group-aware way - be it by *a priori* clustering (e.g. scran) or known cell type annotation (e.g. SCNorm) - result in more reliable size factor estimates, also for very heterogeneous cell populations with strong expression differences[106,99,107,100]. In theory, extrinsic spike-in RNA molecules such as the widely used External RNA Controls Consortium (ERCC)[37,38] allow the decomposition of observed cell-to-cell variability into technical noise and actual biological factors[108,63,109,110,111,112]. More importantly, spike-ins are the only option to also estimate differences in total mRNA content among cells. However, ERCCs have a number of limitations as it is unclear how well they mimic nascent mRNA molecules as they are purified, shorter than the average transcript, have shorter poly-A tails and their concentration ranges deviate from in vivo transcript abundances[111,113]. Even if spike-ins properly capture the underlying dynamic, their usage is restricted to protocols where they can be added which does not include droplet-based capture techniques, yet. These shortcomings should be addressed in future generations of spike-in mRNAs and will likely improve normalisation[114].

Integration of single-cell RNA-seq data

While earlier studies have mostly only quantified the gene expression of single cells derived from an individual sample, nowadays studies are common which profile single cells originating from multiple batches, e.g. scRNA-seq experiments conducted by different labs[115], or individual patients in clinical studies[116]. In those cases data integration becomes essential to ensure comparability[117]. There are already a number of computational approaches available which were initially developed for bulk RNA-seq, e.g. ComBat[118], RUV[111] and limma[119].

One drawback of these methods is the assumption that cellular composition of the sample is the same, e.g. when aliquots of the same cell mixture is processed for sequencing in different labs[120,115].

Given the increased efforts to chart the cellular composition of whole tissues or even organisms, tools have been developed to not only correct for technical batches but also allow the integration of these diverse data sets. For example, mnnCorrect implemented in scran package[99] utilizes a mutual nearest neighbors (MNN) approach where the cosine distance between cells originating from different data sets functions as a measure of similarity to identify cells belonging to the same neighborhood[121]. Another prominent example is implemented in the Seurat package. Here, canonical correlation analysis is carried out in combination with Dynamic Time Warping to align the different data sets in a shared multi-dimensional subspace[122]. Other examples for batch correction and data integration for single-cell RNA-seq include Scanorama[123], scMerge[124] and BBKNN[125]. With the exception of Scanorama, the majority of methods is unable to correctly integrate data in a scenario with dataset-specific cell types (Janßen et al. 2019, unpublished). Ultimately, further comparisons are needed to assess the general applicability and performance of data integration and batch correction methods for single-cell RNA-seq experiments[126,127].

Identification of cell types and states

Bulk RNA-seq experiments usually involve predefined groups of samples, for example diseased and healthy tissue cells, different tissue types or treatment and control groups[29]. It is possible to design scRNA-seq experiments in the same way by sorting cells into known groups based on surface markers, sampling them at a series of time points or comparing treatment groups, but often single-cell experiments are more exploratory, e.g. profiling cell types in tissues such as the mouse retina[58] or cortex[102]. In fact, there are now a number of Cell Atlas projects attempting to produce a reference of the transcriptional profiles of all the cell types in an organism (e.g. human[46], mouse[128,129], *C. elegans*[130] and flatworm[131]).

Identifying similar cells in complex tissues, along lineages or differentiation paths is therefore an integral step in analyzing these data sets and as such, it has been a key focus of methods development with over two hundred tools released so far[132]. There are a number of

unsupervised methods available to identify cell types by grouping, e.g. single-cell Consensus Clustering (SC3)[133], BackSPIN[102] and Seurat[122] developed specifically for single cells, but also general purpose classifiers like general Support Vector Machines (SVM)[134]. The selection of parameters in these unsupervised methods is difficult and can influence the interpretation of results, esp. the number of cell clusters selected[134,135,136]. An alternative is the classification of cells using comprehensive references, e.g. scPred[137], scmap[138], scMatch[139] and SingleR[140]. Using references has the advantage of building on existing, usually well curated markers to quickly identify cellular identities. On the other hand, since the classifier is trained on the reference, it is naturally biased towards the composition of the reference, making the identification of previously unannotated cell types or states difficult[134]. But given the ongoing efforts of atlas projects, these references will improve in terms of completeness and reliability.

Besides the assignment of discrete cellular identities, there are also studies which focus on ordering cells along a continuous trajectory of cellular types, e.g. the differentiation of stem cells[141]. The methods usually rely on a dimensionality reduction technique such as principal component analysis (PCA). This simplified representation is then used to define a graph by e.g. minimal spanning tree (MST) through which a path is determined and the cells are ordered along this continuous trajectory. A recent benchmarking study revealed that the performance of methods depends on the underlying topology of the data and that multiple complementary approaches should be used to infer a robust and comprehensive trajectory[142].

Deciding on which cell assignment approach to use depends on the cellular composition, research questions and experimental design. Both approaches, continuous ordering as well as assignment of distinct cell types, can be informative.

Differential gene expression

After assigning the identity of single cells by prior knowledge, ordering or clustering, the analysis naturally focuses on the identification of differences between groups of cells. For FAC-sorted and clustered cells, this is usually done by identifying genes that are differentially

expressed between the groups or marker genes that are characteristic for a single cell type identity. An important first step in differential analysis is defining an appropriate distribution to allow reliable inference of expression differences. Given that the majority of quantification methods ultimately result in a matrix of counts, common discrete distributions can be considered. One option is the Poisson distribution which describes here the probability of sampling species of RNA out of a pool of RNA molecules at random. A Poisson distribution is a one parameter distribution where the mean of the distribution is equal to the variance, thus accounting for sampling noise only. However, counting noise is not the only source of variance in RNA-seq experiments as technical and biological noise can add additional variance [143]. A better fit is the negative binomial distribution including an extra over-dispersion parameter, allowing the variance in expression to be larger than the mean [144]. Another way of defining the NB distribution is that it is a weighted mixture of Poisson distributions where the rate parameter (i.e. the expected counts) is itself associated with uncertainty following a Gamma distribution, called a Poisson-Gamma mixture distribution. The negative binomial distribution also fits the accepted bursting model of gene expression where transcription can be described in a two-state model, so called molecular-ratchet model, where patterns of gene expression are governed by on- and off-states of genes as well as waiting times between consecutive transcription initiation events [145,146].

Already established methods for the detection of differential expression in bulk have also been applied to scRNA-seq data [69] since the negative binomial distribution has been found to fit the observed read count distribution for the majority of expressed genes in single cells [63]. However, the early analysis of scRNA-seq data might have been limited by filtering to conform to the tools requirements (e.g. minimum mean expression cutoff). Furthermore, concerns have been raised due to the observed differences in distributional characteristics between bulk and single cells, namely dropouts, high variability and outliers [147,86], which might violate the model assumptions of bulk methods. This drove the development of specialized tools for scRNA-seq data. SCDE was one of the first methods addressing the zero count inflation by applying a mixture model of the negative binomial and Poisson distribution and robustifying the estimation in the presence of strong overdispersion by bootstrapping [103]. BPSC and D3E are other examples of mixture modeling approaches whereby a beta-Poisson mixture is used

to capture the bimodality of scRNA-seq expression profiles[148,149].

Instead of mixing distributions to match the observed expression patterns as closely as possible, there are also other possibilities to cope with the excess of zeroes. For example MAST incorporates a two-part generalized model by applying a hurdle model[150]. The first step is to fit the expression rate, i.e. zero vs. larger counts, as a logistic regression and conditioning on the resulting probability, the mean gene expression is modeled as a Gaussian distribution. It is also possible to identify genes that might have the same mean between groups but differ in variance[151] or other characteristics of their expression distribution, e.g. difference in non-zero fraction across cells[152]. On the other hand, standard statistical tests such as Student's t-test or Mann-Whitney U test that had been unsuitable for bulk RNA-seq experiments due to the small number of samples, have been used to identify marker genes in single cell populations[94]. Nevertheless, these statistical tests are limited to pair-wise comparison and cannot accommodate complex experimental designs nor correct for unwanted variation, esp. batch effects.

The detection of differential expression is an essential step in many scRNA-seq experiments but it has been unclear which modeling and testing framework is suitable for scRNA-seq data. The power simulation framework that I established during my PhD has contributed to solving this question[153,154,155].

Imputation

Another approach to tackling the problem of too many zeros is to use zero-inflated versions of count distributions for dimensionality reduction (e.g. ZIFA[156]), factor analysis (e.g. ZINB-WaVE[157]) or differential expression testing (e.g. DEsingle[158], zingeR[158,159]). However, there is still an ongoing debate whether single cell RNA-seq data are truly zero-inflated and therefore, if there is even the need to include zero-inflation in the modeling[160,161,162].

Imputation of zeros has recently received considerable attention as an alternative strategy to compensate for the sparsity of scRNA-seq[163]. For instance, there are a number of methods that aim to identify which observed zeros represent true technical rather than biological zeros using probabilistic models and impute the missing data accordingly. Examples include SAVER[164], bayNorm[165] or scImpute[166]. Data-smoothing methods on the other hand adjust

all gene expression levels based on expression in similar cells, denoising the whole profile, e.g. MAGIC[167], netSmooth[168] or DrImpute[169]. However, one major problem of these methods is that they solely rely on internal information for imputation. This circularity can lead to the introduction of false structures and inflated correlations between genes and cells that are actually not present in the samples[170]. One way to circumvent this issue is to explore complementary types of data that can inform imputation. There are methods that incorporate external information, e.g. SAVER-X[171] uses atlas-type resources in a transfer deep learning method or URSM[172] that borrows information from matched bulk RNA-seq data.

That said, imputation has also been put forward as means to smooth and normalize gene expression profiles across cells[173] which in some cases has shown to improve the reconstruction of cellular differentiation processes[163]. As part of my work in this thesis, we therefore investigated in how far imputation can improve normalisation and how much this change contributes to the overall performance of scRNA-seq analysis pipelines[155].

1.4 Experimental design and power analysis

R.A. Fisher formalized three integral principles needed for a sound experimental design [174]: 1) replication, 2) randomization and 3) blocking. While these factors are essential for any successful experiment and subsequent statistical analysis, their application to high throughput sequencing of RNA have not been straightforward and often neglected, which may lead to incorrect conclusions in scRNA-seq experiments[91].

Experimental design choices concerning replication for genome-wide expression profiling were discussed in the context of microarray studies[175] and updated in the context of RNA-sequencing[176]. While high-throughput sequencing technologies have led to a considerable decrease in costs[177], it is still the biggest cost factor for most expression studies. Hence, one needs to decide on a trade-off between the number of samples and the read depth per sample. Importantly, the majority of studies, particular differential expression analyses, benefit more from more replicates than deeper sequencing[178,179] (Figure 1.8A). Adhering to the blocking principle in RNA-seq experiments is most frequently violated, e.g. technical

limitations require the separation of an experiment into processing batches, which introduces a source of technical unwanted variation (Figure 1.8B). To avoid confounding the experiment by these nuisance factors, blocks of batches should be constructed in such a way that samples per conditions are evenly and preferably randomly spread across these blocks [176,180] (Figure 1.8C).

In the case of single-cell RNA-seq experiments, it is possible to decouple the location and time of sampling from further downstream library preparation by cryopreservation or methanol fixation of cells after FAC-sorting [181,182]. For droplet-based methods, a number of techniques have been developed to enable multiplexing cells from different samples, e.g. patients, on one single processing run [183,184]. Early cellular barcoding by transfection [185] or during cDNA library preparation [186,65] allows pooling of single cells, reducing not only costs but also the number of processing batches per experimental condition [71]. As long as the experimental design is balanced, batch effects can be included as an additional nuisance covariate in differential expression modelling [187], which comes with a loss of degrees of freedom but ensures that biological and known technical effects on gene expression can be distinguished.

Therefore, the use of biological replicates, random assignment of samples and a balanced block design are essential factors underlying any successful scRNA-seq experiment and subsequent statistical analysis [188]. In addition to these experimental design considerations, the choices made for further downstream analysis methods and tools are also important in terms of possible artifacts or biases influencing scientific conclusions.

1.4.1 Evaluation of single-cell RNA-seq methods

The development of scRNA-seq technologies over the last decade has been staggering. Nowadays, one can choose from multiple methods for every experimental step, particularly for computational data analyses. Deciding which one to use can be difficult and depends on a number of factors. These range from expression data generation aspects covering capture efficiency of RNA molecules, accuracy and precision of expression measurements by the library preparation protocol to computational aspects including performance, scalability, accessibility and suitability. Computational methods usually show their effectiveness by demonstrating

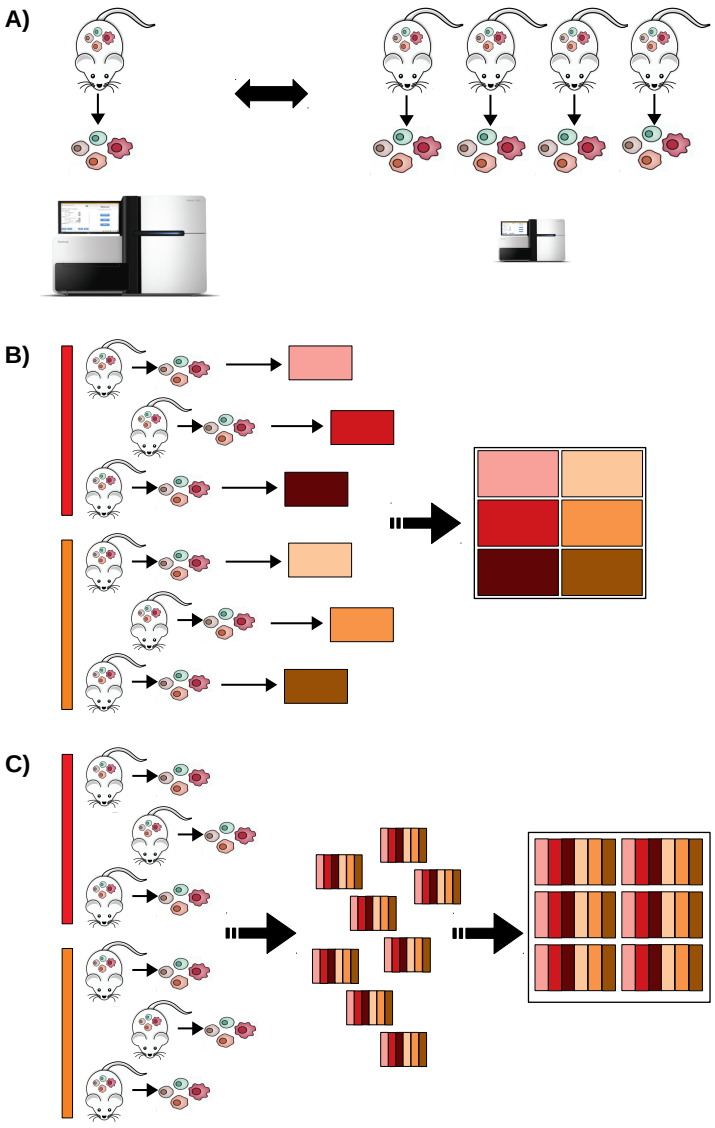


Figure 1.8. Experimental design for scRNA-seq experiments.
A) Few biological replicates with deep sequencing versus many biological replicates with shallow sequencing. B) Confounded design: Cells are isolated from each biological replicate per condition at potentially different times onto separate plates. Prepared libraries are sequenced on separate lanes of the sequencer. C) Balanced design: All samples are evenly distributed across all stages of the experiment, thus reducing the sources of technical variation in the experiment.

their performance for a particular task and that they are at least as good as existing methods. Delivering this proof can be difficult using real data sets alone because the underlying truth is almost always unknown. By generating samples where the true differences are known or have been validated by control measurements[12,189], one can construct a so-called gold standard. However, these standards are usually difficult to produce and inherently can only cover a limited number of possible scenarios. Furthermore, only recently a benchmarking study has provided such a sample mixture of single cells[190].

On the other hand, computer simulations are in principle limitless in their ability to capture real-world scenarios. For bulk RNA-seq, a number of tools have already been developed. These range from simulations assuming an underlying probability distribution for gene expression, e.g. Negative Binomial[191,192] to resampling approaches using observed gene expression profiles[193,194]. Synthetic data sets produced by simulations have already been particularly useful for method development and evaluation of bulk RNA-seq experiments [189,194,195]. Due to these reasons and the aforementioned lack of gold standards, *in silico* simulations based on parametric distributions have also been the approach taken by many early methods for scRNA-seq analysis. Unfortunately, these methods suffered from a lack of documentation and therefore reproducibility. In addition, these simulation frameworks were limited to synthetic data generation alone without the possibility to evaluate experimental designs.

1.4.2 Statistical Hypothesis Testing and Errors

Each hypothesis test results in a binary decision: Accept or reject the null hypothesis stating in the case of differential gene expression analysis that there is no difference in gene expression between groups. When combined with the binary truth, four distinct outcomes result, including correct and wrong conclusions (Figure 1.9A). There are two types of error involved[196]. Type I error or false positivity means that one rejects a true null hypothesis. On the other hand, Type II error or false negativity is the failure to reject a false null hypothesis.

These decision errors happen with a particular probability or "rate". A statistical test is constructed such that it aims at controlling the type I error rate at a specified level, which is referred to as the nominal significance level and denoted by α. The test typically succeeds

in controlling the type I error rate at its nominal level if all assumptions underlying the theoretical construction of the test are fulfilled. The type II error rate, which is denoted by β, is generally not controlled by the test. This error rate depends on the type I error rate, the effect size, the variability in the data and the sample size. It is more convenient to work with the power of a test, which is simply defined as 1 - β. For a given α, effect size, and variance, the sample size can be calculated so that the statistical test also controls $\beta$and hence a certain level of power is ensured.

A test with a low type I error rate is called specific, while a low type II error rate is called sensitive. There is a trade-off between these two types, decreasing one type results in an increase of the other[196] (Figure 1.9B). The standard procedure is to achieve an exact type I error control, meaning that the true error rate should not exceed the prespecified nominal level while minimizing the type II error rate so as to maximize the statistical power $(1 - \beta)$[197].

Multiple Testing Problem

In the case of gene expression data, one is interested in determining the association for a large number of features with the outcome, e.g. response to drug treatment. Assuming that the multiple tests are independent, the probability of making at least one Type I error is equal to 1-(1-α)^m, where m is the number of tests[198]. The impact of this multiple testing problem (MTP) on the error rate has to be accounted for in the decision rule.

In the case of a single hypothesis test, the type I error rate is clearly defined. On the other hand, when moving to multiple hypothesis testing, the definition of type I error rate is not straightforward any more[199]. There are two widely used frameworks that extent these concepts to multiple testing: The Family-Wise Error Rate (FWER) and the False Discovery Rate (FDR)[200,201]. Both aim to control the probability of rejecting a true null hypothesis below the nominal level of the decision rule. The methods actually adjust and combine the raw p-values of the individual tests[197]. For example, controlling FDR at level αmeans that the set of rejected hypotheses is chosen in such a way that the false discovery rate is at most α. Benjamini and Hochberg defined the most well known FDR procedure as the expected proportion of rejected null hypotheses which are falsely rejected among all rejected

A)

Test Outcome	Actual Outcome: Null (EE gene)	Actual Outcome: Alternative (DE gene)	Total
Accept H_0	True Negative	False Negative **Type II Error (β)**	m - R
Reject H_0	False Positive **Type I Error (α)**	True Positive	R
Total	m_0	m_1	m

B)

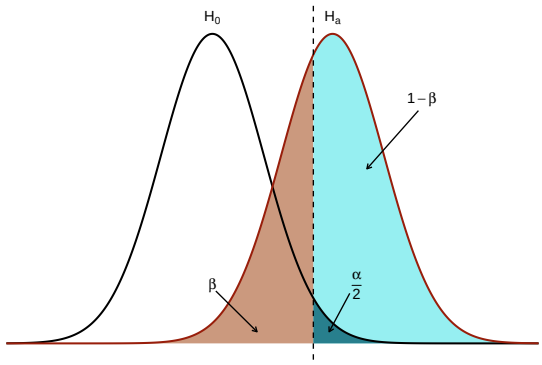


Figure 1.9. Statistical Hypothesis Testing and Errors.
A) Confusion matrix tabulating the four possible outcomes when testing m hypotheses. B) Trade-off between Type I and Type II Errors.

hypotheses conditioning on the number of rejections [202].

The control over the error can be strong or weak [203]. Strong control means that the rate is correctly controlled independent of the proportion of true and false null hypotheses, whereas weak control is only achieved if all null hypotheses are true [198,197]. The available methods also have differing levels of conservatism, i.e. controlling the rate in such a way that less hypotheses are rejected. The loss in sensitivity is greater, i.e. a decreased chance of identifying true differences between treatment groups, when the multiple testing procedure is more conservative. The preference of strong control with moderate conservatism are also the main reasons why the FWER has been basically abandoned in favor of FDR procedures for high-throughput sequencing data analysis.

1.4.3 Statistical Power Analysis for RNA-sequencing experiments

The statistical power of a hypothesis test is the probability to correctly reject the false null hypothesis in favor of the true alternative hypothesis [196]. There are a number of factors that determine the power level in a general univariate test: alpha level, one versus two-tailed test, sample size and effect size. Conventionally, the aim is to achieve a power of 80%. In the case of gene expression data, one needs to extend the definition of power to accommodate the large number of statistical tests performed. As with false positives, the concept of false negatives and therefore power can be defined in multiple ways. The average power, i.e. $E(TP)/m_1$, is commonly employed (Figure 1.9A). There is also the option to consider the global power, i.e. probability of rejecting at least one null hypothesis $Pr(TP \geq 1) = Pr(FN \leq m_1 \breve{} 1)$ [198,204].

A number of sample size estimators and power analysis tools developed for bulk RNA-sequencing are available [205]. The majority rely on pilot or publicly available data to derive parameters for count data simulations based on parametric distributions (e.g. PROPER [206], Scotty [207]) but the count simulations usually assume a constant technical variance per gene resulting in unreliable power calculations for lowly and variably expressed DE genes [205]. Furthermore, some of these tools base their power calculations on testing results of certain DE-tools assuming an ensured FDR control. They also do not consider other steps of the computational pipeline like normalisation in their calculations.

At its core, sample size calculations are an important step in conducting research since it is essential to ensure sufficient statistical power for detecting anticipated effects. In addition, journals are increasingly adopting methods such as reporting statements concerning experimental design plans (e.g. Nature Neuroscience[208]) and editors emphasize the need for thorough experimental design plans in preclinical trials[209]. However, there is a considerable lack of tools for power analysis and sample size calculations for scRNA-seq experiments.

As part of my thesis, I developed the first tool for statistical power analysis and sample size calculations of bulk and single-cell RNA-seq experiments[154] using simulations closely resembling observed gene expression profiles[153,210].

2 | Results

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

Parekh S, Ziegenhain C, **Vieth B**, Enard W, Hellmann I:

"The impact of amplification on differential expression analyses by RNA-seq." (2016)

Scientific Reports 6 (25533).

doi: 10.1038/srep25533

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

https://www.nature.com/articles/srep25533#Sec18

www.nature.com/scientificreports

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 sequenced[6]. 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-Seq[12]. 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 transposase[12]. 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-amplification[16]. 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)

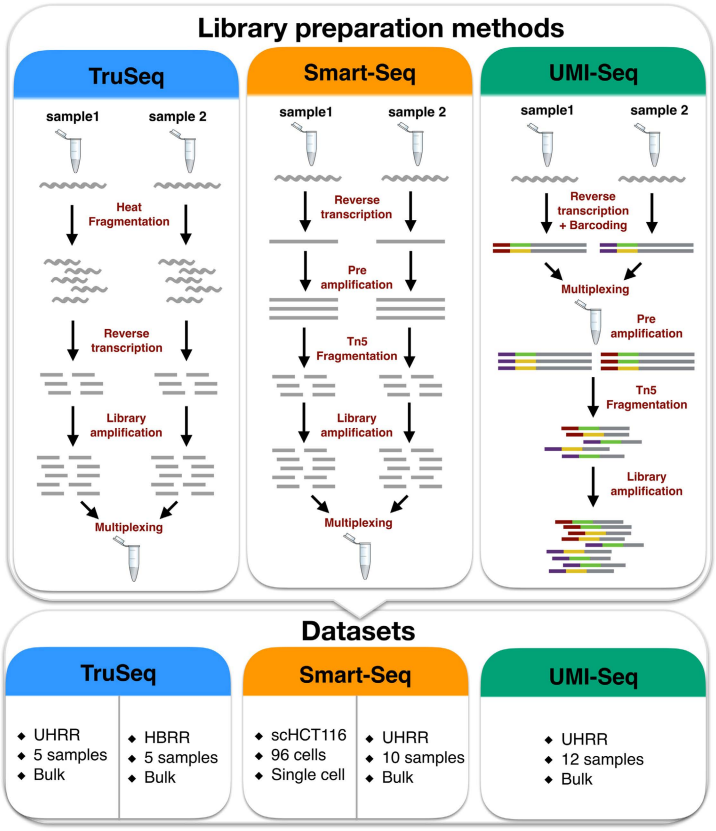


Figure 1. Schematic of library preparation protocols and datasets. The upper panel details the steps for the three sequencing library preparation methods analysed in this study. In the UMI-seq flow-chart red and purple tags represent the sample barcodes and the green and yellow tags the UMIs.

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 PROPER[19].

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 literature[2] 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 ± SD million)	Read Length	PCR cycles
scHCT116 Smart-Seq	GSE51254	Quake	96	1.8 ± 1.1	101	21* + 12
UHRR Smart-Seq	GSE75823	Enard	10	1.5 ± 1.1	50	10* + 12
UHRR UMI-seq	GSE75823	Enard	12	9 ± 1	46	15* + 12
UHRR TruSeq	GSE49712	SEQC	5	125 ± 33	101	15
HBRR TruSeq	GSE49712	SEQC	5	140 ± 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

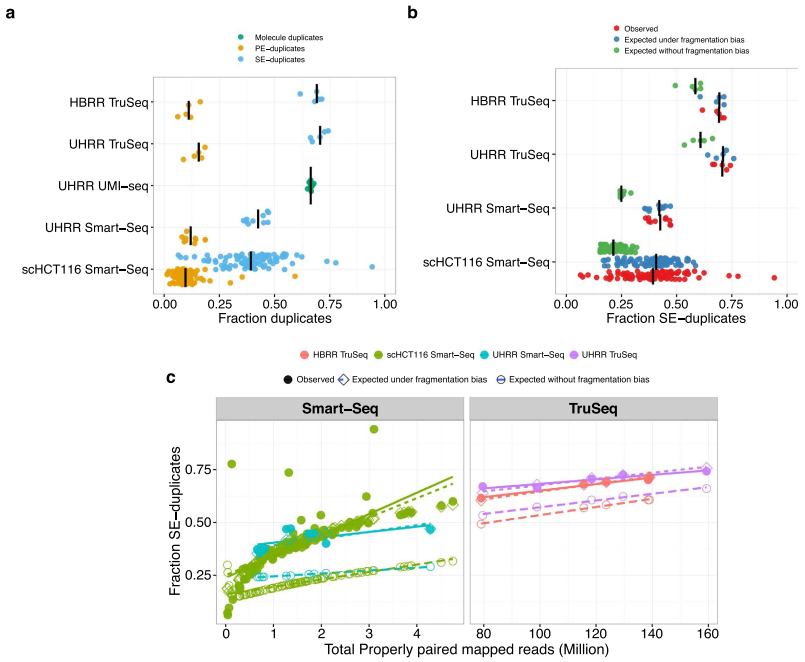


Figure 2. The Fraction of SE-duplicates increases with the total number of reads. In panel (**a**), we plot the fraction of computationally identified SE-duplicates (blue) and PE-duplicates (yellow) per sample. For the UMI-seq data, we identify duplicates only based on the experimental evidence provided 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.

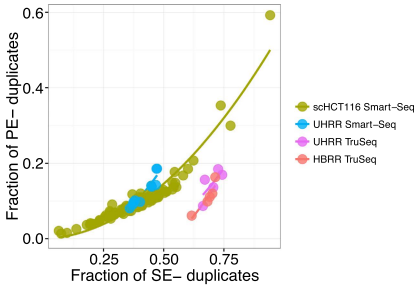


Figure 3. The relation between SE- and PE-duplicates. The relation between SE- and PE-duplicates is expected to follow a quadratic function, if the majority of duplicates are natural, i.e. due to fragmentation and sampling. Here, we show a quadratic fit for the different datasets (UHRR-TruSeq–purple,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

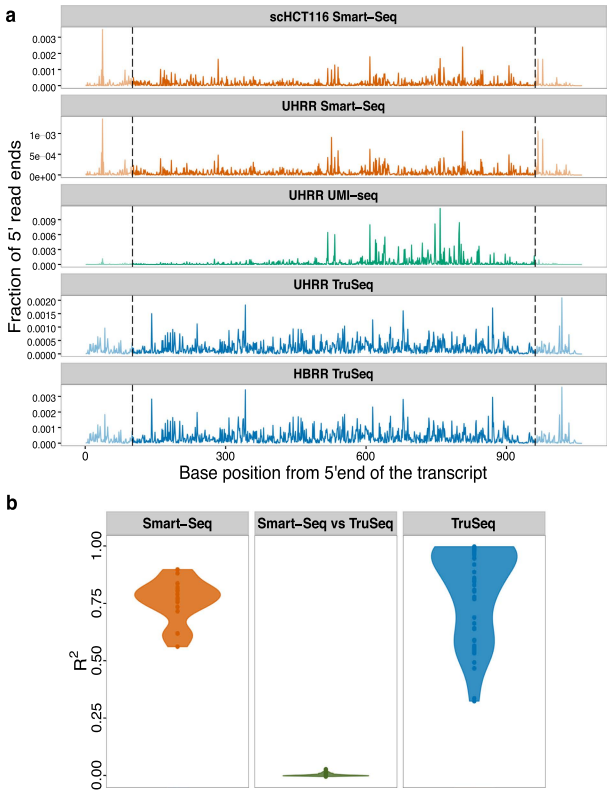


Figure 4. The fragmentation patterns of the ERCCs are highly reproducible for different samples prepared with the same RNA-seq library method. (**a**) Here, we plot the fraction of 5′ read ends per position of ERCC-00002. Because the TruSeq libries (blue) had read lengths of 100 bases, we do not consider 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^2s 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.

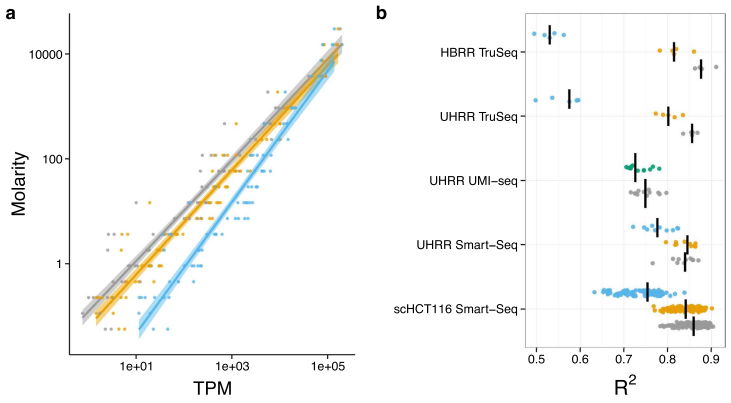


Figure 5. Removing duplicates does not improve the accuracy of expression quantification as measured using the ERCC spike-ins. Expression levels as quantified in transcripts per million reads (TPM) are a good predictor of the concentrations of the ERCC spike-ins. The log-linear fit of TPM vs. Molarity 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 Transposase[21] 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 ~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

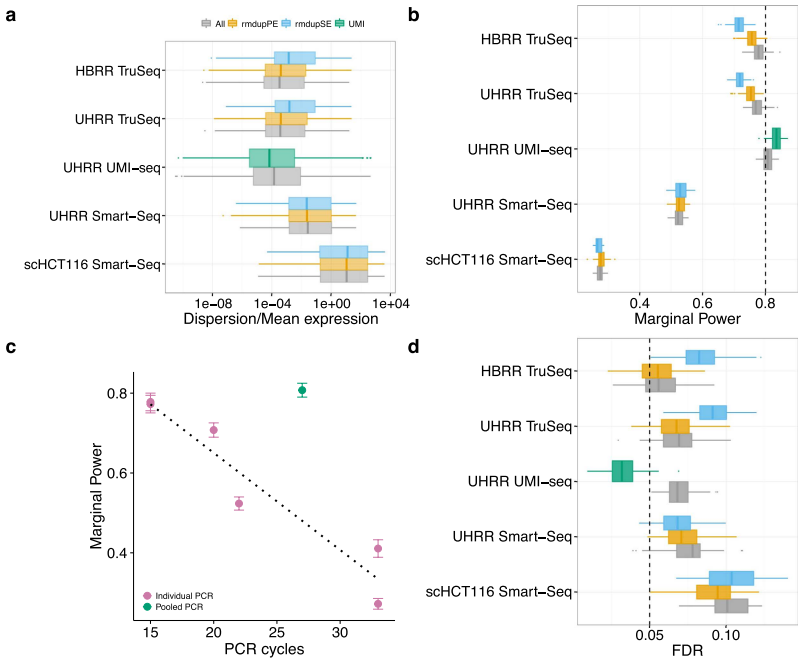


Figure 6. Duplicate removal has little influence on the power and FDR to detect DE-genes in comparison to the library preparation method. We estimated the distributions of mean expression and dispersion across genes for each dataset using DESeq2 after downsampling the datasets to 3 or 1 million reads. 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 PROPER[19]. 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 DESeq2[22].

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 log 2-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 log 2-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 artifical 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/μl). The recommended ERCC spike-in amounts result in 5–10^7 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-III[2]. 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 neccessary and therefore we used NextGenMap for the ERCCs[30]. Except for three parameters, (1) the maximum fragment size which was set to 10 kb, (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 Rsubread[31] (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 NextGenMap[30] (Supplementary text). If either the sample barcode or the UMI had at least one base with sequence quality $\leq$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 ($L_{eG} = L - 2 \times$ read-length). Then the expected number of unique reads can be estiamted as

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

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 $s = 1/\sum_{k \in 1 \ldots r_G} P(X = k)$.

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 downstream[21]. The resulting PWM was then used to calculate motif scores across the ERCCs using the Bioconductor package PWMEnrich[32].

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 DESeq2[22] 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 PROPER[19]. 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).

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

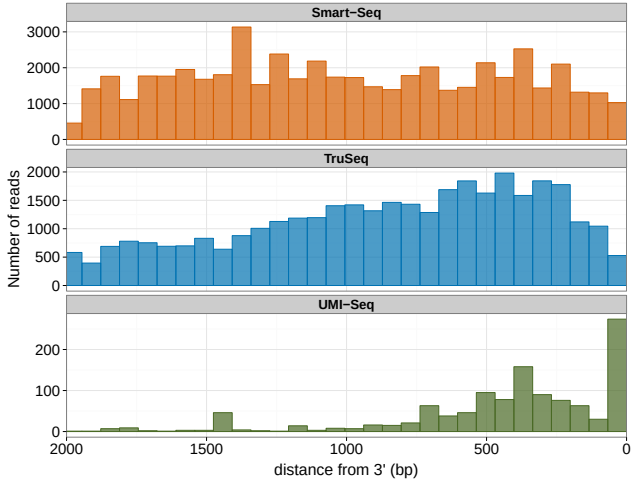


Figure S1: 3' bias in fragmentation site is prominent in UMI-seq. The histogram showing distance of the fragmentation site from 3' end of the gene measured from ERCC spike-ins of length $\sim 2kb$. Colors represent library preparation methods, 'blue' - Smart-Seq, 'orange' - TruSeq, 'green' - UMI-seq.

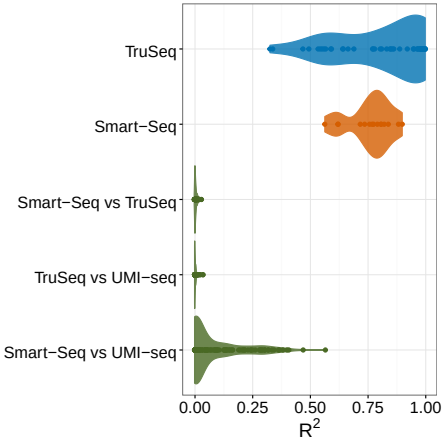


Figure S2: The fragmentation patterns of the most 3' 600bp of ERCCs are relatively reproducible between Smart-Seq and UMI-seq. Violin plots of the adjusted R^2 from a linear model between fraction of 5'read ends from different samples. The adjusted R^2 are calculated considering full length for Smart-Seq 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.

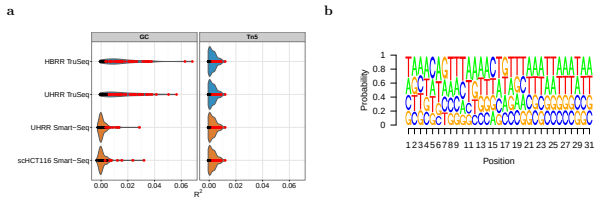


Figure S3: Fragmentation does not appear to have a cutting site preference. Colors of the violin plots represent library preparation methods, 'blue - Smart-Seq, 'orange' - TruSeq and dots are colored by the significance of the fit where 'red' - pvalue ≤ 0.05 and 'black' - pvalue > 0.05. **a)** The left 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.

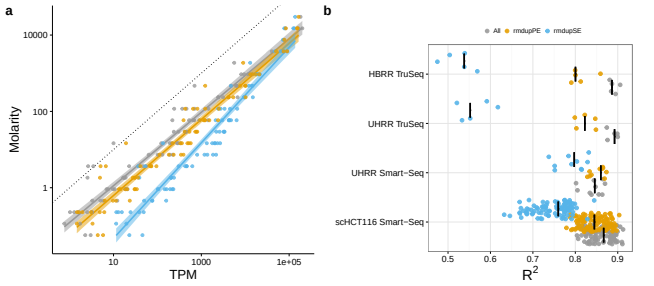


Figure S4: Removing duplicates does not improve the accuracy of expression quantification as measured using the ERCC spike-ins. Expression levels as quantified in transcripts per million reads (TPM) are considered to be good measure of ERCC spike ins. However, other factors like capture and sequencing 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).

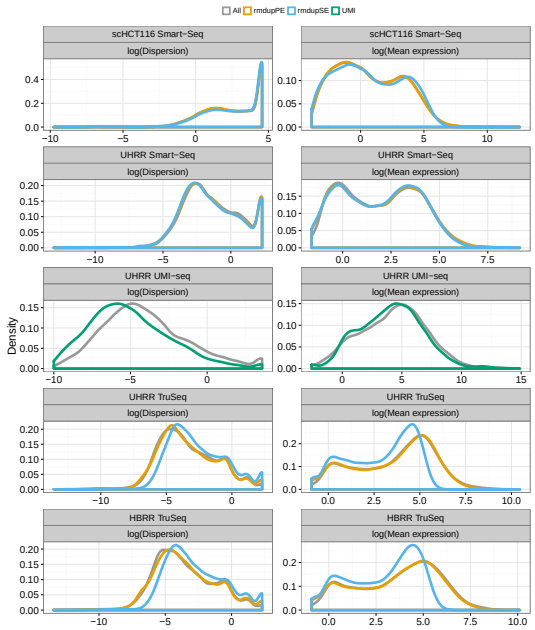


Figure S5: Empirical mean and dispersion distributions are used to estimate power to detect differential expression. The left panel shows density plot of log(dispersion) and the right panel the log(mean baseline expression) measured by DESeq2 for each study. Different duplicates treatments are represented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

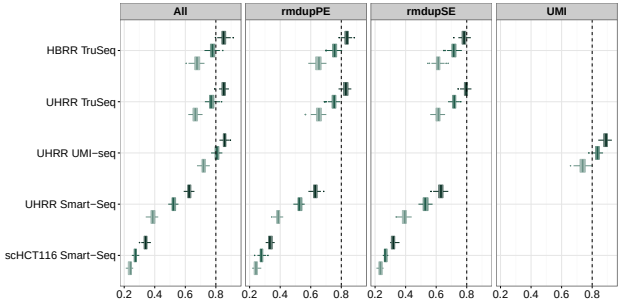


Figure S6: Power to detect differential expression increases with increased sample size. The boxplot shows marginal power to detect 0.5 log2foldchange at 5% nominal FDR for different sample sizes. Colors gradient from light to dark represent sample sizes 3,6 and 12 for the bulk and 30,45 and 90 for the 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-dup ∼ SE-dup+(SE-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 R^2 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(ng)
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	rmdupPE	15	1.00
HBRR TruSeq	3	0.07	0.61	266.98	20.45	0.08	rmdupSE	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	rmdupPE	15	1.00
HBRR TruSeq	6	0.08	0.72	315.48	31.46	0.10	rmdupSE	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	rmdupPE	15	1.00
HBRR TruSeq	12	0.07	0.78	352.17	30.74	0.09	rmdupSE	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	rmdupPE	33	0.00
scHCT116 Smart-Seq	30	0.15	0.23	211.20	37.70	0.18	rmdupSE	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	rmdupPE	33	0.00
scHCT116 Smart-Seq	45	0.10	0.27	251.00	29.35	0.12	rmdupSE	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	rmdupPE	33	0.00
scHCT116 Smart-Seq	90	0.07	0.32	308.55	22.75	0.07	rmdupSE	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	rmdupPE	22	0.25
UHRR Smart-Seq	3	0.05	0.39	283.54	15.46	0.05	rmdupSE	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	rmdupPE	22	0.25
UHRR Smart-Seq	6	0.07	0.53	398.36	30.53	0.08	rmdupSE	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	rmdupPE	22	0.25
UHRR Smart-Seq	12	0.06	0.63	481.09	32.36	0.07	rmdupSE	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	rmdupPE	15	1.00
UHRR TruSeq	3	0.08	0.61	316.45	30.10	0.10	rmdupSE	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	rmdupPE	15	1.00
UHRR TruSeq	6	0.09	0.72	375.37	41.36	0.11	rmdupSE	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	rmdupPE	15	1.00
UHRR TruSeq	12	0.08	0.79	412.77	39.44	0.10	rmdupSE	15	1.00

2.2 Comparative Analysis of Single-Cell RNA Sequencing Methods

Ziegenhain C, **Vieth B**, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W:

"Comparative Analysis of Single-Cell RNA Sequencing Methods." (2017)

Molecular Cell 65 (4): 631–643.e4.

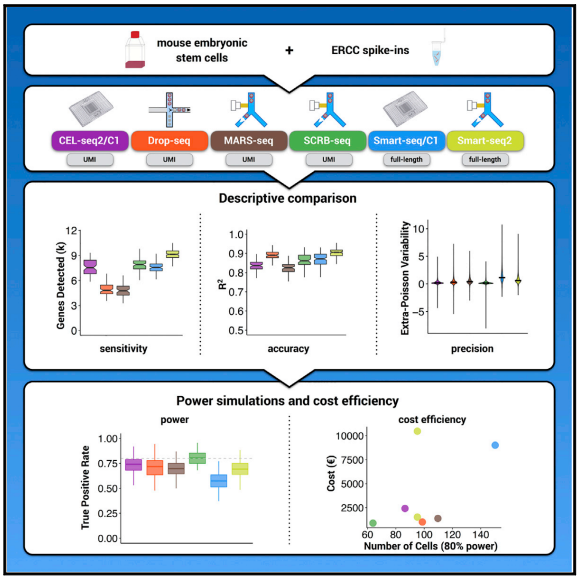
doi: 10.1016/j.molcel.2017.01.023

Article

Molecular Cell

Comparative Analysis of Single-Cell RNA Sequencing Methods

Graphical Abstract



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

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.

Ziegenhain et al., 2017, Molecular Cell 65, 631–643
February 16, 2017
http://dx.doi.org/10.1016/j.molcel.2017.01.023

CellPress

Molecular Cell
Article

CellPress

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,*]
[1]Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians University, Großhaderner Straße 2, 82152 Martinsried, Germany
[2]Ludwig Institute for Cancer Research, Box 240, 171 77 Stockholm, Sweden
[3]Department of Cell and Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden
[4]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain
[5]Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain
[6]Department of Biology II and Center for Integrated Protein Science Munich (CIPSM), Ludwig-Maximilians University, Großhaderner Straße 2, 82152 Martinsried, Germany
[7]Lead 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

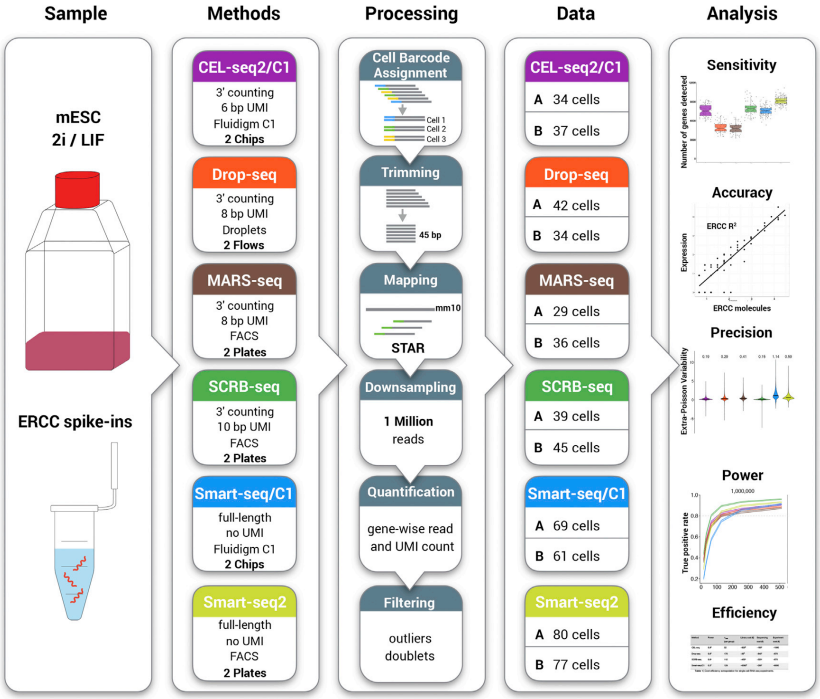


Figure 1. Schematic Overview of the Experimental and Computational Workflow

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, SCRB-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, SCRB-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 SCRB-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.

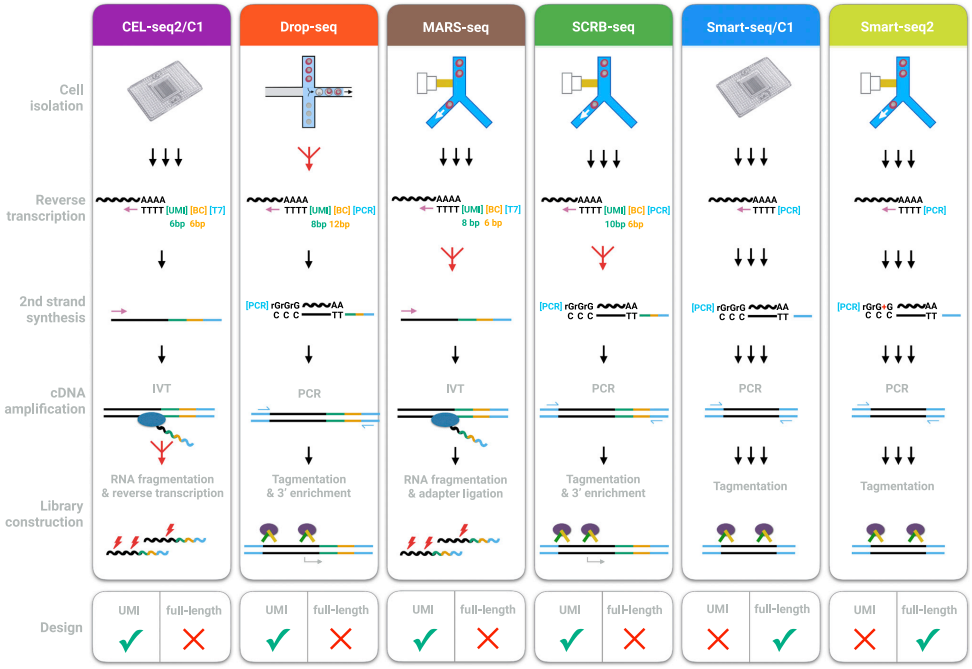


Figure 2. Schematic Overview of Library Preparation Steps
For details, see the text. See also Table S1.

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 SCRB-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. SCRB-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 SCRB-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 SCRB-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, barcoded 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, SCRB-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, SCRB-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 SCRB-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 SCRB-seq (Figure S2C). Notably, CEL-seq2/C1 and MARS-seq showed a steeper slope at low sequencing depths than both Drop-seq and SCRB-seq, potentially due to a less biased amplification by in vitro transcription. Hence, among the UMI methods, CEL-seq2/C1 and SCRB-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 SCRB-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.

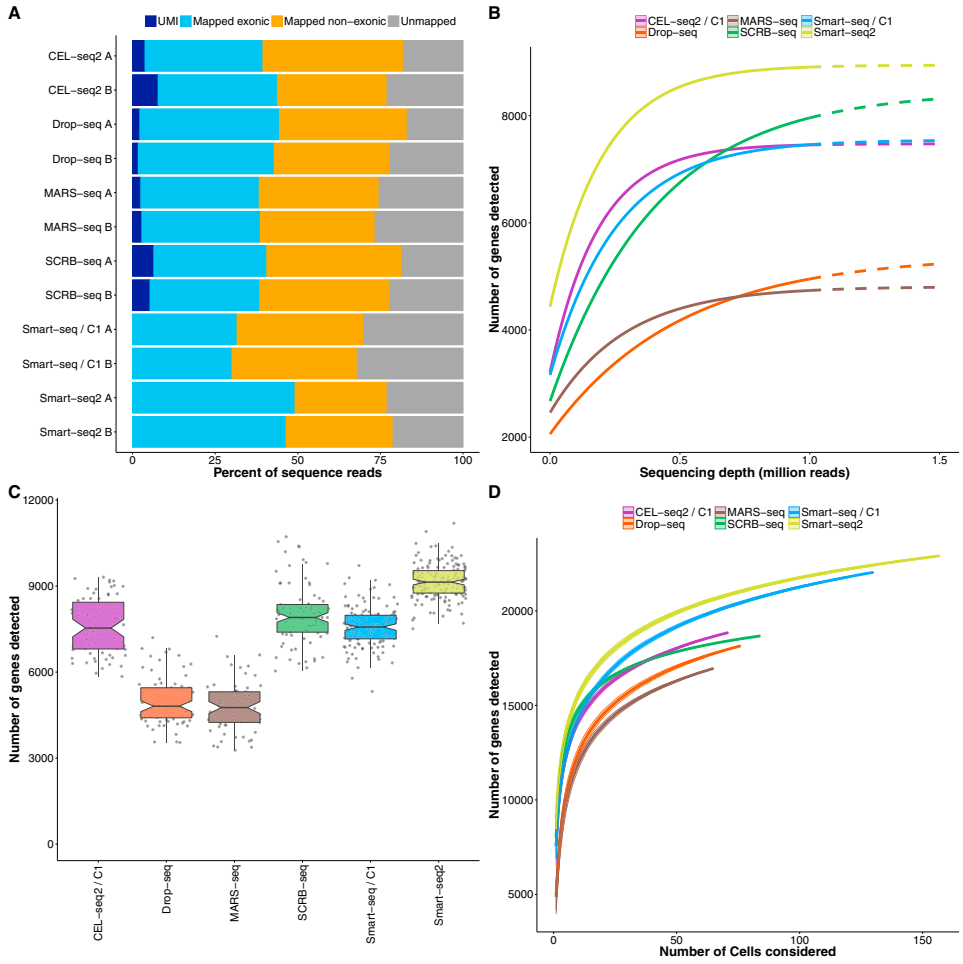


Figure 3. Sensitivity of scRNA-Seq Methods

(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 ± 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

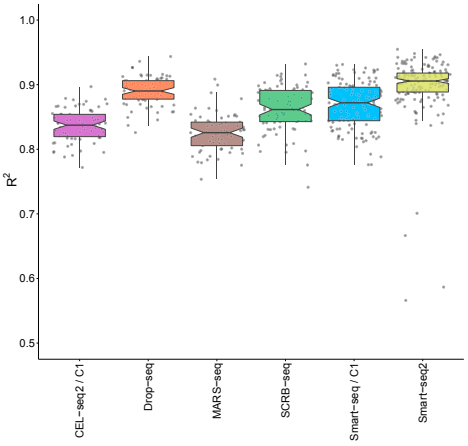


Figure 4. Accuracy of scRNA-Seq Methods

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 ~19,000 genes for CEL-Seq2/C1, ~17,000 for MARS-seq, ~18,000 for Drop-seq and SCRB-Seq, ~20,000 for Smart-seq/C1, and ~21,000 for Smart-seq2 (Figure 3D). While the majority of genes (~13,000) were detected by all methods, ~400 genes were specific to each of the 3′ counting methods, and ~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 ~7, 11, 14, 16, 17, and 28 ERCC molecules for Smart-seq2, Smart-seq/C1, CEL-seq2/C1, SCRB-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 SCRB-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 SCRB-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 SCRB-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

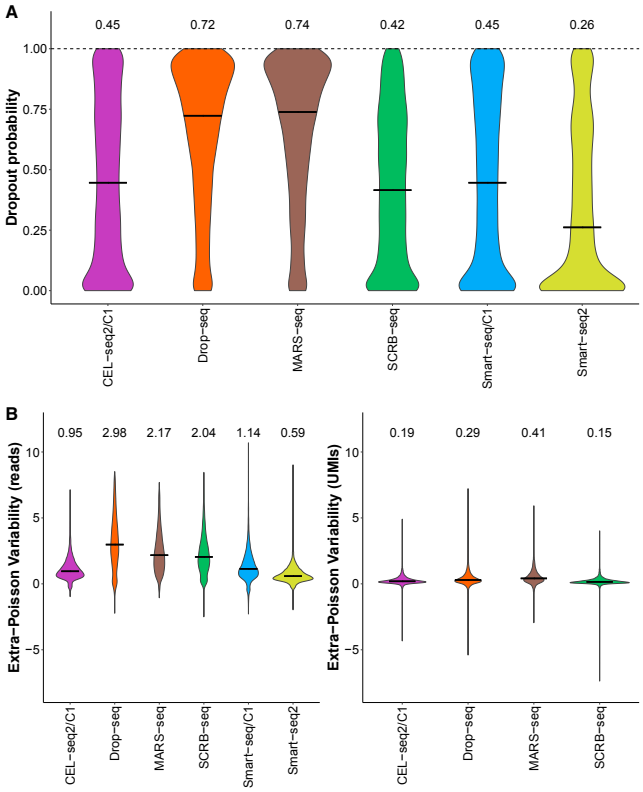


Figure 5. Precision of scRNA-Seq Methods

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, SCRB-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 SCRB-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 SCRB-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% (SCRB-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 SCRB-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 SCRB-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 SCRB-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.

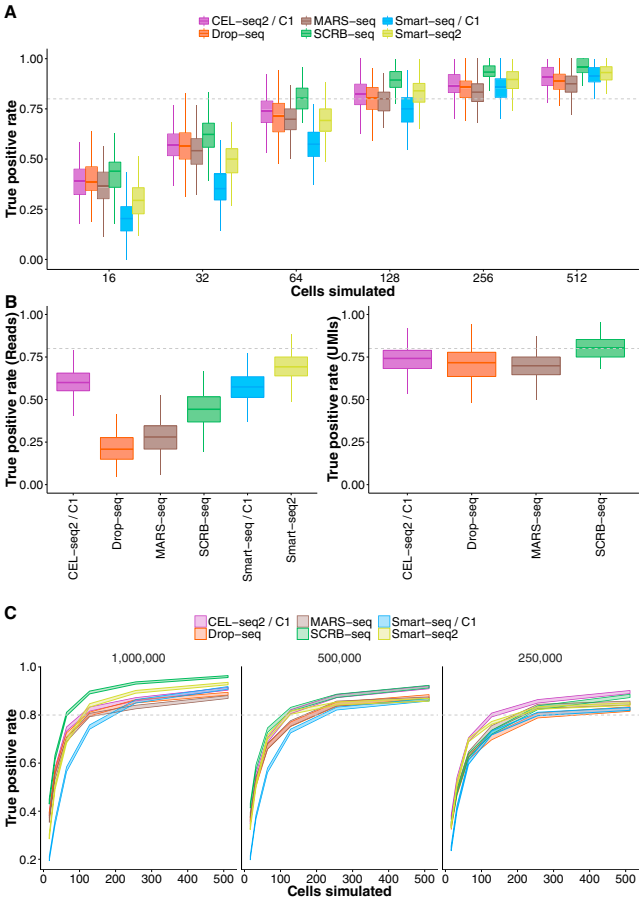


Figure 6. Power of scRNA-Seq Methods

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.

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, SCRB-seq and Drop-seq required 2.6-fold more cells (Table 1). In summary, SCRB-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, 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, SCRB-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, SCRB-seq requires 64 cells per group to reach 80% power. Generating 128 SCRB-seq libraries costs ~260$ and generating 128 million reads costs ~640$. Note that the necessary paired-end reads for CEL-seq2/C1, SCRB-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$) is the most cost-effective method when sequencing 254 cells at a depth of 250,000 reads, and SCRB-seq (810$), 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

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.

[a]True positive rate and false discovery rate are based on simulations (Figure 6; Figure S9).

[b]Sequencing depth of one, 0.5, and 0.25 million reads.

[c]Assuming $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, SCRB-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 comparison 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 SCRB-seq using the ERCCs. In our comparison, we clearly find that Smart-seq2 is the most sensitive method, closely followed by SCRB-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 SCRB-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, SCRB-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 SCRB-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, SCRB-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 SCRB-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 SCRB-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, SCRB-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 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 SCRB-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 SCRB-seq and Drop-seq. Furthermore, the costs of SCRB-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, SCRB-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.

CellPress

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., Kołodziejczyk, 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., Prlic, 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., Nemesh, 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., Edsgärd, 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., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

CellPress

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
Perfluoroctanol	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 μm)	Fluidigm	100-8134
C1 IFC for mRNA-seq (10-17 μm)	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 ACACGACGCTCTTCCGATCT[BC6][N10][T30]VN	IDT	"TruGrade Ultramer"

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SCRB-seq template-switch oligo, iCiGiCACACTCTTTCC CTACACGACGCrGrGrG	Eurogentech	N/A
Drop-seq P5 primer, AATGATACGGCGACCACCGAGA TCTACACGCCT GTCCGCGGAAGCAGTGGTATCAACG 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 μg/ml streptomycin (Sigma-Aldrich), 1000 U/ml recombinant mouse LIF (Millipore) and 2i (1 μM PD032591 and 3 μM 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 μm 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 μl 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, Chemgenes) 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 RNAprotect 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 $MgCl_2$ 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); 10μl 5x KAPA HiFi buffer; 1.5 μl 10mM dNTPs; 2.5μl water) in 10 PCR cycles. Barcoded libraries were purified and pooled at equimolar ratios.

DNA sequencing

For SCRB-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 SCRB-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 SCRB-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/SCRB-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

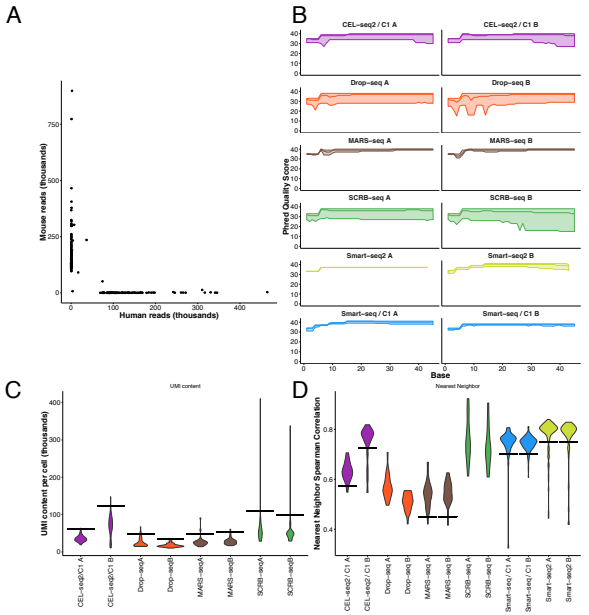


Figure S2

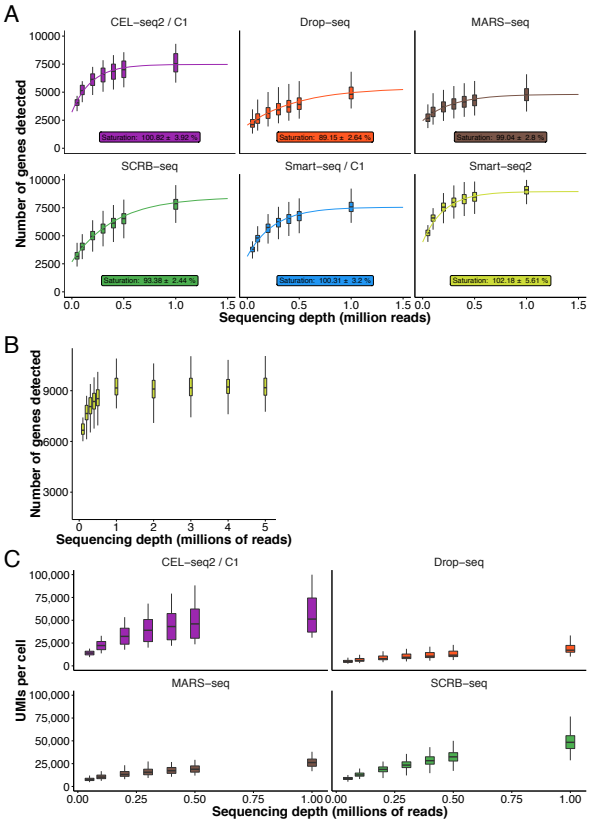


Figure S3

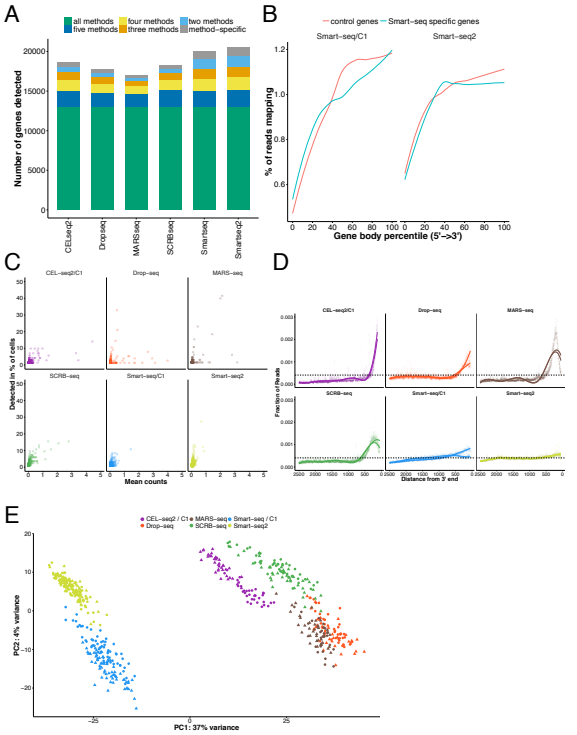


Figure S4

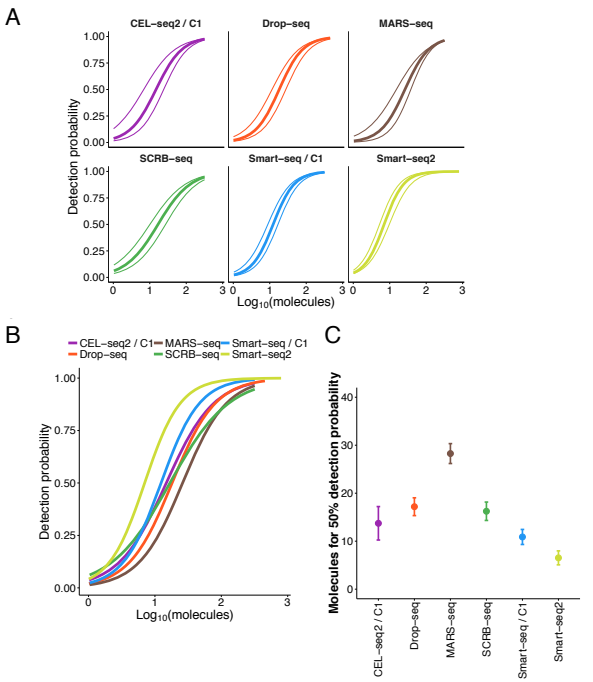


Figure S5

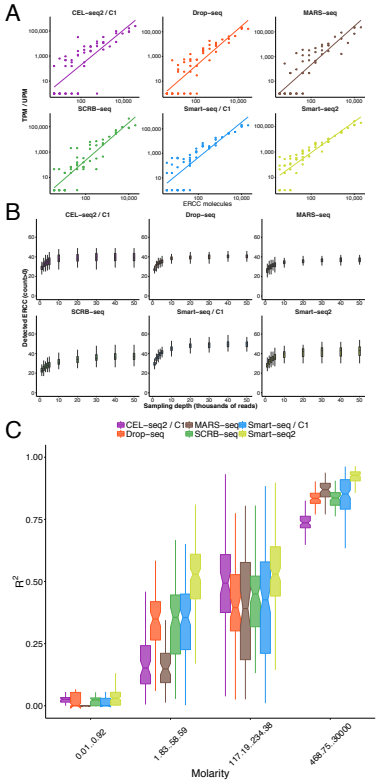


Figure S6

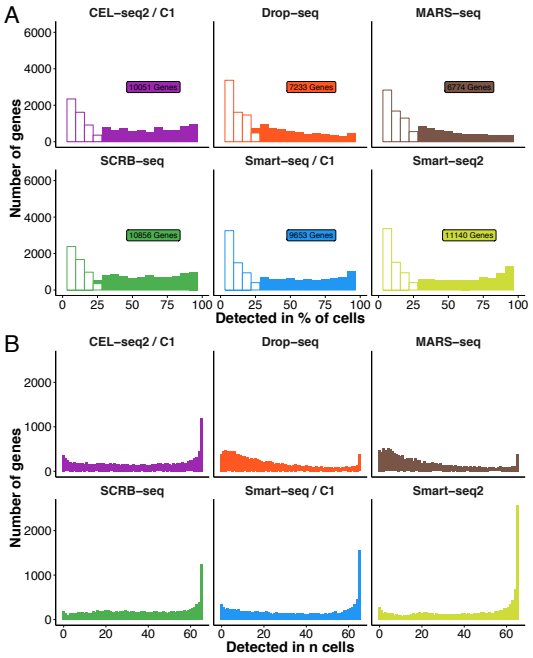


Figure S7

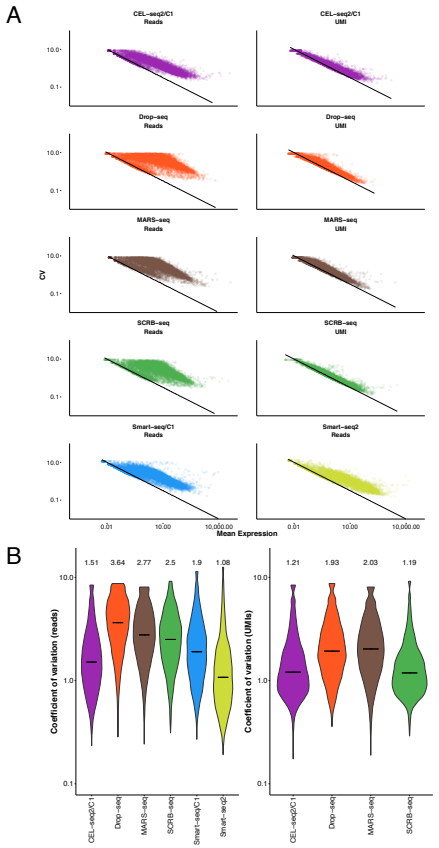


Figure S8

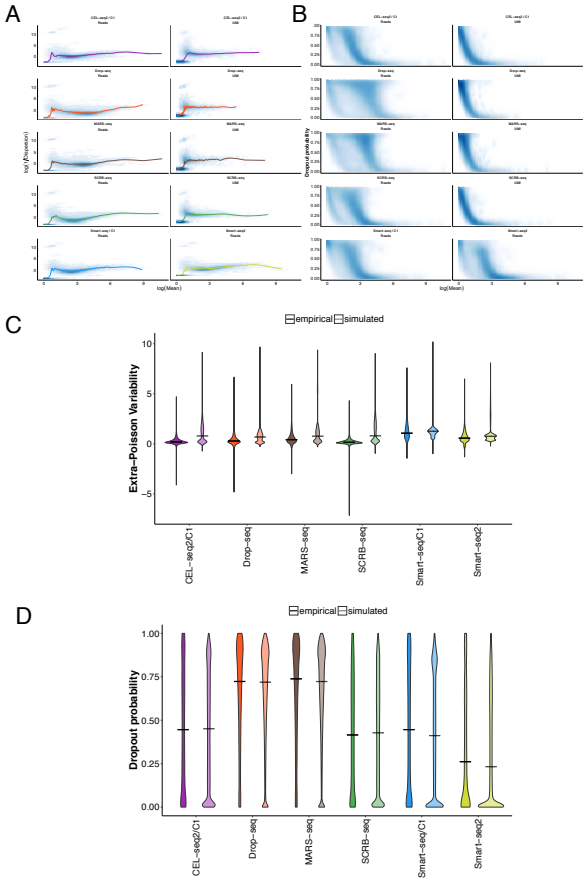


Figure S9

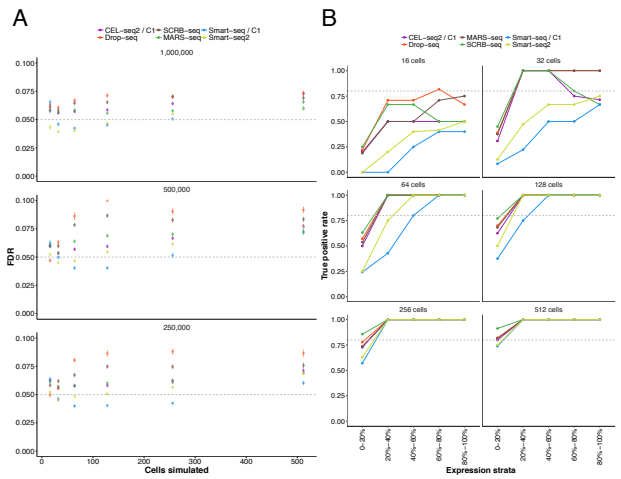
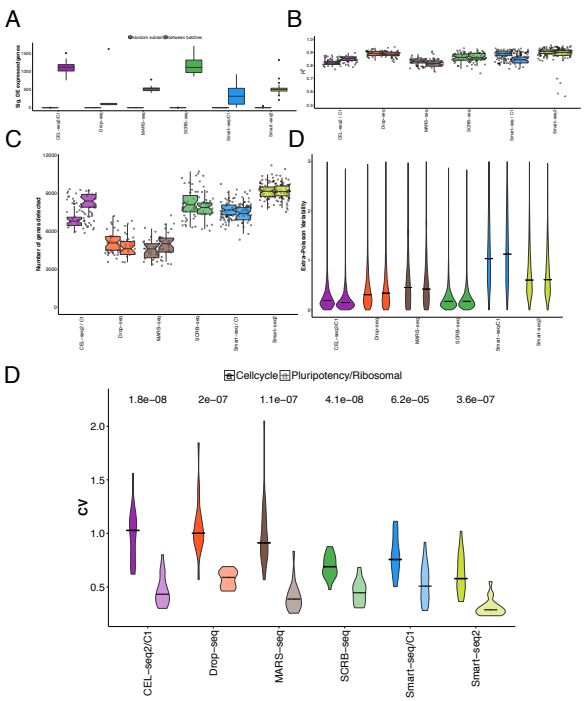


Figure S10



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 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 p0 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

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

Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I:

"powsimR: Power Analysis for Bulk and Single Cell RNA-Seq Experiments." (2017)

Bioinformatics 33 (21): 3486-3488.

doi: 10.1093/bioinformatics/btx435

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

https://academic.oup.com/bioinformatics/article/33/21/3486/3952669#supplementary-data

Bioinformatics, 33(21), 2017, 3486–3488
doi: 10.1093/bioinformatics/btx435
Advance Access Publication Date: 11 July 2017
Applications Note

Gene expression

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

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

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

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on March 15, 2017; revised on June 29, 2017; editorial decision on July 2, 2017; accepted on July 4, 2017

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 and implementation: The R package and associated tutorial are freely available at https://github.com/bvieth/powsimR.
Contact: vieth@bio.lmu.de or 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 (Hashimshony *et al.*, 2012; Picelli *et al.*, 2014) 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 (Klein *et al.*, 2015; Macosko *et al.*, 2015; Zheng *et al.*, 2017).

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 the power to detect differentially expressed genes. For our read count simulations, we (i) reliably model the mean, dispersion and dropout distributions as well as the relationship between those factors from the data. (ii) 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. (iii) 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.

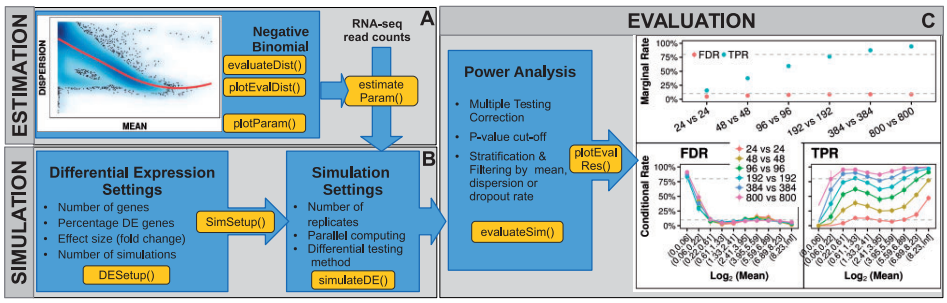


Fig. 1. powsimR schematic overview. (**A**) The mean-dispersion relationship is estimated from RNA-seq data, which can be either single cell or bulk data. The user can provide their own count tables or one of our five example datasets and choose whether to fit a negative binomial or a zero-inflated negative 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)

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; Lun *et al.*, 2016; Mi *et al.*, 2015), 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 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 (Soumillon *et al.*, 2014; Ziegenhain *et al.*, 2017) (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 Fig. S4). Therefore the default sampling distribution in powsimR is the negative binomial (Fig. 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 (Fig. 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 Fig. 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 Fig. 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), EBSeq (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 (Fig. 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 dataset 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).

Conflict of Interest: none declared.

References

Auer,P.L. and Doerge,R.W. (2010) Statistical design and analysis of RNA sequencing data. *Genetics*, **185**, 405–416.

Bacher,R. and Kendziorski,C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, **17**, 63.

Conesa,A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.

Finak,G. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 1–13.

Grün,D. *et al.* (2014) Validation of noise models for single-cell transcriptomics. *Nat. Methods*, **11**, 637–640.

Hardcastle,T.J. (2016) Generalized empirical bayesian methods for discovery of differential data in high-throughput biology. *Bioinformatics*, **32**, 195–202.

Hashimshony,T. *et al.* (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, **2**, 666–673.

Kharchenko,P.V. *et al.* (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.

Klein,A.M. *et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.

Kolodziejczyk,A.A. *et al.* (2015) Single cell RNA-Sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, **17**, 471–485.

Korthauer,K.D. *et al.* (2016) A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.*, **17**, 222.

Leng,N. *et al.* (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**, 1035–1043.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Lun,A.T.L. *et al.* (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**, 75.

Macosko,E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Mi,G. *et al.* (2015) Goodness-of-fit tests and model diagnostics for negative binomial regression of RNA sequencing data. *PLoS One*, **10**, e0119254.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Picelli,S. *et al.* (2014) Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, **9**, 171–181.

Poplawski,A. and Binder,H. (2017) Feasibility of sample size calculation for RNA-seq studies. *Brief. Bioinform.* pii: bbw144.

Qiu,X. *et al.* (2017) Single-cell mRNA quantification and differential analysis with census. *Nat. Methods*, **14**, 309–315.

Ritchie,M.E. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Robinson,M.D. *et al.* (2010) edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Schurch,N.J. *et al.* (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*.

Seyednasrollah,F. *et al.* (2015) ROTS: reproducible RNA-seq biomarker detector—prognostic markers for clear cell renal cell cancer. *Nucleic Acids Res.*, gkv806.

Soumillon,M. *et al.* and others (2014) Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*.

Tarazona,S. *et al.* (2015) Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.*, **43**, e140.

Vu,T.N. *et al.* (2016) Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*, **32**, 2128–2135.

Wu,A.R. *et al.* (2014) Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, **11**, 41–46.

Wu,H. *et al.* (2013) A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, **14**, 232–243.

Zheng,G.X.Y. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.

Ziegenhain,C. *et al.* (2017) Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, **65**, 631–643.e4.

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

Supplementary Information

by

Beate Vieth[1], Christoph Ziegenhain[1], Swati Parekh[1], Wolfgang Enard[1] and Ines Hellmann[1]

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

1 Determining the best fitting distribution per gene

To determine the best fitting distribution to the observed RNA-seq count data, we compare the theoretical fit of the Poisson, negative binomial (NB), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) and Beta-Poisson (BP) distribution to the empirical RNA-seq read counts [2, 8, 3]. We used the following statistics to evaluate which distribution fits best:

- goodness of fit (GOF) statistics based on Chi-square statistic using residual deviances and degrees of freedom (Chi-square test).
- Akaike Information Criterium (AIC).
- Likelihood Ratio Test (LRT) for nested models, i.e. testing whether estimating a dispersion parameter in the NB models is appropriate.
- Vuong Test (VT) for non-nested models, i.e. testing whether assuming zero-inflation results in a better fit.
- Comparing the observed dropouts to the zero count prediction of the models.

Note that the goodness of fit statistics could not be calculated for the BP, however, since it already the AIC statistic suggested that the BP fit worse than the other distributions and could neither predict the dropouts correctly (Figure S1, Supplementary File S2), we did not follow this further.

We analyzed 8 published single cell RNA-seq studies ([1, 9, 11, 6, 7, 14, 13, 15]) produced using 9 different RNA-seq library preparation methods (Smart-seq/C1, Smart-seq2, MARS-seq, SCRB-seq, STRT, STRT-UMI, Drop-seq, 10XGenomics, CEL-seq2). For illustrative purposes, we focus on Kolodziejczk et al. (2015) [9], but the distribution analysis for all can be found in Supplementary File S2.

For the Kolodziejczk et al. (2015) data, we found that the NB distribution is an adequate fit (Figure S1): The Chi-Square test indicates that the NB is appropriate for at least 40 % of the genes (Figure S1 A). Moreover, the AIC suggests that the NB is in 60% of the cases better than the Poisson, ZIP, ZINB and BP (Figure S1 B). The ZINB is the only of the commonly used distributions that comes close, providing the best fit for 40% of all compared genes, however this difference is only significant for 6% (Figure S1D).

One of the major differences between the methods is the use of Unique Molecular Identifiers (UMIs) that allow for confident removal of PCR-duplicates [5, 15]. For all protocols considered, we evaluated the fit of the 5 different distributions, and for the vast majority the NB would be the distribution of choice (Figure S2). 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 (Figure S4).

Next, we assess the fit of the dropout rate by comparing expected and predicted zero counts per gene. Interestingly, even though the negative binomial does not model dropouts explicitly, the deviation of predicted zero counts from the expected under the NB distribution is relatively small (Figure S1 C). The ZINB only gives

a small advantage with respect to dropouts. The comparison of models by LRT and VT illustrates the small improvement of the model fit by assuming a ZINB distribution (10%) (FigureS1 D) for the Kolodziejczk data, which is comparable to the average for non-UMI methods, and much lower for the UMI-methods (<5%)(Figure S4 and Figure S3).

We thus refrain from using a mixture distribution, however for some of the protocols that do not utilize UMIs, such as e.g. Smart-Seq2, the ZINB might provide a better fit and should be used as a sampling distribution in the power simulations.

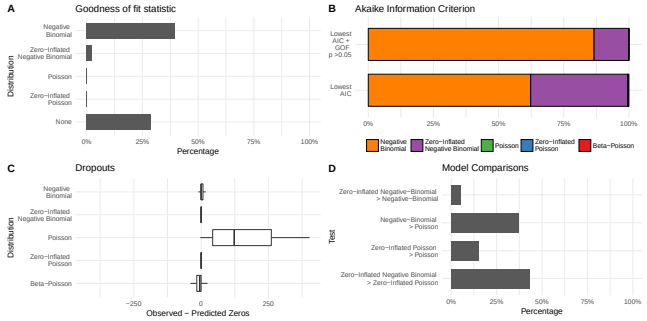


Figure S1: A) Goodness of fit of the model per gene assessed with a Chi-square test based on residual deviance and degrees of freedom. B) The fraction of genes for which the respective distribution has the lowest AIC and additionally the distribution with the lowest AIC as well as not rejected by the goodness of fit statistic. C) Observed versus predicted dropouts per distributional model and gene. D) Model assessment per gene based on Likelihood Ratio Test for nested models and Vung Test for non-nested models.
The same plot representing other datasets can be found in Supplementary File S2.

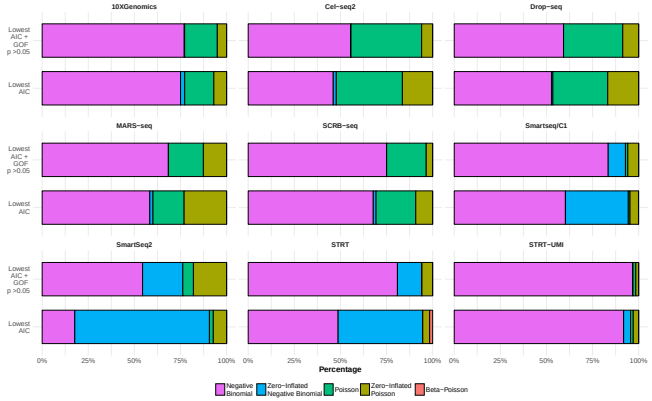


Figure S2: The negative binomial gives the best fit for the majority of genes (i.e. lowest AIC) for all UMI datasets. For protocols that do not account for PCR duplicates, the zero-inflated negative binomial often has a lower AIC, however this is mainly due to genes that cannot be fitted very well in general (GOF p-value<=0.05).

2 Read Count Simulation Framework

We have implemented a read count simulation framework assuming an underlying negative binomial distribution. To predict the dispersion θ given a random draw of an observed mean expression value μ, we apply a locally weighted polynomial regression fit. Furthermore, to capture the variability of the observed dispersion estimates, a local variability prediction band is applied (R package msir [12]). The read count for gene i in sample j is then given by:

$$X_{ij} \sim NB(\mu, \theta) \tag{1}$$

The mean, dispersion and dropout rates of an example read count simulation closely resembles the observed estimates for the Kolodziejczk data set (Figure S5).

For bulk RNA-seq experiments, the negative binomial alone is not able to capture the observed number of dropouts appropriately. Here, we predict the dropout probability (p_0) using a decreasing constrained B-splines regression (CRAN R package cobs [10]) of dropout rate against mean expression to determine the mean expression value μ_{DP5}, where the dropout probability is expected to fall below 5%. For all genes with

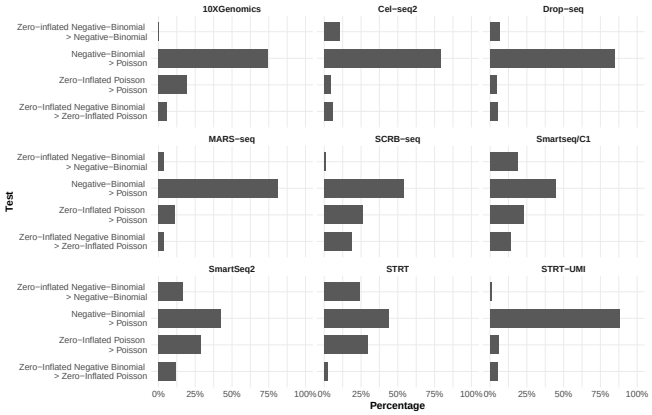


Figure S3: Model assessment per gene based on likelihood ratio test for nested models and Vuong test for non-nested models shows that zero-inflated negative binomial significantly improves the fit for maximally 25% of the genes (STRT protocol).

$\mu_i < \mu_{DP5}$ we do not estimate a gene specific dropout probability, but sample the dropout probability from all genes with $< \mu_{DP5}$. With these parameters, the read count for a gene i in a sample j is modeled as a product of a negative binomial multiplied with an indicator whether that sample was a dropout or not, which is determined using binomial sampling:

$$X_{ij} \sim I * NB(\mu, \theta), \text{where } I \in \{0, 1\} \tag{2}$$

$$P(I = 0) = B(1 - p_0) \tag{3}$$

The necessity of this apparently unintuitive zero inflation for bulk data is illustrated by the dataset from Eizirik et al. 2012 [4]. Note that dropouts occur across genes with different mean expression levels so that there is only a very weak relationship between mean expression and dropout probabilities (Figure S6).

For the simulations of expression changes, the user can freely define a distribution, a list of log_2-fold changes or simply a constant. We recommend to simulate with a realistic log_2-fold change distribution, which we determined for the Kolodziejczyk et al. (2015) [9] as a narrow $\Gamma(\alpha, \beta)$- distribution plus $-1 \times \Gamma(\alpha, \beta)$ (Figure S7).

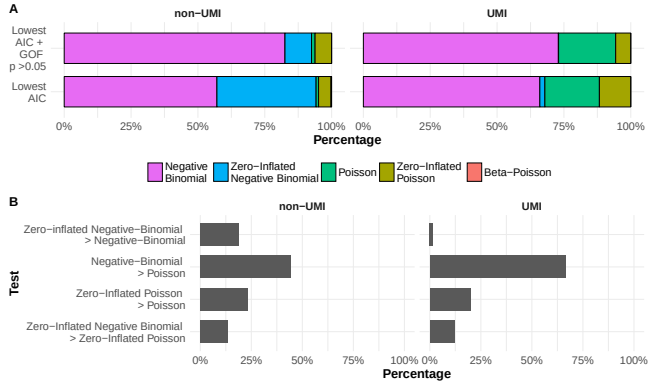


Figure S4: 6 UMI-protocols (STRT-UMI, Cel-SEq2, Drop-seq, MARS-seq, SCRB-seq,10XGenomics) are compared to 3 protocols not using UMIs (Smartseq/C1, SmartSeq2, STRT), showing that zero-inflation is only relevant for non-UMI-methods. A) The fraction of genes for which the respective distribution has the lowest AIC and additionally the distribution with the lowest AIC is not to rejected by the goodness of fit statistic. D) Model assessment per gene based on likelihood ratio test for nested models and Vuong test for non-nested models.

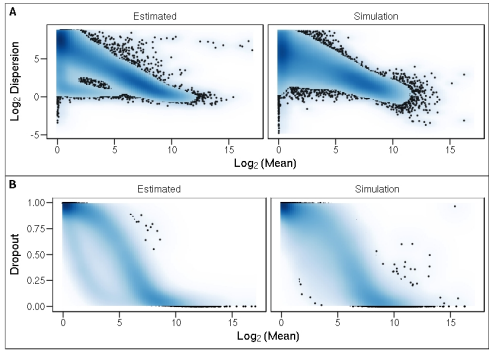


Figure S5: A) Dispersion versus mean. B) Dropout versus mean.

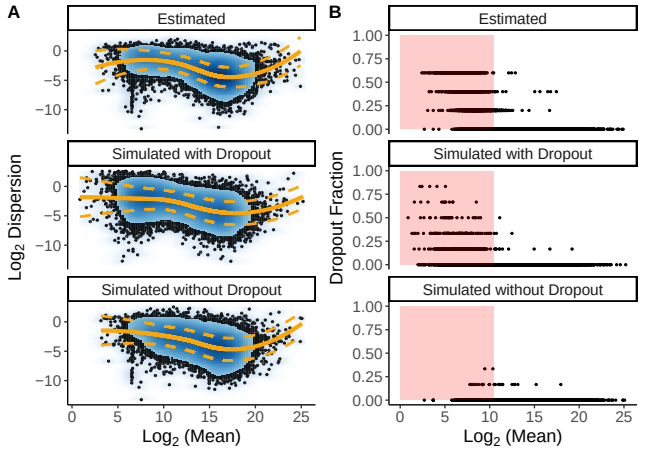


Figure S6: For bulk RNA-seq, the simulations include dropout sampling to better mimic the observed mean-dropout relation. A) Dispersion versus mean with locally weighted polynomial regression fit (orange line) and variability prediction band (dashed orange line). B) Dropout versus mean with red box indicating genes with $< \mu_{DP5}$ from which the dropout probability will be sampled from.

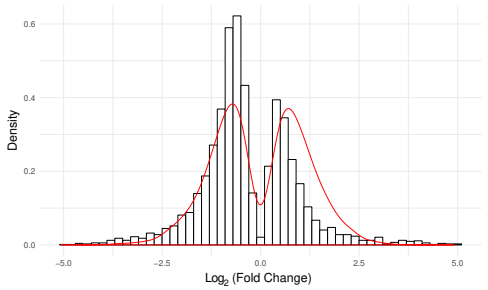


Figure S7: Log2 fold changes between serum+LiF and 2i+LiF cultured cells (Kolodziejczk et al. 2015). Red line indicates the density of a theoretical narrow gamma distribution (shape and rate equal to 3).

3 Included RNA-seq Experiments

We provide raw count matrices for several published single cell data sets (Table S1 on github (https://github.com/bvieth/powsimRData). Furthermore, the vignette gives an example on how to access RNA-seq datasets in online repositories such as recount (https://jhubiostatistics.shinyapps.io/recount/).

Table S1: Key properties of the example data-sets included in powsimR.

	Study	Accession	Species	No. Cells	Cell-type*	Library preparation	UMI	Remarks
1	Kolodziejczk et al. (2015) [9]	E-MTAB-2600	Mouse	869	ESC	Smart-seq C1	no	different growth media
2	Islam et al. (2011) [6]	GSE29087	Mouse	48	ESC	STRT-seq	no	-
3	Islam et al. (2014) [7]	GSE46980	Mouse	96	ESC	STRT-seq C1	yes	-
4	Buettner et al. (2015) [1]	E-MTAB-2805	Mouse	288	ESC	Smart-seq C1	no	FACs-sorted for cell-cycle
5	Soumillon et al. (2014) [13]	GSE53638	Human	12,000	adipocytes	SCRB-seq	yes	time-series

* ESC - embryonic stem cells

References

[1] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, advance online publication, 19 January 2015.

[2] A Colin Cameron and Pravin K Trivedi. *Regression Analysis of Count Data (Econometric Society Monographs)*. Cambridge University Press, 2 edition edition, 27 May 2013.

[3] Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (D3E)–a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, 17:110, 29 February 2016.

[4] Décio L Eizirik, Michael Sammeth, Thomas Bouckenooghe, Guy Bottu, Giorgia Sisino, Mariana Igoillo-Esteve, Fernanda Ortis, Izortze Santin, Maikel L Colli, Jenny Barthson, Luc Bouwens, Linda Hughes, Lorna Gregory, Gerton Lunter, Lorella Marselli, Piero Marchetti, Mark I McCarthy, and Miriam Cnop. The human pancreatic islet transcriptome: Expression of candidate genes for type 1 diabetes and the impact of Pro-Inflammatory cytokines. *PLoS Genet.*, 8(3):e1002552, 2012.

[5] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat. Methods*, 11(6):637–640, June 2014.

[6] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, 21(7):1160–1167, 1 July 2011.

[7] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11(2):163–166, February 2014.

[8] Jong Kyoung Kim and John C Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.*, 14(1):R7, 28 January 2013.

[9] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Jason C H Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N Natarajan, Alex C Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, John C Marioni, and Sarah A Teichmann. Single cell RNA-Sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17(4):471–485, 1 October 2015.

[10] Pin Ng and Martin Maechler. A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling*, 7(4):315–328, 2007.

[11] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, Naveen Ramalingam, Gang Sun, Myo Thu, Michael Norris, Ronald Lebofsky, Dominique Toppani, Darnell W Kemp, Ii, Michael Wong, Barry Clerkson, Brittnee N Jones, Shiquan Wu, Lawrence Knutsson, Beatriz Alvarado, Jing Wang, Lesley S Weaver, Andrew P May, Robert C Jones, Marc A Unger, Arnold R Kriegstein, and Jay A A West. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, 32(10):1053–1058, October 2014.

[12] Luca Scrucca. Model-based sir for dimension reduction. *Computational Statistics & Data Analysis*, 5(11):3010–3026, 2011.

[13] 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*, page 003236, 5 March 2014.

[14] Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, 16 January 2017.

[15] 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.

2.3.1 Updates to powsimR

I have continued to develop the powsimR package, improving the package infrastructure and implementing additional methods. This section briefly describes changes to the software since its publication. The most significant changes include the extension of the simulation framework as well as implementing additional methods. There is now the possibility to simulate technical batches of samples or cells. The simulation of spike-in ERCCs [110] is now also possible so that users can now utilize spike-in-aware normalisation and DE testing methods. In addition, I implemented additional normalisation and DE testing methods so that the user can now choose between 12 and 15 methods, respectively. Due to the increased attention for imputing scRNA-seq data, I included five imputation methods that can be applied prior to normalisation as well as DE testing can now be run on the imputed data. While working on "A Systematic Evaluation of scRNA-seq pipelines", I considered additional performance metrics related to DE testing results which are now also available in the package. Besides the traditional metrics of statistical power analysis (TPR, FPR, FDR, etc.), users can also use composite measures such as TPR versus FDR curve as advocated by Soneson and Robinson 2016 [211]. Estimating the deviance in simulated versus estimated library size factors as well as the error in log2 fold changes is now also available to evaluate normalisation, imputation and DE-testing choices. Given the increased throughput of scRNA-seq in the last couple of years, researchers might be interested in statistical power analysis where more cells are included but without an increase in total sequencing. Thus, I have included flexible downsampling of simulated read counts.

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

Parekh S*, Ziegenhain C*, **Vieth B**, Hellmann I, Enard W:

"zUMIs: A fast and flexible pipeline to process RNA sequencing data with UMIs." (2018)

GigaScience 7 (6).giy059.

doi: 10.1093/gigascience/giy059

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

https://academic.oup.com/gigascience/article/7/6/giy059/5005022#supplementary-data

OXFORD (GIGA)n SCIENCE

GigaScience, 7, 2018, 1–9

doi: 10.1093/gigascience/giy059
Advance Access Publication Date: 26 May 2018
Technical Note

TECHNICAL NOTE

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

Swati Parekh [*,†,‡], Christoph Ziegenhain [†,§], Beate Vieth, Wolfgang Enard and Ines Hellmann [*]

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

*Correspondence address. Ines Hellmann. E-mail: hellmann@bio.lmu.de http://orcid.org/0000-0003-0588-1313 and Swati Parekh. E-mail: sparekh@age.mpg.de https://orcid.org/0000-0002-4826-1651
†Contributed equally.
‡Present address: Max Planck Institute for Biology of Ageing, 50931 Cologne, Germany.
§Present address: Department of Cell & Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden.
Technical Note

Abstract

Background: Single-cell RNA-sequencing (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 bar codes (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. **Findings:** *zUMIs* is a pipeline that can handle both known and random BCs and also efficiently collapse UMIs, either just for exon mapping reads or for both exon and intron mapping reads. If BC annotation is missing, *zUMIs* can accurately detect intact cells from the distribution of sequencing reads. Another unique feature of *zUMIs* is the adaptive downsampling function that facilitates dealing with hugely varying library sizes but also allows the user to evaluate whether the library has been sequenced to saturation. To illustrate the utility of *zUMIs*, we analyzed a single-nucleus RNA-seq dataset and show that more than 35% of all reads map to introns. Also, we show that these intronic reads are informative about expression levels, significantly increasing the number of detected genes and improving the cluster resolution. **Conclusions:** *zUMIs* flexibility makes if possible to accommodate data generated with any of the major scRNA-seq protocols that use BCs and UMIs and is the most feature-rich, fast, and user-friendly pipeline to process such scRNA-seq data.

Keywords: single-cell RNA-sequencing; digital gene expression; unique molecular identifiers; pipeline

Introduction

The recent development of increasingly sensitive protocols allows for the generation of RNA-sequencing (RNA-seq) libraries of single cells [1]. The throughput of such single-cell RNA-seq (scRNA-seq) protocols is rapidly increasing, enabling the profiling of tens of thousands of cells [2, 3] and opening exciting possibilities to analyze cellular identities [4, 5]. As the required amplification from such small 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 expression differences between cells [8, 9]. To increase the throughput, many protocols also incorporate sample-specific bar codes (BCs) to label all cDNA molecules of a single cell with a nucleotide sequence before library generation [10]. This allows for early pooling, which further decreases amplification noise [6]. Addition-

Received: 17 October 2017; **Revised:** 16 March 2018; **Accepted:** 15 May 2018

Table 1: Features of available UMI pipelines for the quantification of gene expression data.

Name	Reference	Open source	Quality filter	UMI collapsing	Mapper	BC detection	Intron	Down-sampling	Compatible UMI library protocols
Cell Ranger	[2]	yes	BC+UMI	Hamming distance	STAR	A	no	yes	[2]
CEL-seq	[15]	yes	BC+UMI	Identity only	bowtie2	WL	no	no	[15, 46]
dropEst	[16]	yes	BC	Frequency-based	TopHat2 or Kallisto	WL,top-n,EM	yes	no	[2, 13, 19]
Drop-seq-tools	[13]	no	BC+UMI	Hamming distance	STAR	WL,top-n	no	no	[13, 15, 17]
scPipe	[47]	yes	BC+UMI	Hamming distance	subread	WL,top-n	no	no	[13, 17, 18, 46]
umis	[14]	yes	BC	Frequency-based	Kallisto	WL,top-n,EM	no	no	[2, 13, 17–19, 46, 48]
UMI-tools	[25]	yes	BC+UMI	Network-based	BWA	WL	no	no	[17, 19]
zUMIs	This work	yes	BC+UMI	Hamming distance	STAR	A,WL,top-n	yes	yes	[2, 3, 12, 13, 15, 17, 18, 21, 46, 48]

We consider whether the pipeline is open source, has sequence quality filters for cell BCs and UMIs, mappers, UMI-collapsing options, options for BC detection (A, automatically infer intact BCs; WL, extract only the given list of known BCs; top-n, order BCs according the number of reads and keep the top n BCs; EM, merge BCs with given edit distance), whether it can count intron mapping reads, whether it offers a utility to make varying library sizes more comparable via downsampling, and finally with which RNA-seq library preparation protocols is it compatible

ally, for cell types such as primary 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 (Table 1). For example, the Drop-seq-tools is not an open source [13]. While Cell Ranger is open, it is exceedingly difficult to adapt the code to new or unknown sample BCs and other library types. Other tools are specifically designed to work with one mapping algorithm and focus mainly on transcriptome references [14, 15]. Furthermore, the only other UMI-RNA-seq pipeline providing the utility to also consider intron mapping reads, dropEst [16], is only applicable to droplet-based protocols. Here, we present *zUMIs*, a fast and flexible pipeline that overcomes these limitations.

Findings

zUMIs is a pipeline to process RNA-seq data that were multiplexed using cell BCs and also contain UMIs. Read-pairs are filtered to remove reads with low-quality BCs or UMIs based on sequence and then mapped to a reference genome (Fig.1). Next, *zUMIs* generates UMI and read count tables for exon and exon+intron counting. We reason that very low input material such as from single nuclei sequencing might profit from including reads that potentially originate from nascent RNAs. Another unique feature of *zUMIs* is that it allows for downsampling of reads before collapsing UMIs, thus 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 BCs and UMIs, supporting protocols that have both sequences in one read [2, 3, 12, 13, 15, 17, 18] as well as protocols that provide UMI and BC in separate reads [19–21]. This makes *zUMIs* the only tool that is easily compatible with all major UMI-based scRNA-seq protocols.

Implementation and Operation

Filtering and mapping

The first step in our pipeline is to filter reads that have low-quality BCs according to a user-defined threshold (Fig.1). This step eliminates the majority of spurious BCs and thus greatly reduces the number of BCs that need to be considered for counting. Similarly, we also filter low-quality UMIs.

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.

Transcript counting

Next, reads are assigned to genes. In order to distinguish exon and intron counts, we generate two mutually exclusive annotation files from the provided gtf, one detailing exon positions, the other introns. Based on those annotations, `Rsubread featureCounts` [23] is used to first assign reads to exons and afterward to check whether the remaining reads fall into introns, in other words, if a read is overlapping with intronic and exonic sequences, it will be assigned to the exon only. The output is then read into R using `data.table` [24], generating count tables for UMIs and reads per gene per BC. We then collapse UMIs that were mapped either to the exon or intron of the same gene. Note that only the processing of intron and exon reads together allows for properly collapse of UMIs that can be sampled from the intronic as well as from the exonic part of the same nascent mRNA molecule.

Per default, we only collapse UMIs by sequence identity. If there is a risk that a large proportion of UMIs remains undercollapsed due to sequence errors, *zUMIs* provides the option to collapse UMIs within a given Hamming distance. We compare

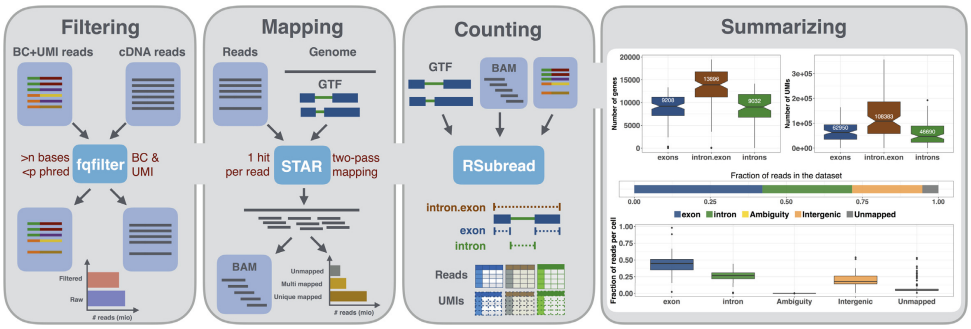


Figure 1: Schematic of the zUMIs pipeline. Each of the gray panels from left to right depicts a step of the *zUMIs* pipeline. First, fastq files are filtered according to user-defined bar code (BC) and unique molecular identifier (UMI) quality thresholds. Next, the remaining cDNA reads are mapped to 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 BCe and distribution of reads into mapping feature categories.

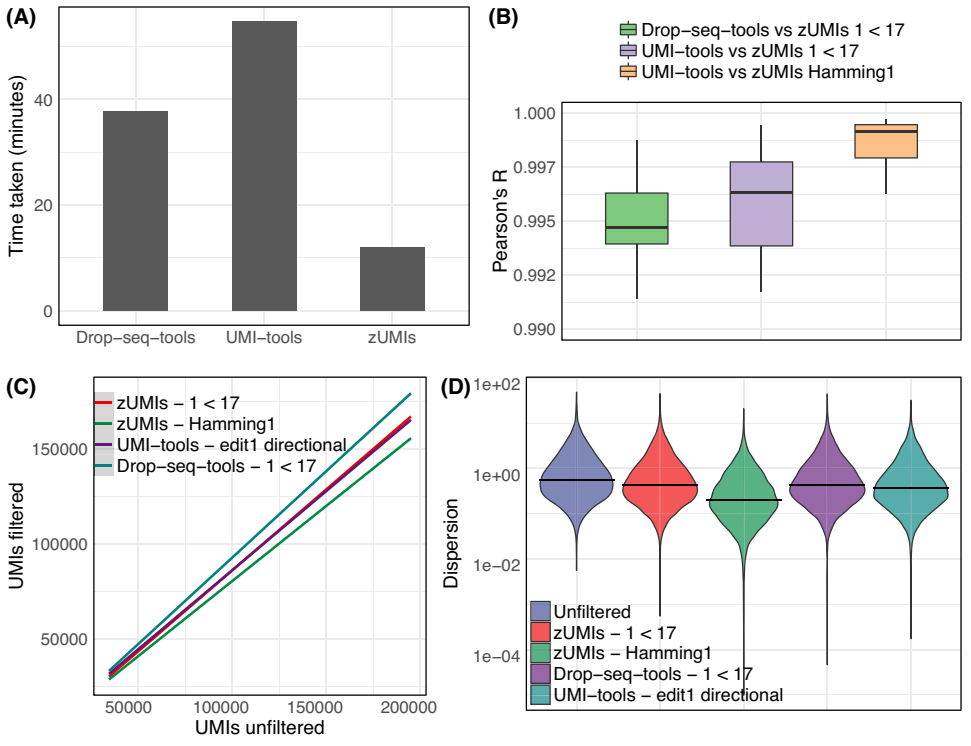


Figure 2: Comparison of different UMI collapsing methods. We compared Drop-seq-tools and UMI-tools with zUMIs using our HEK dataset (227 mio reads). **(A)** Run time to count exonic UMIs using zUMIs (hamming distance = 0), UMI-tools ("unique" mode) and Drop-seq-tools (edit distance = 0). **(B)** Box plots of correlation coefficients of gene-wise UMI counts of the same cell generated with different methods. UMI counts generated using zUMIs (quality filter "1 base under phred 17" or hamming distance = 1) were correlated to UMI counts generated using Drop-seq-tools (quality filter "1 base under phred 17") and UMI-tools ("directional adjacency" mode). **(C)** Comparison of the total number of UMIs per cell derived from different counting methods to "unfiltered" counts. **(D)** Violin plots of gene-wise dispersion estimates with different quality filtering and UMI collapsing methods.

the two *zUMIs* UMI-collapsing options to the recommended directional adjacency approach implemented in UMI-tools [25] using our in-house example dataset (see Methods section). *zUMIs* identity collapsing yields nearly identical UMI counts per cell as UMI-tools, while Hamming distance yields increasingly fewer UMIs per cell with increasing sequencing depth (Fig.2C). Smith et al [25] suggest that edit distance collapsing without considering the relative frequencies of UMIs might indeed overreach and overcollapse the UMIs. We suspect that this is indeed what happens in our example data, where we find that gene-wise dispersion estimates appear suspiciously truncated as expected if several counts are unduly reduce to one, the minimal number after collapsing (Fig.2D).

However, note that the above-described differences are minor. By and large, there is good agreement between UMI counts obtained by UMI-tools [25], the Drop-seq pipeline [13], and *zUMIs*. The correlation between gene-wise counts of the same cell is >0.99 for all comparisons (Fig. 2B). In light of this, we consider the >3 times higher processing speed of *zUMIs* to be a decisive advantage (Fig.2A).

Cell BC 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 BCs mark good cells from the data (Fig.3A, 3B). To this end, we fit a k-dimensional multivariate normal distribution using the R-package `mclust` [26, 27] for the number of reads/BC, where k is empirically determined by mclust via the Bayesian information criterion. We reason that only the kth normal distribution with the largest mean contains BCs that identify reads originating from intact cells. We exclude all BCs that fall in the lower 1% tail of this kth normal distribution to exclude spurious BCs. The HEK dataset used here contains 96 cells with known BCs and *zUMIs* identifies 99 BCs as intact, including all the 96 known BCs. 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, we recommend to always check the elbow plot generated by zUMIs (Fig.3B) to confirm that the cutoff used by zUMIs is valid for a given dataset. In cases where the number of BCs or BC sequences are known, it is preferable to use this information. If *zUMIs* is either given the number of expected 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 [28, 29]. A straight-forward solution for this issue is to downsample overrepresented libraries [30]. *zUMIs* has an built-in function for downsampling datasets to a user-specified number of reads or a range of reads. By default, *zUMIs* downsamples all selected BCs to be within three absolute deviations from the median number of reads per BC (Fig.3C). Alternatively, the user can provide a target sequencing depth, and *zUMIs* will downsample to the specified read number or omit the cell from the downsampled count table if fewer reads were present. Furthermore, *zUMIs* also allows the user to specify a 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 example dataset, the number of detected genes starts leveling off at 1 million reads. Sequencing double that amount would only increase the number of detected genes from 9,000 to 10,600 when counting exon reads (Fig.3D). In line with previous findings [8, 14], the saturation curve of exon+intron counting runs parallel to the one for exon counting, both indicating that a sequencing depth of 1 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 counting, one for intron and one for exon+intron counts. If a user chooses the downsampling option, six 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 (Fig.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 less than 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 (Fig.3E). We also observed that the peak random access memory usage for processing datasets of 227, 500, and 1,000 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 30 billion read-pairs generated with 10xGenomics Chromium (see Methods section) on a 22-core processor in only 7 days.

Intron counting

Recently, it has been shown that intron mapping reads in RNA-seq likely originate from nascent mRNAs and are useful for gene expression estimates [31, 32]. Additionally, novel approaches leverage the ratios of intron and exon mapping reads to infer information on transcription dynamics and cell states [33]. To address this new aspect of analysis, *zUMIs* also counts and collapses intron-only mapping reads as well as intron and exon mapping reads from the same gene with the same UMI. To assess the information gain from intronic reads to estimate gene expression levels, we analyzed a publicly available DroNc-seq dataset from mouse brain ([12]; see Methods section). 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 from 1,041 for exon counts to 1,995 for exon+intron counts. Next, we used the resulting UMI count tables to investigate whether exon+intron counting improves the identification of cell types, as suggested by Lake et al. [11]. The validity and accuracy of counting introns for single-nucleus sequencing methods has recently been demonstrated [34]. Following the Seurat pipeline to cluster cells [35, 36], we find that using exon+intron counts discriminates 28 clusters, while we could only discriminate 19 clusters using exon counts (Fig.4A, 4B). The larger number of clusters is not simply due to the increase in the counted UMIs and genes. When we permute the intron counts across cells and add them to the exon counts, the added noise actually reduces the number of identifiable clusters (Fig.4E).

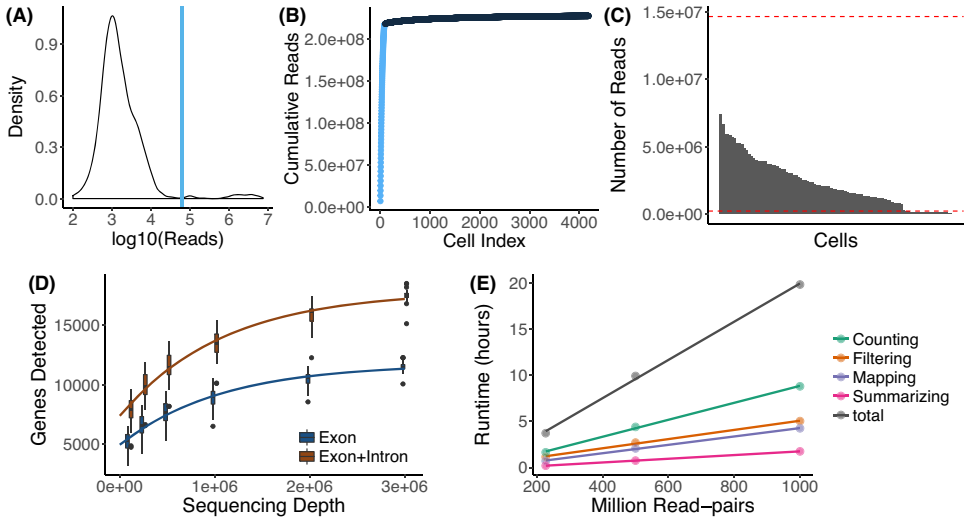


Figure 3: Utilities of zUMIs. Each of the panels shows the utilities of *zUMIs* pipeline. The plots from A–D show the results from the example HEK dataset used here. **(A)** The plot shows a density distribution of reads per BC. Cell BCs with reads right of the blue line are selected. **(B)** The plot shows the cumulative read distribution in the example HEK dataset where the BCs in light blue are the selected cells. **(C)** The bar plot shows the number of reads per selected cell BC with the red lines showing upper and lower median absolute deviation (MAD) 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 are shown. **(E)** Runtime for three datasets with 227, 500, and 1,000 million read-pairs. The runtime is divided in the main steps of the *zUMIs* pipeline as follows: filtering, mapping, counting, and summarizing. Each dataset was processed using 16 threads ("-p 16").

We continue to further characterize the seven clusters that were subdivided by the addition of intron counts (Fig.4D). First, we identify DE genes between the newly formed clusters. If we count only exon reads, there appear to be, on average, only 10 DE genes between the subgroups, while exon+intron counting yields ~10 times more DE genes, thus corroborating the signal found with clustering. The log2-fold changes of those additional DE genes estimated with either counting strategy are generally in good agreement; especially large log2-fold changes are detected with both exon and exon+intron counting (Fig.4F). Genes that are detected as DE in only one of our counting strategies have small log2-fold changes, and there are more of these small changes detected using exon+intron counting.

Detecting more genes naturally increases the chance to also detect more informative genes. Here, we cross-reference the gene list with marker genes for transcriptomic subtypes detected for major cell types of the mouse brain [37] and find that ~5% of the additional genes are also marker genes, which corresponds well to the general frequency of marker genes among the detected genes (4%). In the same vein, we also detect proportionally more DE genes with exon+intron counting compared to exon counting. Thus, including introns simply allows us to better detect present transcripts, while leaving the proportions of interest unaltered. Having a closer look at cluster 7, it was split into a bigger (7) and a smaller cluster (24) using exon+intron counting (Fig.4A-C), we find one marker gene (Il1rapl2) to be DE between the subclusters using exon+intron counting, while Il1rapl2 had only spurious counts using exon counts. Il1rapl2 is a marker for transcriptomic subtypes of GABAergic Pvalb-type neurons [37], suggesting that the split of cluster 7 might be biologically meaningful (Fig.4E).

In order to evaluate the power gained by exon+intron counting in a more systematic way, we perform power simulations using empirical mean and dispersion distributions from the largest and most uniform cluster (~1,500 cells) [9]. For a fair comparison, we include all detected genes, which is equivalent to the number of genes detected with exon+intron counting. Also, since we call a gene detected as soon as one count is associated, exon counting is necessarily a subset of exon+intron. Thus, there are, on average, 4 times more genes in the lowest expression quantile for exon counting than for exon+intron counting (Fig.4H). For those genes, expression is too spurious to be used for differential expression analysis; for exon+intron counting, we have, on average, 60% power to detect a DE gene in the first mean expression bin with a well-controlled false discovery rate (FDR) (Fig.4G). In summary, the increased power for exon+intron counting and probably also the larger number of clusters are due to better detection of lowly expressed genes. Furthermore, we think that although potentially noisy, the large number of additionally detected genes makes exon+intron counting worthwhile, especially for single-nuclei sequencing techniques that are enriched for nuclear nascent RNA transcripts, such as DroNc-seq [12]. Additionally, exon+intron counting may help in extracting as much information as possible from low coverage data as generated in the context of high-throughput cell atlas efforts (e.g., 10,000–20,000 reads/cell [38, 39]. Last, users should always exclude the possibility of intronic reads stemming from genomic DNA contamination in the library preparation by confirming low inter-

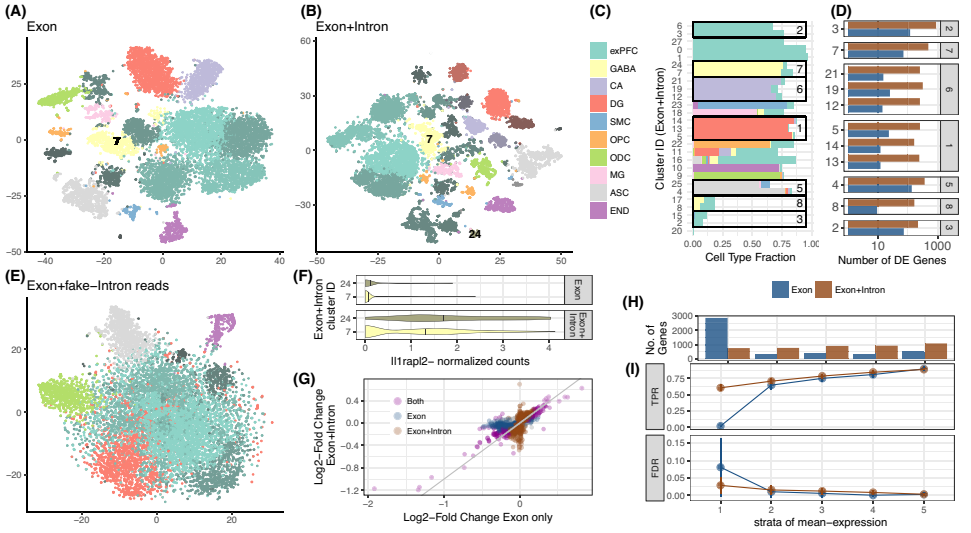


Figure 4: Contribution of intron reads to biological insights. We analyzed published single-nucleus RNA-seq data from mouse prefrontal cortex (PFC) and hippocampus [12] to assess the utility of counting intron in addition to exon reads. We processed the raw data with *zUMIs* to obtain expression tables with exon reads as well as exon+intron reads and then used the R-package Seurat [35, 36] to cluster cells. With exon counts, we identified 19 clusters **(A)**, and with exon+intron counts we identified 27 clusters **(B)**. Clusters are represented as t-SNE plots and colored according to the most frequent cell-type assignment in the original article [12]: glutamatergic neurons from the prefrontal cortex (exPFC), GABAergic interneurons (GABA), pyramidal neurons from the hippocampal CA region (CA), granule neurons from the hippocampal dentate gyrus region (DG), astrocytes (ASC), microglia (MG), oligodendrocytes (ODC), oligodendrocyte precursor cells (OPC), neuronal stem cells (NSC), smooth muscle cells (SMC) and endothelial cells (END). Different shades of those clusters indicate that multiple clusters had the same major cell type assigned. If we randomly sample counts from the intron data and add them to the exon counting, the noise reduces the number of clusters and the Seurat pipeline can only identify 9–11 clusters **(E)**. The composition of each cluster based on exon+intron is detailed in panel **(C)**, and cells that were not assigned a cell type in [12] are displayed as empty. The boxes mark the clusters that were not split when using exon data only. For example, cluster 7 from exon counting, which mainly consists of GABAergic neurons, was split into clusters 7, 24 (506, 66 cells) when using exon+intron counting. In **(D)**, we show the numbers of genes that were differentially expressed (DE) (limma p-adj <0.05) between the clusters found only with exon+intron counts. The panel numbers represent the exon counting cluster numbers and the y-axis the exon+intron counting cluster number. The log2-fold changes corresponding to these contrasts are also used in **(G)**. Among the genes that were additionally detected to be DE by exon+intron counting was the marker gene Il1rapl2 (limma p-adj $= 10^{-5}$). In **(F)**, we present a violin plot of the normalized counts for Il1rapl2 in cells of the GABAergic subclusters 7 and 24. Log2-fold changes calculated with exon+intron counts correlate well with exon counts **(G)**. Note that for exon counting only, half as many genes could be evaluated as for exon+intron counting and thus only half of the exon+intron genes are depicted in **(G)**. Large $\log_2$ fold changes (LFCs) are found to be significant with both counting strategies (purple points are close to the bisecting line). We conducted simulations based on mean and dispersion measured using exon cluster 0 (1,616 cells, ~90% exPFC). In **(I)** we show the expected true positive rate and the false discovery rate for a scenario comparing 300 vs 300 cells. Results for exon and exon+intron counting were stratified into five quantiles according to the mean expression of genes, where stratum 1 contains lowly expressed genes and stratum 5 the most highly expressed genes. The numbers of genes falling into each of the bins using exon+intron and exon counting are depicted in **(H)**.

genic mapping fractions using the statistics output provided by *zUMIs*.

Conclusion

zUMIs is a fast and flexible pipeline for 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 BC 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.

Methods

Analyzed RNA-seq datasets

HEK293T cells were cultured in DMEM high glucose with L-glutamine (Biowest) supplemented with 10% fetal bovine serum (Thermo Fisher) and 1% penicillin/streptomycin (Sigma-Aldrich) in a 37°C incubator with 5% carbon dioxide. Cells were passaged and split every 2 or 3 days. For single-cell RNA-seq, HEK293T cells were dissociated by incubation with 0.25% Trypsin (Sigma-Aldrich) for 5 minutes at 37°C. The single-cell suspension was washed twice with phosphate-buffered saline, and dead cells were stained with Zombie Yellow (Biolegend) according to the manufacturer's protocol. Single cells were sorted into DNA LoBind 96-well polymerase chain reaction (PCR) plates (Eppendorf) containing lysis buffer with a Sony SH-800 cell sorter in 3-drop purity mode using a 100-µmnozzle. Next, single-cell RNA-seq libraries were constructed from one 96-well plate using a slightly modified version of the mcSCRB-seq protocol. Reverse transcription was performed as described previously [40], with the only change being the use of KAPA HiFi HotStart enzyme for

PCR amplification of cDNA. Resulting libraries were sequenced using an Illumina HiSeq1500 with 16 cycles in Read 1 to decode cell BCs (6 bases) and UMIs (10 bases) and 50 cycles in Read 2 to sequence into the cDNA fragment, obtaining ∼227 million reads. Raw fastq files were processed using *zUMIs*, mapping to the human genome (hg38) and Ensembl gene models (GRCh38.84).

In addition, we analyzed data from 1.3 million mouse brain cells generated on the 10xGenomics Chromium platform [2]. Sequences were downloaded from the National Center for Biotechnology Information Sequence Read Archive under accession number SRP096558. The data consist of 30 billion read-pairs from 133 individual samples. In these data, read 1 contains 16 bp for the cell BC and 10 bp for the UMI and read 2 contains 114 bp of cDNA. *zUMIs* was run using default settings, and we allowed 7 threads per job for a total of up to 42 threads on an Intel Xeon E5-2699 22-core processor.

Finally, we obtained mouse brain DroNc-seq read data [12] from the Broad Institute Single Cell Portal [41]. This dataset consists of ∼1,615 million read-pairs from ∼11,000 single nuclei. Read 1 contains a 12 bpcell BC and a 8 bpUMI and read 2 60 bpof cDNA.

The two mouse datasets were mapped to genome version mm10 and applying Ensembl gene models (GRCm38.75).

Power simulations and DE analysis

We evaluated the power to detect differential expression with the help of the powsimR package [9]. For the DroNc-seq dataset, we estimated the parameters of the negative binomial distribution from one of the identified clusters, namely, cluster 0, compromising 1,500 glutamatergic neuronal cells from the prefrontal cortex (Fig.4D). Since we detect more genes with exon+intron counting (4,433 compared to 1,782), we included this phenomenon in our read count simulation by drawing mean expression values for a total of 4,433 genes. This means that the table includes sparse counts for the exon counting. Log_2-fold changes were drawn from a gamma distribution with shape equal to 1 and scale equal to 2. In each of the 25 simulation iterations, we draw an equal sample size of 300 cells per group and test for differential expression using limma-trend [42] on log_2 counts per million (CPM) values with scran [43] library size correction. The true positive rate and FDR are stratified over the empirical mean expression quantile bins.

For the differential expression analysis between clusters, we use the same DE estimation procedure as in the simulations: scran normalization followed by limma-trend DE-analysis (c.f. [44]).

Cluster identification

After processing the DroNc-seq data [12] with zUMIs as described above, we cluster cells based on UMI counts derived from exons only and exons+introns reads using the Seurat pipeline [35, 36]. First, cells with fewer than 200 detected genes were filtered out. The filtered data were normalized using the LogNormalize function. We then scale the data by regressing out the effects of the number of transcripts and genes detected per cell using the ScaleData function. The normalized and scaled data are then used to identify the most variable genes by fitting a relationship between mean expression (ExpMean) and dispersion (LogVMR) using the FindVariableGenes function. The identified variable genes are used for principle component analysis, and the top 20 principle components are then used to find clusters using graph-based clustering as implemented in FindClusters. To illustrate that the additional clusters found by counting exon+intron reads are not spurious, we use intron-only UMI counts from the same data to add to the observed exon-only counts. More specifically, to each gene we add scran-size factor-corrected intron counts from the same gene after permuting them across cells. We assessed the cluster numbers from 100 such permutations.

Comparison of UMI collapsing strategies

In order to validate *zUMIs* and compare different UMI collapsing methods, we used the HEK dataset described above. We ran *zUMIs* (1) without quality filtering, (2) filtering for onebase under Phred 17, and (3) collapsing similar UMI sequences within a hamming distance of 1. To compare with other available tools, we ran the same dataset using the Drop-seq-tools version 1.13 [13] and quality filter "1 base under Phred 17" without edit distance collapsing. Last, the HEK dataset was used with UMI tools [25] in (1) "unique" and (2) "directional adjacency" mode with edit distance set to 1. Also, we compared the output of *zUMIs* from the DroNc-seq dataset when using default parameters ("1 base under Phred 20") to UMI-tools in (1) "unique," (2) "directional adjacency," and (3) "cluster" settings. For each setting and tool combination, we compared per-cell/per-nuclei UMI contents in a linear model fit.

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, Rsubread >= 1.26.1, pigz >= 2.3 & samtools >= 1.1
- License: GNU GPLv3.0
- Research Resource Identification Initiative ID: SCR_016139

Availability of supporting data

All data that were generated for this project were submitted to GEO under accession GSE99822. An archival copy of the source code and test data are available via the *GigaScience* repository GigaDB [45].

Abbreviations

BC: barcode; DE: differentially expressed; FDR: false discovery rate; MAD: median absolute deviation; PCR: polymerase chain reaction; PFC: prefrontal cortex; scRNA-seq: single-cell RNA sequencing; UMI: unique molecular identifier.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) through SFB1243 subprojects A14/A15.

Author contributions

S.P. and C.Z. designed and implemented the pipeline. B.V. tested the pipeline and helped in power simulations. All authors contributed to writing the manuscript.

References

1. Sandberg R. Entering the era of single-cell transcriptomics in biology and medicine. Nat Methods 2014;**11**(1):22–4.
2. Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;**8**:14049.
3. Rosenberg AB, Roco CM, Muscat RA, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science 2018;**360**(6385):176–82.
4. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol 2016;**34**(11):1145–60.
5. Regev A, Teichmann SA, Lander ES, et al. The Human Cell Atlas. Elife, 2017; 6.
6. Parekh S, Ziegenhain C, Vieth B, et al. The impact of amplification on differential expression analyses by RNA-seq. Sci Rep 2016;**6**:25533.
7. Kivioja T, Vähärautio A, Karlsson K, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 2012;**9**(1):72–4.
8. Ziegenhain C, Vieth B, Parekh S, et al. Comparative analysis of single-cell RNA sequencing methods. Mol Cell 2017;**65**(4):631–43.e4.
9. Vieth B, Ziegenhain C, Parekh S, et al. powsimR: power analysis for bulk and single cell RNA-seq experiments. Bioinformatics 2017;**33**(21):3486–3488.
10. Ziegenhain C, Vieth B, Parekh, et al. Quantitative single-cell transcriptomics. Brief Funct Genomics 2018; doi:10.1093/bfgp/ely009.
11. Lake BB, Ai R, Kaeser GE, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science 2016;**352**(6293):1586–90.
12. Habib N, Avraham-Davidi I, Basu A, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods 2017;**14**(10):955–8.
13. Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015;**161**(5):1202–14.
14. Svensson V, Natarajan KN, Ly LH, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Methods 2017;**14**(4):381–7.
15. Hashimshony T, Senderovich N, Avital G, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-seq. Genome Biol 2016;**17**(1):77.
16. Petukhov V, Guo J, Baryawno N, et al. Accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. bioRxiv 2017;p. 171496.
17. Soumillon M, Cacchiarelli D, Semrau S, et al. Characterization of directed differentiation by high-throughput single-cell RNA-seq. bioRxiv 2014.
18. Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 2014;**343**(6172):776–9.
19. Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 2015;**161**(5):1187–1201.
20. Zilionis R, Nainys J, Veres A, et al. Single-cell barcoding and sequencing using droplet microfluidics. Nat Protoc 2017;**12**(1):44–73.
21. Hochgerner H, Lönnerberg P, Hodge R, et al. STRT-seq-2i: dual-index 5' single cell and nucleus RNA-seq on an addressable microwell array. Sci Rep 2017;**7**(1):16327.
22. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;**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;**30**(7):923–30.
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. Smith TS, Heger A, Sudbery I. UMI-tools: modelling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome Res. 2017.
26. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 2002;**97**(458):611–31.
27. Fraley C, Raftery AE Enhanced Model-Based Clustering, Density Estimation and Discriminant Analysis Software: MCLUST. , J. Classification , 2003, **20**, 263–286.
28. Vallejos CA, Risso D, Scialdone A, et al. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods 2017;**14**(6):565–71.
29. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. Brief Bioinform 2017.
30. Grün D, van Oudenaarden A. Design and analysis of single-cell sequencing experiments. Cell 2015;**163**(4):799–810.
31. Hendriks GJ, Gaidatzis D, Aeschimann F, et al. Extensive oscillatory gene expression during *C. elegans* larval development. Mol Cell 2014;**53**(3):380–92.
32. Gaidatzis D, Burger L, Florescu M, et al. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. Nat Biotechnol 2015;**33**(7):722–9.
33. La Manno G, Soldatov R, Hochgerner H, et al. RNA velocity in single cells. bioRxiv 2017;p. 206052.
34. Lake BB, Codeluppi S, Yung YC, et al. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. Sci Rep 2017;**7**(1):6031.
35. Satija R, Farrell JA, Gennert D, et al. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015;**33**(5):495–502.
36. Butler A, Satija R. Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. bioRxiv 2017;p. 164889.
37. Tasic B, Menon V, Nguyen TN, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci 2016;**19**(2):335–46.
38. The Tabula Muris Consortium, Quake SR, Wyss-Coray T, et al. Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a Tabula Muris. bioRxiv 2018;p. 237446.
39. Han X, Wang R, Zhou Y, et al. Mapping the mouse cell atlas by microwell-seq. Cell 2018;**172**(5):1091–1107.e17.
40. Bagnoli JW, Ziegenhain C, Janjic A, et al. mcSCRB-seq: sensitive and powerful single-cell RNA sequencing. bioRxiv 2017;p. 188367.
41. Broad Institute Single Cell Portal. https://portals.broadinstitute.org/single_cell/study/dronc-seq-single-nucleus-rna-seq-on-mouse-archived-brain.

42. Law CW, Chen Y, Shi W, et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014;**15**(2):R29.
43. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res 2016;**5**.
44. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods 2018;**15**(4):255–61.
45. Parekh S, Ziegenhain C, Vieth B, et al. Supporting data for 'zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs.' GigaScience Database 2018;http://dx.doi.org/10.5524/100447.
46. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods 2014;**11**(6):637–40.
47. Tian L, Su S, Amann-Zalcenstein D, et al. scPipe: a flexible data preprocessing pipeline for single-cell RNA-sequencing data. bioRxiv 2017;p. 175927.
48. Islam S, Zeisel A, Joost S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 2014;**11**(2):163–6.

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

Supplementary Information

by

Swati Parekh [1,2*], Christoph Ziegenhain [1,*], Beate Vieth [1], Wolfgang Enard [1] and Ines Hellmann [1,2]

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

[*]Contributed equally

[2]Corresponding author

1 Tools for UMI counting in RNA-seq

Name	Reference	Open Source	Quality UMI/BC	Mapper	intron	Down-sampling	Compatibility with UMI library protocols
Cell Ranger	Zheng et al. 2017	yes	yes	STAR	no	yes	Zheng et al. 2017
Drop-seq	Macosko et al. 2015	no	yes	STAR	no	no	Macosko et al. 2015, Soumillon et al. 2014, Hashimshony et al. 2016
CEL-seq	Hashimshony et al. 2016	yes	yes	bowtie2	no	no	Grün et al. 2014, Hashimshony et al. 2016
umis	Svensson et al. 2017	yes	no	Kallisto	no	no	Soumillon et al. 2014, Grün et al. 2014, Islam et al. 2014, Jaitin et al. 2014, Macosko et al. 2015, Klein et al. 2015, Zheng et al. 2017
UMI-tools	Smith et al. 2017	yes	yes	BWA	no	no	Soumillon et al. 2014, Klein et al. 2015
zUMIs	This work	yes	yes	STAR	yes	yes	Soumillon et al. 2014, Grün et al. 2014, Islam et al. 2014, Jaitin et al. 2014, Macosko et al. 2015, Hashimshony et al. 2016, Hochgerner et al. 2017, Habib et al. 2017, Rosenberg et al. 2017, Zheng et al. 2017

Table S1: Features of available tools that can handle UMIs for the quantification of gene expression data. The evaluated features are whether the tool is open source, it considers the sequence quality for cell barcode (BC) and UMI, which mapper it uses, whether it can consider intron mapping reads for counting, whether it offers a utility to make varying library sizes more comparable via downsampling and finally with which RNA-seq library preparation protocols it is compatible.

2 Characterization of zUMIs

To demonstrate the utility of *zUMIs*, we processed data generated from 96 HEK cells using the SCRB-seq protocol Soumillon et al., 2014; Ziegenhain et al., 2017.
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.

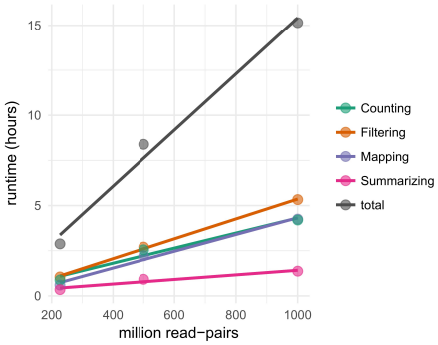


Figure S1: *zUMIs* runtime for three datasets with 227, 500 and 1000 million read-pairs. The runtime is divided in the main steps of the *zUMIs* pipeline: Filtering, Mapping, Counting and Summarizing. Each dataset was processed using up to 16 threads ("-p 16").

3 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).

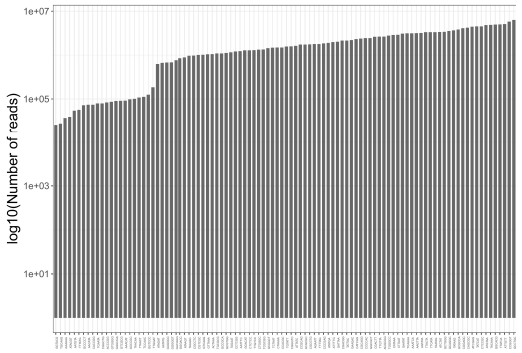


Figure S2: Reads per barcode. Bars show the number of reads assigned to each sample barcode.

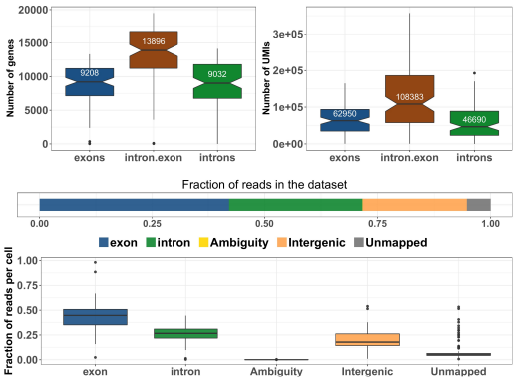


Figure S3: Summary statistics. The boxplot in the left panel shows number of genes (left) and number of UMIs(right) detected per barcode while considering only intronic/exonic counts and intornic+exonic counts. The horizontal relative barplot in the middle indicates total fraction of reads assignment 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.

4 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 Grün and van Oudenaarden 2015. 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).

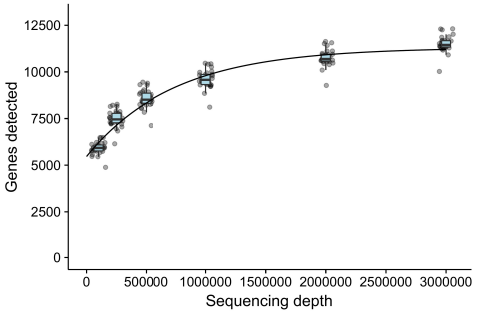


Figure S4: Downsampling. Cells were downsampled to six depths from 100,000 to 3,000,000 reads. For each sequencing depth the reads detected per cell is shown. Here the increase in the number of genes detected using 1 million as compared to 3 million reads is small, suggesting that 1 million reads per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

5 Cell barcode selection

In order to be compatible with well-based and droplet-based scRNA-seq methods, *zUMIs* is flexible with handling of cell barcodes. As default behavior, zUMIs tries to estimate the relevant barcodes from the data using the cumulative read distribution presented as "knee plot" in (Figure S5). For this, the cumulative read numbers of all observed barcodes are calculated, ordered descending by number of obtained reads per barcode. Only those barcodes are retained as real, as long as their difference to the neighboring barcode accounts for at least 1% of the maximal difference observed. If this does not account for at least 90% of total sequenced reads, the threshold is lowered to 0.1%. The dataset used in this paper has known 96 cells and when we let zUMIs select the barcodes with this method, it correctly identifies these cells. To override automatic detection of barcodes, users can either give a fixed number of barcodes to consider using "-b 100" key or refer to a plain text file containing known expected barcodes as "-b barcodefile.txt" where the file contains a list of barcodes without a header.

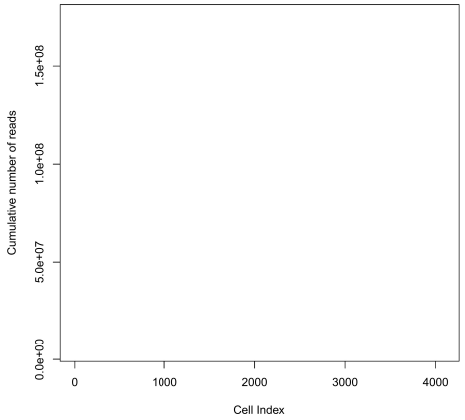


Figure S5: Knee plot. The plot shows cumulative read distribution in the example HEK dataset where the barcodes in red are the selected cells.

References

Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, 16 January 2017.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 21 May 2015. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2015.05.002.

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.

Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, Yuval Dor, Aviv Regev, and Itai Yanai. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, 17(1):77, 28 April 2016.

Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat. Methods*, 11(6):637–640, June 2014.

Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods*, 6 March 2017.

Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11(2):163–166, February 2014.

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, and Ido Amit. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 14 February 2014.

Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 21 May 2015.

Tom Sean Smith, Andreas Heger, and Ian Sudbery. UMI-tools: Modelling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.*, 18 January 2017.

Hannah Hochgerner, Peter Lönnerberg, Rebecca Hodge, Jaromir Mikes, Abeer Heskol, Herman Hubschle, Philip Lin, Simone Picelli, Gioele La Manno, Michael Ratz, Jude Dunne, Syed Husain, Ed Lein, Maithreyan Srinivasan, Amit Zeisel, and Sten Linnarsson. STRT-seq-2i: dual-index 5 single cell and nucleus RNA-seq on an addressable microwell array. 20 April 2017.

Naomi Habib, Anindita Basu, Inbal Avraham-Davidi, Tyler Burks, Sourav R Choudhury, Francois Aguet, Ellen Gelfand, Kristin Ardlie, David A Weitz, Orit Rozenblatt-Rosen, Feng Zhang, and Aviv Regev. DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq. 9 March 2017.

Alexander B Rosenberg, Charles Roco, Richard A Muscat, Anna Kuchina, Sumit Mukherjee, Wei Chen, David J Peeler, Zizhen Yao, Bosiljka Tasic, Drew L Sellers, Suzie H Pun, and Georg Seelig. Scaling single cell transcriptomics through split pool barcoding. 2 February 2017.

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.

Dominic Grün and Alexander van Oudenaarden. Design and analysis of Single-Cell sequencing experiments. *Cell*, 163(4):799–810, 5 November 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.10.039.

2.4.1 Updates to zUMIs

My colleagues Swati Parekh, Christoph Ziegenhain and Ines Hellmann regularly improve and update zUMis since its publication. I help with testing implementations and changes as well as with documentation.
The most significant changes are listed below:

- Setup of all parameters in a convenient YAML config file, also implemented as a R shiny app.
- Compatibility with non-UMI protocols, such as Smart-seq2, and paired end cDNA reads.
- Increased processing speed (at least 2x faster)
- Parallelize filtering step
- Parallelize Hamming distance UMI collapsing
- Extensive use of the R package data.table and its pipes for constructing gene expression matrices
- Mapping by piping STAR SAM output into threaded samtools BAM compression
- Possibility to integrate transgenes or external references like ERCC spike ins on the fly.

Minor changes include:

- Optional downstream velocyto analysis (La Manno et al. 2018)
- Optional output of zUMIs results in loomR format (Butler et al. 2018)
- Automatic barcode detection guided by incorporating cell barcode whitelists
- Extracting cell barcode sequences from fastq files with frameshifts (e.g. ddSeq protocol [212])

2.5 A Systematic Evaluation of Single Cell RNA-Seq Analysis Pipelines

Vieth B, Parekh S, Ziegenhain C, Parekh S, Enard W, Hellmann I:

"A Systematic Evaluation of Single Cell RNA-Seq Analysis Pipelines" (2019)

Nature Communications 10 (1):4667-4678.

doi: 10.1038/s41467-019-12266-7

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

https://www.nature.com/articles/s41467-019-12266-7#Sec16

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-12266-7 OPEN

A systematic evaluation of single cell RNA-seq analysis pipelines

Beate Vieth [1], Swati Parekh [2], Christoph Ziegenhain[3], Wolfgang Enard [1] & Ines Hellmann [1*]

The recent rapid spread of single cell RNA sequencing (scRNA-seq) methods has created a large variety of experimental and computational pipelines for which best practices have not yet been established. Here, we use simulations based on five scRNA-seq library protocols in combination with nine realistic differential expression (DE) setups to systematically evaluate three mapping, four imputation, seven normalisation and four differential expression testing approaches resulting in ~3000 pipelines, allowing us to also assess interactions among pipeline steps. We find that choices of normalisation and library preparation protocols have the biggest impact on scRNA-seq analyses. Specifically, we find that library preparation determines the ability to detect symmetric expression differences, while normalisation dominates pipeline performance in asymmetric DE-setups. Finally, we illustrate the importance of informed choices by showing that a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the sample size.

[1] Anthropology and Human Genomics, Department of Biology II, Ludwig-Maximilians University, Munich, Germany. [2] Max Planck Institute for Biology of Ageing, Cologne, Germany. [3] Department of Cell and Molecular Biology, Karolinska Institutet, SE-171 65, Stockholm, Sweden. *email: hellmann@bio.lmu.de

Many experimental protocols and computational analysis approaches exist for single cell RNA sequencing (scRNA-seq). Furthermore, scRNA-seq analyses can have different goals including differential expression (DE) analysis, clustering of cells, classification of cells and trajectory reconstruction[1]. All these goals have the first analysis steps in common in that they require expression counts or normalised counts. Here, we focus on these important first choices made in any scRNA-seq study, using DE-inference as performance readout. Benchmarking studies exist only separately for each analysis step, which are library preparation protocols[2,3], alignment[4,5], annotations[6], count matrix preprocessing[7,8] and normalisation[9]. However, the impact of the combined choices of the separate analysis steps on overall pipeline performance has not been quantified. In order to achieve a fair and unbiased comparison of computational pipelines, simulations of realistic data sets are necessary. This is because the ground truth of real data is unknown and alternatives, such as concordance analyses are bound to favour similar and not necessarily better methods.

To this end, we integrate popular methods for each analysis step into our simulation framework powsimR[10]. As the basis for simulations, powsimR uses raw count matrices to describe the mean-variance relationship of gene expression measures. This includes the variance introduced during the experiment itself as well as extra variance due to the first to computational steps of expression quantification. Adding DE then provides us with detailed performance measures based on how faithfully DE-genes can be recovered.

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

Realistic simulations in conjunction with a wide array of scRNA-seq methods, allow us not only to quantify the performance of individual pipeline steps, but also to quantify interdependencies among the steps. Moreover, the relative importance of the various steps to the overall pipeline can be estimated. Hence, our analysis provides sound recommendations regarding the construction of an optimal computational scRNA-seq pipeline for the data at hand.

Results

scRNA-seq data and simulations. The starting point for our comprehensive pipeline comparison is a representative selection of scRNA-seq library preparation protocols (Fig. 1a). Here, we included one full-length method (Smart-seq2[13]) and four UMI methods[2,14–16]. The UMI strategies encompass two plate-based (SCRB-seq, CEL-seq2) and the most common non-commercial and commercial droplet-based protocols (Drop-seq, 10X Chromium). CEL-seq2 differs from SCRB-seq in that it relies on linear amplification by in vitro transcription, while SCRB-seq relies on PCR amplification using the same strategy as 10X Chromium (see Ziegenhain et al.[2,17] for a detailed discussion). We then combine the library preparation protocols with three mapping approaches[18–20] and three annotation schemes[21–23] resulting in 45 distinct raw count matrices (see "Methods"). We simulated 27 distinct DE-setups per matrix, each with 20 replicates, resulting in a total of 19,980 simulated data sets (Fig. 1b).

Genome-mapping quantifies gene expression with high accuracy. We first investigated how expression quantification is affected by different alignment methods using our selection of scRNA-seq experiments. For each of the three following strategies we picked one the most popular methods (Supplementary Fig. 2): (1) alignment of reads to the genome using splice-aware alignment (STAR[18]), (2) alignment to the transcriptome (BWA[19]) and (3) pseudo-alignment of reads guided by a transcriptome (kallisto[24]).We then combined these with three annotation schemes including two curated schemes (RefSeq[21] and Vega[23]) and the more inclusive GENCODE[22] (Supplementary Table 2).

First, we assessed the performance by the number of reads or UMIs that were aligned and assigned to genes (Fig. 2a and Supplementary Fig. 3). Alignment rates of reads are comparable across all scRNA-seq protocols. Assignment rates on the other hand show some interaction between mapper and protocol. All mappers, aligned and assigned more reads using GENCODE as compared to RefSeq annotation, whereas the pseudo-aligner kallisto profited most from the more comprehensive annotation of GENCODE and here in particular the 3'UMI protocols (Figure 2A). Generally, STAR in combination with GENCODE aligned (82–86%) and assigned (37–63%) the most reads, while kallisto assigned consistently the fewest reads (20-40%) (Figure 2D). BWA assigned an intermediate fraction of reads (22–44%), but—suspiciously—these were distributed across more UMIs. As reads with the same UMI are more likely to originate from the same mRNA molecule and thus the same gene, the average number of genes with which one UMI sequence is associated, can be seen as a measure of false mapping. Indeed, we find that the same UMI is associated with more genes when mapped by BWA than when mapped by STAR (Fig. 2b). This indicates a high false mapping rate, that probably inflates the number of genes that are detected by BWA (Fig. 2c and Supplementary Fig. 4).

This said, it remains to be seen what impact the differences in read or UMI counts obtained through the different alignment strategies and annotations have on the power to detect DE-genes.

As already indicated from the low fraction of assigned reads, kallisto has the lowest mean expression and the highest gene dropout rates (Fig. 2d and Supplementary Figs. 7 and 8) and, as expected from a high fraction of falsely mapped reads, BWA has the largest variance. To estimate the impact that these statistics have on the power to detect DE-genes, we use the mean-variance relationship to simulate data sets with DE-genes (Fig. 2d, e). As previously reported[2], UMI protocols have a noticeably higher power than Smart-seq2 (Fig. 2f). Moreover for Smart-seq2, we find that kallisto, especially with RefSeq annotation, performs slightly better than STAR, while for UMI-methods STAR performs better (Fig. 2f and Supplementary Fig. 9).

In summary, using BWA to map to the transcriptome introduces noise, thus considerably reducing the power to detect DE-genes as compared to genome alignment using STAR or the pseudo-alignment strategy kallisto, but given the lower mapping rate of kallisto STAR with GENCODE is generally preferable.

Many asymmetric changes pose a problem without spike-ins. The next step in any RNA-seq analysis is the normalisation of the count matrix. The main idea here is that the resulting size factors correct for differing sequencing depths. In order to improve normalisation, spike-ins as an added standard can help, but are not feasible for all scRNA-seq library preparations. Another avenue to improve normalisation would be to deal with sparsity

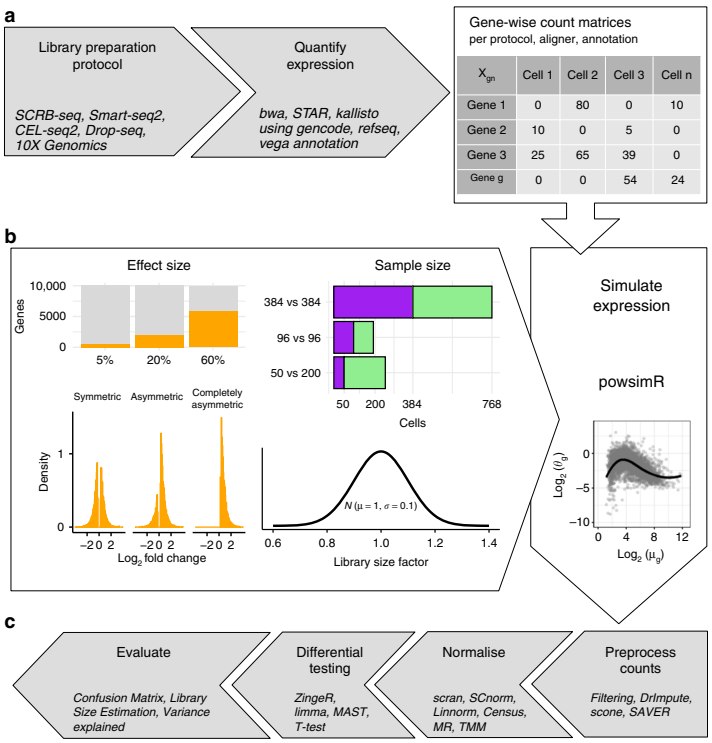


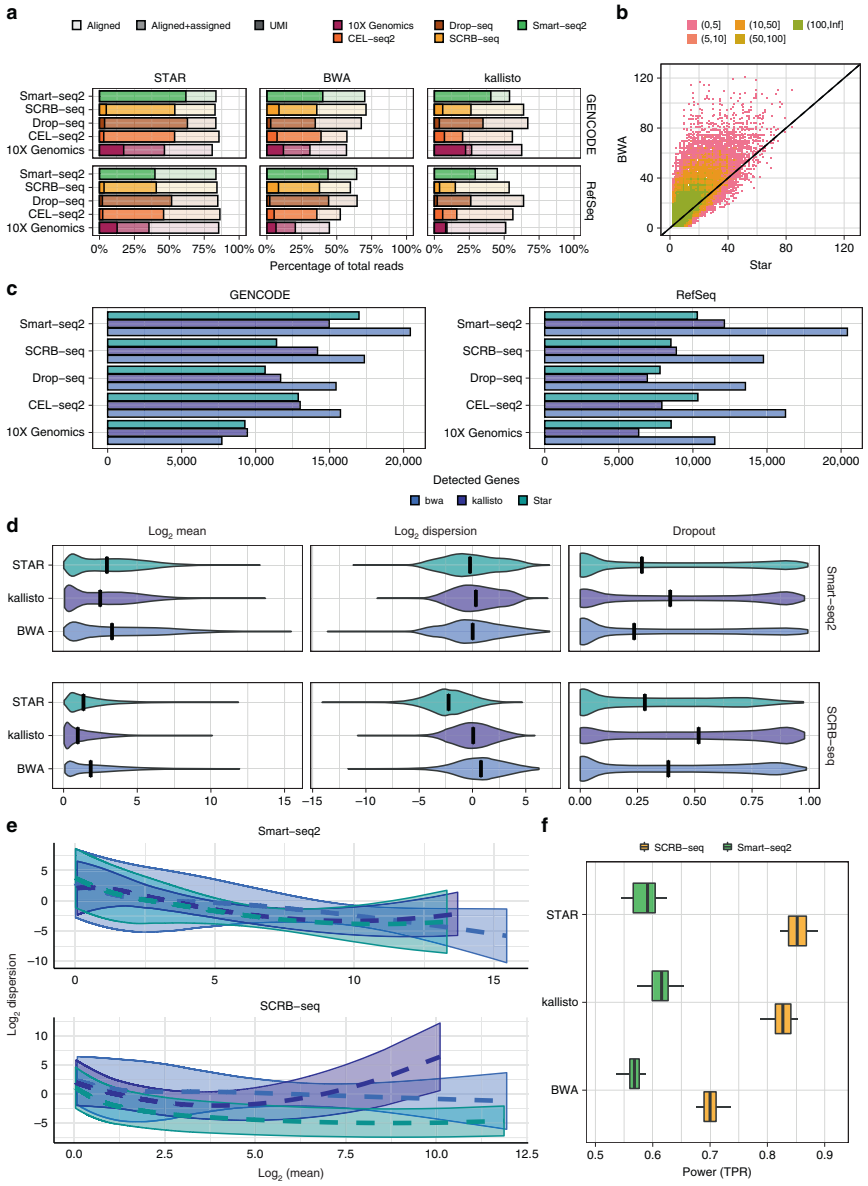
Fig. 1 Study Overview. **a** The data sets yielding raw count matrices: We use scRNA-seq data sets[2,16] representing 5 popular library preparation protocols. For each data set, we obtain multiple gene count matrices that result from various combinations of alignment methods and annotation schemes (see also Supplementary Figs. 1 and 2, and Supplementary Tables 1 and 2). **b** The simulation setup: Using powsimR[10] distribution estimates from real count matrices, we simulate the expression of 10,000 genes for two groups with 384 vs 384, 96 vs. 96 and 50 vs. 200 cells, where 5, 20 or 60% of genes are DE between groups. The magnitude of expression change for each gene is drawn from a narrow gamma distribution ($X \sim \Gamma(\alpha = 1, \beta = 2)$) and the directions can either be symmetric, asymmetric or completely asymmetric. To introduce slight variation in expression capture, we draw a different size factor for each cell from a narrow normal distribution. **c** The analysis pipeline: The simulated data sets are then analysed using combinations of four count matrix preprocessing, seven normalisation and four DE approaches. The evaluation of these pipelines focuses on the outcome of the confusion matrix and its derivatives (TPR, FDR, pAUC, MCC), deviance in library size estimates (RMSE) and computational run time

by imputing missing data prior to normalisation as discussed in the next chapter (Fig. 1c). To begin with, we compare how much the estimated size factors deviate from the truth. As long as there is only a small proportion of DE-genes or if the differences are symmetric, estimated size factors are not too far from the simulated ones and there are no large differences among methods (Fig. 3a and Supplementary Figs. 10 and 11). However with increasing asymmetry, size factors deviate more and more and the single cell methods scran[25] and SCnorm[26] perform markedly better than the bulk methods TMM[27], MR[28] and Positive Counts as well as the single cell method Linnorm[29]. Census[30] is an outlier in that it has a constant deviation of 0.1, which is due to filling in 1 when library sizes could not be calculated.

To determine the effect of these deviations on downstream analyses, we evaluated the performance of DE inference using different normalisation methods (Fig. 3b and Supplementary Figs. 12–15). Firstly, the differences in the TPR across normalisation methods are only minor, only Linnorm performed consistently worse (Supplementary Fig. 13). In contrast, the ability to control the FDR heavily depends on the normalisation method (Supplementary Fig. 14). For small numbers of DE-genes or symmetrically distributed changes, the FDR is well controlled for all methods except Linnorm. However, with an increasing number and asymmetry of DE-genes, only SCnorm and scran keep FDR control, provided that cells are grouped or clustered prior to normalisation. In our most extreme scenario with 60% DE-genes and complete asymmetry, all methods except Census loose FDR control. SCnorm, scran, Positive Counts and MR regain FDR control with spike-ins for 60% completely asymmetric DE-genes (Supplementary Fig. 14). Given similar TPR of the methods, this FDR control determines the pAUC (Fig. 3b, c).

Since in real data it is usually unknown what proportion of genes is DE and whether cells contain differing levels of mRNA, we recommend a method that is robust under all tested scenarios. Thus, for most experimental setups scran is a good choice, only for Smart-seq2 data without spike-ins, Census might be a better choice.

Imputation has little impact on pipeline performance. If the main reason why normalisation methods perform worse for



scRNA-seq than for bulk data is the sparsity of the count matrix, reducing this sparsity by either more stringent filtering or imputation of missing values should remedy the problem[31]. Here, we test the impact of frequency filtering and three imputation approaches (DrImpute[32], scone[33], SAVER[34]) on normalisation performance. Note, that we use the imputation or filtering only to obtain size factor estimates, that are then used together with the raw count matrix for DE-testing.

We find that simple frequency filtering has no effect on normalisation results (Fig. 3d). Performance as measured by pAUC is identical to using raw counts. In contrast, imputation can have an effect on performance and there are large differences

Fig. 2 Expression Quantification. **a** Read alignment and assignment rates per library preparation protocol stratified over aligner and annotation. The lighter shade represents the percentage of the total reads that could be aligned and the darker shade the percentage that also was uniquely assigned (see also Supplementary Fig. 3). For comparability, cells were downsampled to 1 million reads/cell, with the exception of 10× Genomics data that were only sequenced to on average 60,000 reads/cell. Hence, these data are farther from saturation and have a higher UMI/read ratio. **b** Number of genes per UMI with >1 reads for BWA and STAR alignment using the SCRB-seq data set and GENCODE annotation. Colours denote number bins of UMIs. **c** Number of genes detected per Library Preparation Protocol stratified over Aligner and Annotation (i.e. at least 10% nonzero expression values) (see also Supplementary Fig. 4). **d** Estimated mean expression, dispersion and gene dropout rates for SCRB-seq and Smart-seq2 data using STAR, BWA or kallisto alignments with GENCODE annotation (see also Supplementary Fig. 7). **e** Mean-dispersion fitting line applying a cubic smoothing spline with 95% variability bands for SCRB-seq and Smart-seq2 data using STAR, BWA or kallisto alignments with GENCODE annotation (see also Supplementary Fig. 8). **f** The effect of quantification choices on the power (TPR) to detect differential expression stratified over library preparation and aligner. The expression of 10,000 detected genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation per protocol. Five percent of the simulated genes are differentially expressed following a symmetric narrow gamma distribution. Unfiltered counts were normalised using scran. Differential expression was tested using limma-trend (see also Supplementary Fig. 9)

among methods. Imputation with DrImpute and scone rarely increased the pAUC and occasionally as in the case of SCRB-seq with MR normalisation, the pAUC even decreased by 100 and 76%, respectively due to worse FDR control relative to using raw counts (Supplementary Fig. 18). In contrast, these two imputation methods achieved an appreciable increase in pAUC together with scran normalisation, ~28, 4 and 9% for 10× Genomics, SCRB-seq and Smart-seq2 data, respectively. SAVER on the other hand never made things worse, irrespective of data set and normalisation method but was able to rescue FDR control for MR normalisation of UMI data, even in a completely asymmetric DE-pattern.

These observations suggest that data sets with a high gene dropout rate might benefit more from imputation than data sets with a relatively low gene dropout rate (Supplementary Figs. 16–18). In order to further investigate the effect of imputation on sparse data, we downsampled the Smart-seq2 and SCRB-seq data, which were originally based on 1 million reads/cell, to make them more comparable to the 10X-HGMM data with on average of 60,000 reads/cell. A radical downsampling to 10% of the original sequencing depth decreases the number of detected genes for SCRB-seq by only 1%, suggesting that the original RNA-seq library was sequenced to saturation. In contrast, the Smart-seq2 data were much less saturated at 1 million reads/cell: Downsampling reduced the number of detected genes by 34%. However, the relative effect of imputation on performance remains small. This is probably due to the fact that the main effect of downsampling is a reduction in the detected genes, which also cannot be imputed. Thus, if a good normalisation method is used to begin with (e.g. scran with clustering), the improvement by imputation remains relatively small.

Good normalisation removes the need for specialised DE-tools. The final step in our pipeline analysis is the detection of DE-genes. Recently, Soneson et al.[31] benchmarked 36 DE approaches and found that edgeR[27], MAST[35], limma-trend[36] and even the T-Test performed well. Moreover, they found that for edgeR, it is important to incorporate an estimate of the dropout rate per cell. Therefore, we combine edgeR here with zingeR[37].

Both edgeR-zingeR and limma-trend in combination with a good normalisation reach similar pAUCs as using the simulated size factors (Fig. 4). However, in the case of edgeR-zingeR this performance is achieved by a higher TPR paid while loosing FDR control (see Supplementary Figs. 19–21), even in the case of symmetric DE-settings (Supplementary Figs. 22–24).

Nevertheless, we find that DE-analysis performance strongly depends on the normalisation method and on the library preparation method. In combination with the simulated size factors or scran normalisation, even a *T*-Test performs well. Conversely, in combination with MR or SCnorm, the *T*-Test has an increased FDR (Supplementary Fig. 20). SCnorms bad performance with a *T*-Test was surprising given SCnorms good performance with limma-trend (Fig. 3b). One explanation could be the relatively large deviation of SCnorm derived size factors (Fig. 3a and Supplementary Fig. 11) which inflate the expression estimates.

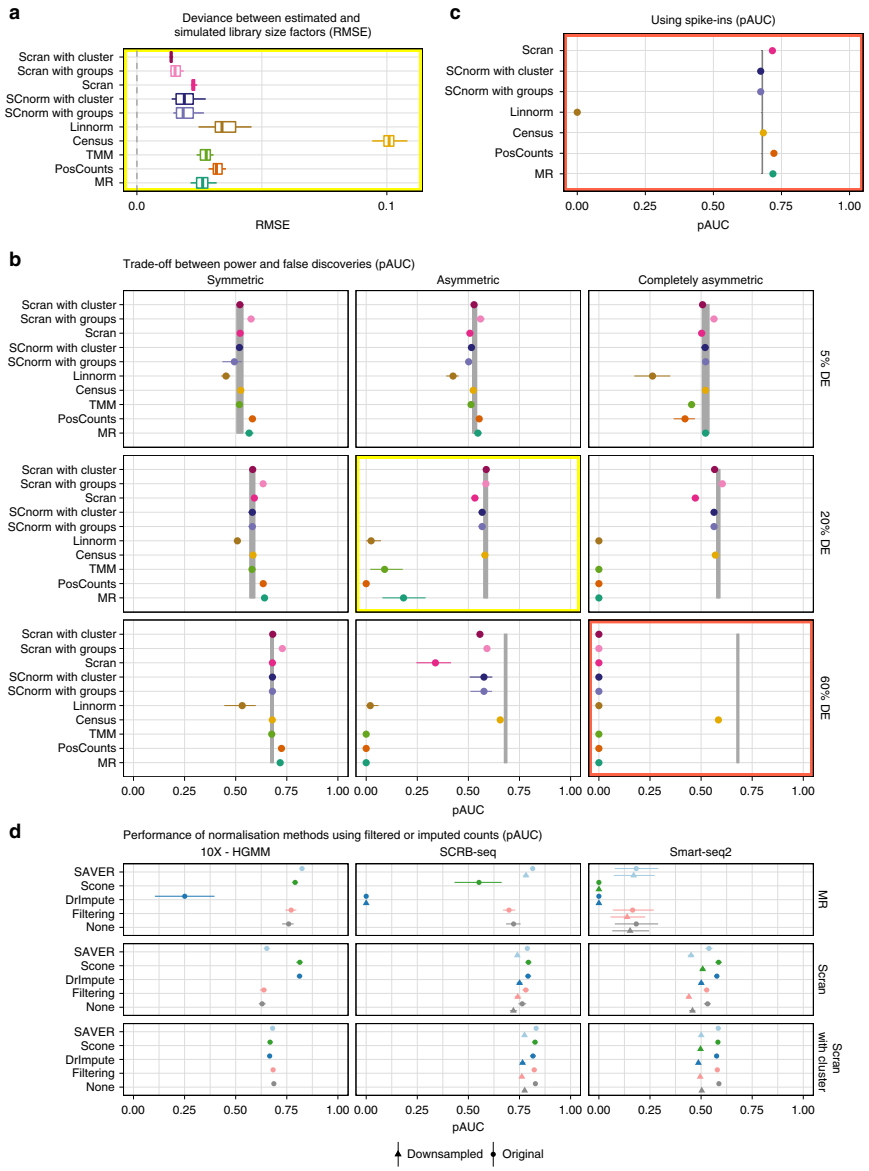
Furthermore, we find that MAST performs consistently worse than the other DE-tools when applied to UMI-based data, but -except in combination with SCnorm- it is doing fine with Smart-seq2 data. Interestingly, Census normalisation in combination with edgeR-zingeR outperformed limma-trend with Smart-seq2 (Supplementary Fig. 25).

In concordance with Soneson et al.[31], we found that limma-trend, a DE-tool developed for bulk RNA-seq data showed the most robust performance. Moreover, library preparation and normalisation appeared to have a stronger effect on pipeline performance than the choice of DE-tool.

Normalisation is overall the most influential step. Because we tested a nearly exhaustive number of ~3000 possible scRNA-seq pipelines, starting with the choice of library preparation protocol and ending with DE-testing, we can estimate the contribution of each separate step to pipeline performance for our different DE-settings (Fig. 1b). We used a beta regression model to explain the variance in pipeline performance with the choices made at the seven pipeline steps (1) library preparation protocol, (2) spike-in usage, (3) alignment method, (4) annotation scheme, (5) preprocessing of counts, (6) normalisation and (7) DE-tool as explanatory variables. We used the difference in pseudo-R^2 between the full model including all seven pipeline steps and leave-one-out reduced models to measure the contribution of each separate step to overall performance.

All pipeline choices together (the full model) explain ~50 and ~60% of the variance in performance, for 20 and 60% DE-genes, respectively (Fig. 5a). Choices of preprocessing the count matrix contribute very little ($\Delta R^2 \leq 1\%$). The same is true for annotation ($\Delta R^2 \leq 2\%$) and aligner choices ($\Delta R^2 \leq 5\%$). For aligner and annotation, it is important to note that these are upper bounds, because our simulations do not include differences in gene detection rates (Fig. 2c).

Surprisingly, the choice of DE-tool only matters for symmetric DE-setups ($\Delta R^2_{\mathrm{DE}=0.2} = 15\%$; $\Delta R^2_{\mathrm{DE}=0.6} = 11\%$), and the choice of library preparation protocol has an even bigger impact on performance for symmetric DE-setups ($\Delta R^2_{\mathrm{Symmetric}} = 17 - 29\%$) and additionally for 5% asymmetric changes ($\Delta R^2_{5\%\mathrm{Asymmetric}} = 17\%$). Normalisation choices have overall a large impact in all DE-settings ($\Delta R^2 = 12$–38%), where the importance increases with increasing levels of DE-genes and increasing asymmetry. Spike-ins are



only necessary if many asymmetric changes are expected and have little or no impact if only 5% of the genes are DE or the changes are symmetric (Fig. 5a). Moreover, for completely asymmetric DE-patterns, the regression model did not converge without normalisation and spike-ins, because their absence or presence alone pushed the MCCs to the extremes.

For the best performing pipeline [SCRB-seq + STAR + GENCODE + SAVER imputation + scran with clustering + limma-trend], using 384 cells per group instead of 96 improves performance only by 6.5–8%. Sample size is more important if a naive pipeline is used. For [SCRB-seq + BWA + GENCODE + no count matrix preprocessing + MR + T-Test] the performance gain by increasing sample size is 10–12% and even worse, for many asymmetric DE-genes, lower sample sizes occasionally appear to perform better (Fig. 5b and Supplementary Fig. 26). Next, we tested our pipeline on publicly available 10× Genomics data set containing the expression profiles of approx. 1000 human peripheral mononuclear blood cells

Fig. 3 Normalisation choices determines DE-analysis performance, not count preprocessing. The data in panels **a**–**c** are based on Smart-seq2 data, all panels are based on two groups of 384 cells, STAR alignment with GENCODE annotation was used for quantification. **a** The root mean squared error (RMSE) of estimated library size factors per normalisation method is plotted for 20% asymmetric DE-genes (see also Supplementary Fig. 11) (Box and whisker plot with centre line = median, bounds of box = 25th and 75th percentile, whiskers = 1.5 * interquartile range from the lower and upper bounds of the box). **b** The discriminatory ability determined by the partial area under the curve (mean pAUC ± s.d.) based on DE testing with limma-trend for normalisation without spike-ins per DE-pattern. The grey ribbon indicates the mean pAUC ± s.d. given simulated size factors (see also Supplementary Figs. 13–15). **c** Using spike-ins for normalisation for 60% completely asymmetric DE-genes. **d** Effect of preprocessing the count matrix for 20% asymmetric DE-genes without spike-ins. Counts were either left asis ('none'), filtered or imputed prior to normalisation. The derived scaling factors were then used for normalisation and DE testing was performed on raw counts using limma-trend (see also Supplementary Figs. 16–18). This procedure was applied to the full count matrix (circle) and to the count matrix downsampled to 10% of its original sequencing depth (triangular). Missing data points are due to failing imputation runs with the sparser data

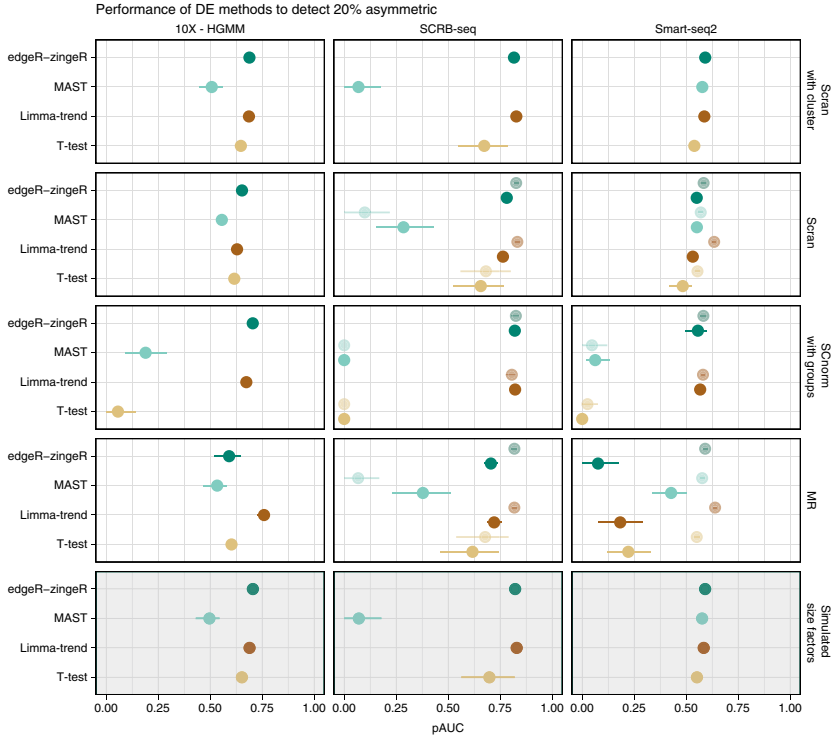
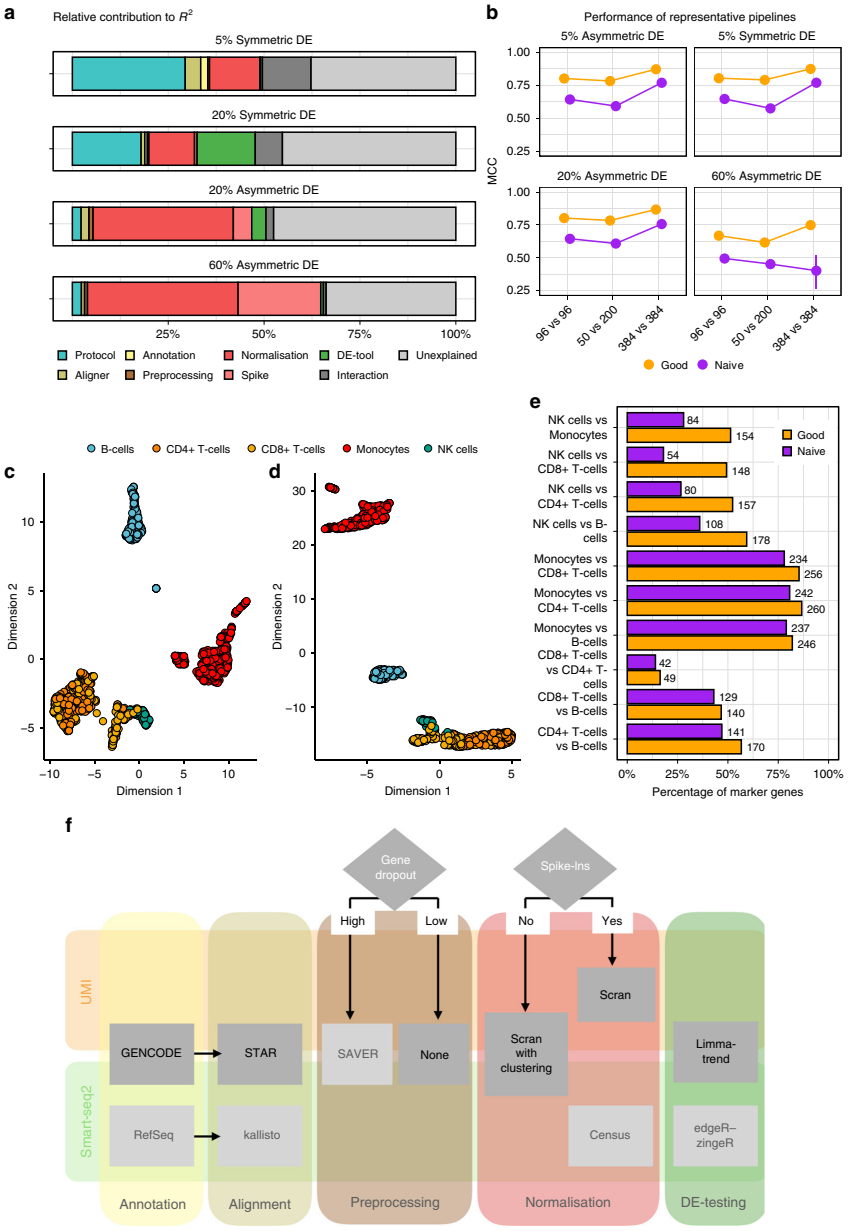


Fig. 4 Evaluation of DE tools. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation per protocol. Twenty percent of the simulated genes are differentially expressed following an asymmetric narrow gamma distribution. Unfiltered counts were normalised using simulated library size factors or applying normalisation methods. Differential expression was tested using *T*-Test, limma-trend, MAST or edgeR-zingeR. The discriminatory ability of DE methods is determined by the partial area under the curve (mean pAUC ± s.d.) for the TPR-FDR curve (see also Supplementary Figs. 19–25)

(PBMC)[16]. First, we classified the cells using SingleR[38] into the celltypes available in the Blueprint Epigenomics Reference[39] distinguishing Monocytes, NK-cells, CD8 + T-cells, CD4 + T-cells and B-cells (Fig. 5c, d). We applied the previously defined good (STAR + gencode + SAVER imputation + scran with clustering + limma-trend) and naive (BWA + gencode + no preprocessing + MR + *T*-Test) pipeline to identify DE-genes between the cell types. Cross-referencing the identified DE-genes with known differences in marker gene expression[39], we find that the good pipeline always identifies a higher fraction of the marker genes as DE than the naive pipeline (Fig. 5e). Comparing NK-cells and CD8 + T-cells, the good pipeline identifies 148 known markers as DE, while the naive pipeline finds only 54. The diminished separation between those two cell-types using the naive pipeline is already visible in the UMAP (Fig. 5d).



In summary, we identify normalisation and library preparation as the most influential choices and the observation that differences in computational steps alone can significantly lower the required sample size nicely illustrates the importance of bioinformatic choices.

Discussion

Here we evaluate the performance of complete computational pipelines for the analysis of scRNA-seq data under realistic conditions with large numbers of DE-genes and differences in total mRNA contents between groups (Fig. 1). Furthermore, our

Fig. 5 Evaluation of analysis pipeline. **a**, **b** The expression of 10000 genes over 768 cells were simulated and 5, 20 or 60% of the genes were differentially expressed following a symmetric or asymmetric narrow gamma distribution. This simulation setup was applied to protocols, alignments, annotations, preprocessing of counts, normalisation and DE tools. For each analysis set, the Matthew Correlation Coefficient (mean MCC ± s.d.) was averaged over 20 simulations and rescaled to [0, 1] interval. The MCC was used as a response variable in beta regression models with log-log link function. **a** The contribution of each covariate in the full model (~Protocol + Aligner + Annotation + Preprocessing + Normalisation + DE-Tool). **b** Performance according to sample size, 1 good and 1 naive pipeline (see also Supplementary Fig. 26). **c**–**e** The expression of ~1000 human PBMcs profiled with 10× Genomics were processed using the good and naive pipeline. Cell types were identified with SingleR classification using the Blueprint Epigenomics Reference. Cell types are represented in a UMAP, for good **c** and naive **d** pipeline, respectively. True marker genes, i.e., given by the reference, per pairwise comparison of cell types for the good and naive pipeline are given in **e** where genes needed to have a adjusted p-value < 0.1, absolute log2 fold change threshold (>0.1) and expressed in at least 10% of the cells to be considered. **f** Pipeline recommendations for UMI and Smart-seq2 data

simulations allow us not only to investigate the influence of choices made at each pipeline step separately, but also to estimate the relative importance and interactions between different steps of an entire scRNA-seq analysis pipeline. We implemented all assessed computational methods and more in powsimR, so that users can easily evaluate pipeline performance given their own data and expected DE-settings.

Beginning with the creation of the raw count matrix, we find that transcriptome mapping with BWA[19] appears to recover the largest number of genes. However, many of these are probably due to falsely mapped reads, also increase expression variance which ultimately results in a lower sensitivity (Fig. 2c–f). In contrast, the pseudo-alignment method kallisto[24] appears to assign reads precisely, but looses a lot of reads leading to a lower mean expression. Finally, a genome mapping approach using the splice-aware aligner STAR[18] in conjunction with GENCODE annotation recovers the most reads with the highest accuracy (Fig. 5f).

Concerning the preprocessing of the count matrix, we found in concordance with Andrews et al.[40] that in particular for sparse data such as 10X, SAVER[34] imputation before normalisation improves performance, while filtering genes has no effect with our data sets and combinations of normalisation and DE-testing methods.

The choice that had the largest impact on performance throughout all tested DE-settings is the choice of normalisation method. Only for symmetric changes, the choice of library preparation protocol had a slightly larger impact than normalisation. In line with Evans et al. (2018)[11], we found that normalisation performance of bulk methods and also some of the single cell methods declined with asymmetry (Fig. 3b). In particular, for 60% completely asymmetric DE-genes only Census retained FDR control. Unfortunately, Census is not recommended for the use with UMI-counts. Thus, for UMI-counts and 60% completely asymmetric changes, only the use of spike-ins could restore test performance. In the debate about the usefulness of spike-ins[17,41], we land on the pro side: Our simulations clearly show that spike-ins are useful in DE-testing settings with asymmetric changes which is likely to be a common phenomenon in scRNA-seq data. Due to good performance across DE-settings and its speed (Supplementary Figs. 22 and 27) we would recommend scran with prior clustering as the best choice for normalisation (Fig. 5f).

The choice in DE-testing method, our final pipeline step had relatively little impact on overall pipeline performance. A good normalisation prior to DE-testing alleviates the need for more complex and thus vulnerable methods, such as for example MASTs hurdle model which implicitly assumes that the CPM values were generated from zero inflated negative binomial count distribution. Indeed, we previously showed that also scRNA-seq data fit a negative binomial distribution rather well and that the previously reported zero-inflation in scRNA-seq data is mainly due to amplification noise which is removed in UMI-data[10]. Hence, it is not surprising that in concordance with Soneson et al.[31], we find that relatively straight forward DE-testing methods adapted from bulk RNA-seq perform well with scRNA-seq data.

Finally, we want to remark that paying attention to the details in a computational pipeline and in particular to normalisation pays off. The effect of using a good pipeline as compared to a naively compiled one has a similar or even greater effect on the potential to detect a biological signal in scRNA-seq data as an increase in cell numbers from 96 to 384 cells per group (Fig. 5b).

Methods

Single cell RNA-seq data sets. The starting point for our comprehensive pipeline comparison is the scRNA-seq library preparation (Fig. 1a). In our comparison, we included the gene expression profiles of mouse embryonic stem cells (mESC) as published in Ziegenhain et al.[2] (Supplementary Fig. 1). We selected four data sets for our comparison: Smart-seq2[13] a well-based full-length scRNA-seq protocol, CEL-seq2[15] a well-based 3′ UMI-protocol using linear amplification, SCRB-seq a well-based 3′ UMI-protocol with PCR amplification[2,42] and Drop-seq[14] a droplet-based 3′ UMI-protocol. In addition, 92 poly-adenylated synthetic RNA transcripts of known concentration designed by the External RNA Control Consortium (ERCCs)[43] were spiked in for all methods except Drop-seq. All raw cDNA sequencing reads were cut to a length of 45 bases and downsampled to one million cDNA reads per cell (Supplementary Table 1 and Supplementary Fig. 1).

Finally, we added a 10X Chromium data set sequencing mouse NIH3T3 cells[16], yielding ~400 good cells with on average ~60,000 reads/cell and another 10X data set analysing ~1000 human peripheral blood mononuclear cells (PBMCs).

These choices of library preparation protocols cover the diversity in current protocols without imposing partiality due to biological differences and technical handling.

Gene expression quantification. For genome mapping and quantification of the UMI-data with a splice-aware aligner, we used the zUMIs[44] (v.0.0.3) pipeline with STAR[18] (v.2.5.3a) and the mouse genome (Mus_musculus.GRm38) together with annotation files (gtf) for GENCODE (vM15), Vega (VEGA68) and RefSeq (Release 85) (Supplementary Table 2). zUMIs is a fast and flexible pipeline for processing scRNA-seq data where cell barcode or UMI reads with low sequence quality reads are filtered out prior to UMI collapsing by sequence identity which yields identical count results as UMI-tools[44,45]. For Smart-Seq2 we use the same pipeline settings as in zUMIs, simply omitting the UMI collapsing step (Supplementary Table 3).

For transcriptome alignment, we downloaded transcriptome fasta files corresponding to the annotations listed above. We used BWA[19] (v0.7.12) to align the scRNA-seq reads to these transcriptomes. We only removed reads that aligned equally well to transcripts of different genes as truly multi-mapped. The remaining reads were tallied up per gene. For UMI data, the reads were collapsed per gene by identity, similar to the strategy recommended in zUMIs.

For kallisto[24] (v0.43.1), a transcriptome-guided pseudo-alignment method, we followed the recommended quantification procedure for scRNA-seq data to yield abundance estimates per equivalence class. To be comparable with other alignment methods, the counts per equivalence class were collapsed per gene. The counts of equivalence classes representing multiple genes were filtered out. For SCRB-seq, CEL-seq2, Drop-seq and 10× Genomics libraries, we chose the UMI-aware quantification option. The ERCC spike-in sequences were appended to the genome or transcriptome sequences for quantification.

Simulations. We used powsimR to estimate, simulate and evaluate single cell RNA-seq experiments[10]. PowsimR has been independently validated for benchmarking DE-approaches[31] and consistently reproduces the mean-variance relationship and dropout rates of genes of scRNA-seq data (see also Supplementary Fig. 28). The gene expression quantification using three different aligners in combination with three annotations per library preparation protocol produced 45 count matrices. These count matrices are the basis for our estimation in powsimR.

Genes needed at least one read or UMI count in at least one cell to be considered in the estimation for simulation parameters. Since we[10] and others[46,47] have found previously, we assume that UMI counts follow a negative binomial distribution and only Smart-seq2 needs the inclusion of zero-inflation. To simulate spike-in data, we added an implementation of the simulation framework for pure technical variation of spike-ins described in Kim et al.[48] to powsimR. The parameters required for these simulations were estimated from 92 ERCC spike-ins in the SCRB-seq, CEL-seq2 and Smart-seq2 data, respectively[2]. To evaluate the effect of differing sequencing depths, we added a new module to powsimR that estimates the degree of PCR amplification for UMI data. This allows the user to downsample a read count matrix by binomial thinning as implemented in edgeR thinCounts()[27] and then to reconstruct the corresponding UMI count matrix base on the estimated PCR amplification rates.

For a detailed evaluation of the pipelines, we simulated two groups of cells for pairwise comparisons with the following three sample size setups: 96 vs. 96, 384 vs. 384 or 50 vs. 200 cells (Fig. 1b). For simplicity, we kept the number of genes that we simulated constant at 10,000. To introduce slight variation in expression capture, we draw a different size factor for each cell from a narrow normal distribution ($X \sim N(\mu = 1, \sigma = 0.1)$) (Fig. 1b). This distribution fits the considered data sets well, irrespective of the applied library preparation method. Furthermore, the two groups of cells can have 5, 20 or 60% differentially expressed genes. To capture the asymmetry of observed expression differences, we considered three setups of DE-patterns: symmetric (50% up- and 50% downregulated), asymmetric (75% up- and 25% downregulated) or completely asymmetric (100% upregulated). The magnitude of expression change is drawn from a narrow gamma distribution ($X \sim \Gamma(\alpha = 1, \beta = 2)$) defining the log2 fold change, which is then added to the sampled mean expression. The combination of these parameters results in a total of 27 DE-setups that were then applied to the parameter estimates from 37 different count matrices to simulate 20 replicates for each setting, producing a total of 19,980 simulated data sets.

These data sets were then analysed by a nearly exhaustive number of combinations of four imputation strategies (scone, SAVER, DrImpute), gene dropout filtering (remove genes with more than 80% zero expression values) together with seven normalisation approaches (TMM, MR, Linnorm, Census, Linnorm, scran, SCnorm) with or without spike-ins, depending on library preparation protocol and method (Fig. 1c). Normalisation factors were then derived as described in Soneson et al.[31] and used in conjunction with the raw count matrices for DE-testing using four representative approaches (*T*-Test, limma-trend, edgeR-zingeR, MAST). The resulting p-values were corrected for multiple testing with Benjamini-Hochberg FDR and we applied a threshold level of 10% to define positive test results. All these steps were seamlessly implemented into powsimR (github: https://github.com/bvieth/powsimR). In total we analysed 2,979 different RNA-seq pipelines.

Evaluation metrics. To evaluate the normalisation results, we determined the root mean squared error (RMSE) of a robust linear model using the difference between estimated and simulated library size factors as response variable in rlm() implemented in R-package MASS[49] (v.7.3–51.1) (Supplementary Fig. 10)[9].

All other measures are based on the final results of an entire scRNA-seq pipeline ending with DE-testing. Knowing the identity of the genes that were simulated to show differing expression levels and the results of the DE-testing, we used a number of metrics related to the confusion matrix tabulating the number of true positives, false positives, true negatives and false negatives. We define the power to detect DE with the TPR ($\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$). The false discovery rate is defined as $\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP+TP}}$. We combine these two measures in a TPR versus FDR curve to quantify the trade-off between true and false discoveries in a genome-wide multiple testing setup as advocated by[50]. We then summarise these curves by their partial area under curve (pAUC) of TPR versus observed FDR that still ensures FDR control at the nominal level of 10% (Supplementary Fig. 11). This way of calculating the AUC is ideal for data with relatively high true negative rates as the partial integration does not punish methods that are over-conservative, i.e. that stay way below the nominal FDR.

To summarise the whole confusion matrix in one representative value we chose the Matthews Correlation Coefficient ($\mathrm{MCC} = \frac{\mathrm{TP*TN} - \mathrm{FP*FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$), because it is a balanced measure ensuring a reliable comparison of method performance across all DE-settings[50,51]. As for the pAUC, we calculated the maximal value of MCC where the cutoff still ensured FDR control at the nominal level of 10%.

To quantify the relative contribution of each step in the analysis pipeline, we used the MCC as a response variable in a beta regression model implemented in R-package betareg (v.3.1–1)[52] with each individual pipeline step. Because the MCC assumes the extremes of 0 and 1 in some DE-settings, we applied the recommended transformation, namely $\mathrm{MCC}_{\mathrm{transformed}} = \frac{\mathrm{MCC}*(n-1)+0.5}{n}$, where n is the sample size[53]. The contribution is then given by the difference between the full model pseudo-R^2 containing all covariates versus a model leaving one step out at a time. This is then scaled to the total variance explained to give relative ΔR^2 percentages.

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

Data availability

Any relevant data are available from the authors upon reasonable request. The scRNA-seq data used in this manuscript are all publicly available, and they are summarised in Supplementary Table 1. The SCRB-seq, Smart-seq2, Drop-seq, CEL-seq2 data are available at the Gene Expression Omnibus (GEO) under accession code GSE75790. The HGMM and PBMC data sets are available at 10x Genomics's official website (https://support.10xgenomics.com/single-cell-gene-expression/datasets). The data produced by the analysis in this manuscript is freely available from the following zenodo data repository (https://doi.org/10.5281/zenodo.3364466).

Code availability

The software and code used are summarised in Supplementary Tables 3 and 4. A compendium containing processing scripts and detailed instructions to reproduce the analysis for this manuscript is freely available from the following GitHub repository (https://github.com/bvieth/scRNA-seq-pipelines).

Received: 27 March 2019; Accepted: 28 August 2019;
Published online: 11 October 2019

References

1. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
2. Ziegenhain, C. et al. Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell* **65**, 631–643 (2017).
3. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
4. Baruzzo, G. et al. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods* **14**, 135–139 (2017).
5. Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M. & Wilke, C. O. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics* **19**, 510–524 (2018).
6. Zhao, S. & Zhang, B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics* **16**, 97 (2015).
7. Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. *F1000Res.* **7**, 1740–1776 (2018).
8. Zhang, L. & Zhang, S. Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* https://doi.org/10.1109/TCBB.2018.2848633 (2018).
9. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* **14**, 565–571 (2017).
10. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimr: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33**, 3486–3488 (2017).
11. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* **19**, 776–792 (2018).
12. Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
13. Picelli, S. et al. Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
14. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
15. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
16. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049–14051 (2017).
17. Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I. & Enard, W. Quantitative single-cell transcriptomics. *Brief. Funct. Genomics* **17**, 220–232 (2018).
18. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
20. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. kallisto. https://github.com/pachterlab/kallisto/tree/v0.43.1 (2017).
21. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
22. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
23. Wilming, L. G. et al. The vertebrate genome annotation (vega) database. *Nucleic Acids Res.* **36**, D753–D760 (2008).

24. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
25. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75–89 (2016).
26. Bacher, R. et al. SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* **14**, 584–586 (2017).
27. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25–R34 (2010).
28. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106–R118 (2010).
29. Yip, S. H., Wang, P., Kocher, J.-P. A., Sham, P. C. & Wang, J. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.* **45**, e179–e191 (2017).
30. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with census. *Nat. Methods* **14**, 309–315 (2017).
31. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
32. Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. & Garry, D. J. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinforma.* **19**, 220–230 (2018).
33. Cole, M. B. et al. Performance assessment and selection of normalization procedures for Single-Cell RNA-Seq. *Cell Syst.* **8**, 315–328 (2019).
34. Huang, M. et al. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**, 539–542 (2018).
35. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 1–13 (2015).
36. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29–R46 (2014).
37. Van den Berge, K. et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* **19**, 24–41 (2018).
38. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
39. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The international human epigenome consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**, 1145–1149 (2016).
40. Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. *Mol. Asp. Med.* **59**, 114–122 (2018).
41. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinforma.* **12**, 480–497 (2011).
42. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. Preprint at https://doi.org/10.1101/003236v1 (2014).
43. Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
44. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs - a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, 1–9 (2018).
45. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
46. Amrhein, L., Harsha, K. &Fuchs, C. A mechanistic model for the negative binomial distribution of single-cell mRNA counts (2019).
47. Svensson, V. Droplet scRNA-seq is not zero-inflated (2019).
48. Kim, J. K., Kolodziejczyk, A. A., Illicic, T., Teichmann, S. A. & Marioni, J. C. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **6**, 8687–8695 (2015).
49. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S.* fourth edn (Springer, New York, 2002. http://www.stats.ox.ac.uk/pub/MASS4
50. Soneson, C. & Robinson, M. D. iCOBRA: open, reproducible, standardized and live method benchmarking. *Nat. Methods* **13**, 283 (2016).
51. Boughorbel, S., Jarray, F. & El-Anbari, M. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PLoS ONE* **12**, e0177678 (2017).
52. Cribari-Neto, F. & Zeileis, A. Beta regression in R. *J. Stat. Softw.* **34**, 1–24 (2010).
53. Smithson, M. & Verkuilen, J. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods* **11**, 54–71 (2006).

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through LMUexcellent, SFB1243 (Subproject A14/A15) and DFG grant HE 7669/1-1. C.Z. is recipient of an EMBO long-term fellowship (ALTF 673-2017).

Author contributions

B.V. and I.H. conceived the study. B.V. prepared and analysed the scRNA-seq data. B.V. implemented and conducted the simulation and evaluation framework. S.P. and C.Z. helped in data processing and power simulations. W.E. and I.H. supervised the work and provided guidance in data analysis. B.V., I.H. and W.E. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-12266-7.

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

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

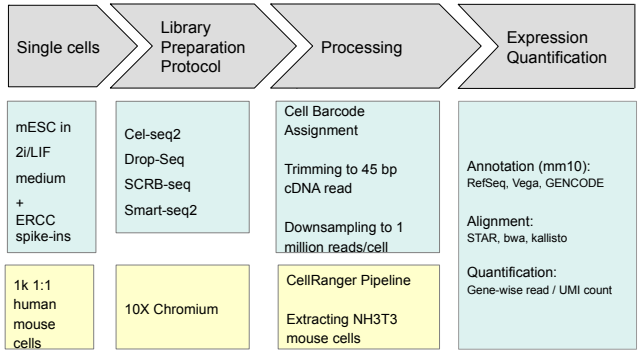
Supplementary Information

A Systematic Evaluation of Single Cell RNA-Seq Analysis Pipelines

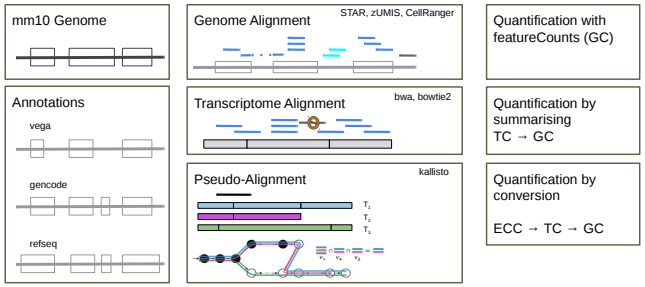
by

Vieth et al.

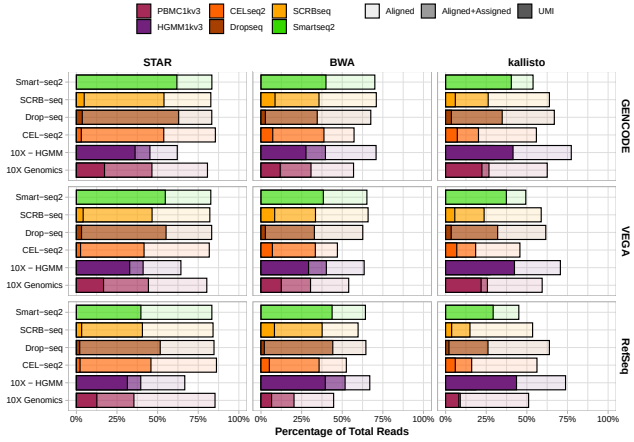
Supplementary Figures



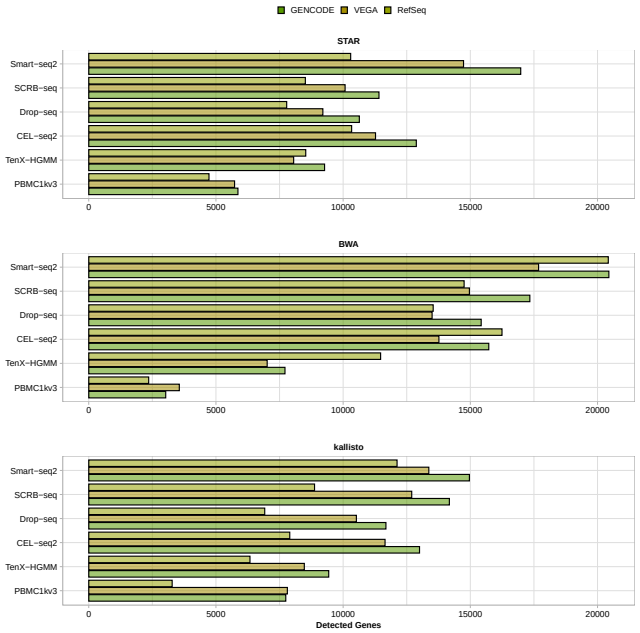
Supplementary Figure 1: Schematic overview of scRNA-seq data sets. In our comparison, we included the gene expression profiles of mouse embryonic stem cells (mESC) as published in [1]. Briefly, the mESC were cultured under two inhibitor/leukemia inhibitory factor (2i /LIF) conditions to ensure rather homogeneous cell populations [2]. We selected four scRNA-seq methods (Smart-seq2, SCRB-seq, Drop-seq, CEL-seq2) that were used to construct libraries in two independent replicate batches. In addition, 92 poly-adenylated synthetic RNA transcripts of known concentration designed by the External RNA Control Consortium (ERCCs) [3] were spiked in for all methods except Drop-seq. All raw sequencing reads were cut and downsampled to 45 base long one million cDNA reads per cell. We included two commonly used expression quantification approaches, namely reference-guided alignment in STAR and bwa as well as pseudoalignment in kallisto in combination with three annotations. Furthermore, we downloaded a scRNA-seq data set from 10X Genomics Support, namely the 1k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NH3T3) Cells [4] generated using the v2 gene expression chemistry. We proceeded with approx. 400 mouse cells with 70000 reads/cell on average.



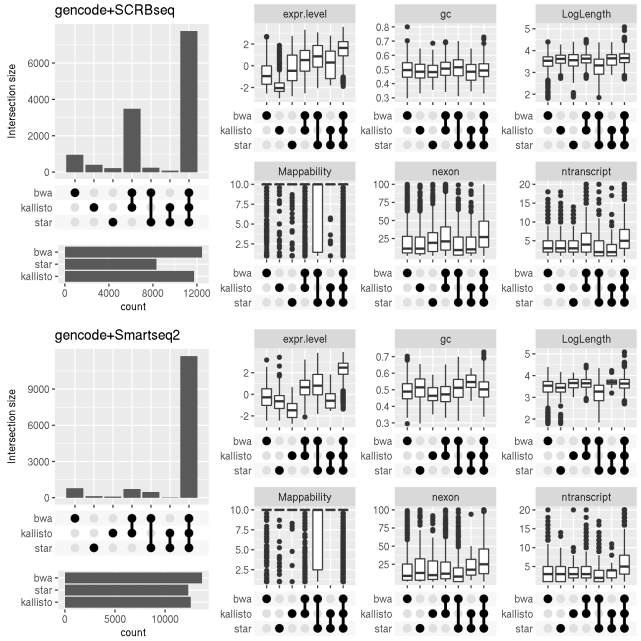
Supplementary Figure 2: Schematic overview of alignment and annotation. **Left** The annotation RefSeq, Vega and GENCODE for the mm10 genome **Middle** Alignment of reads to the genome using STAR allowing alignment to genic and intergenic sequences; Alignment of reads to the transcriptome using BWA; Pseudo-alignment of reads to de Bruijn Graph representation of the transcriptome using kallisto (figure adapted from[5]). **Right** STAR: Estimation of expression as read/UMI counts per gene (GC) using featureCounts for genome alignments; BWA: Estimation of expression as read/UMI counts per gene by summarising counts over transcripts (TC) belonging to one gene; kallisto: Estimation of expression given as equivalence class counts (ECC) converted to trancript counts (TC) and summarised to gene counts (GC).



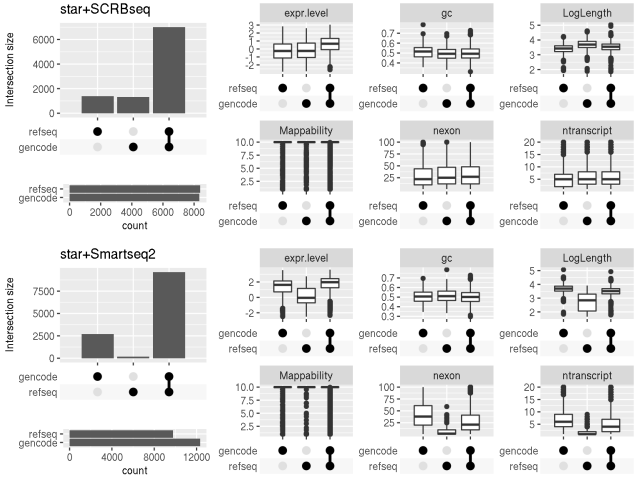
Supplementary Figure 3: Read Alignment and Assignment Rates per Library Preparation Protocol stratified over Aligner and Annotation. The lighter shade represents the percentage of the total reads that could be aligned and the darker shade the percentage that also was uniquely assigned.



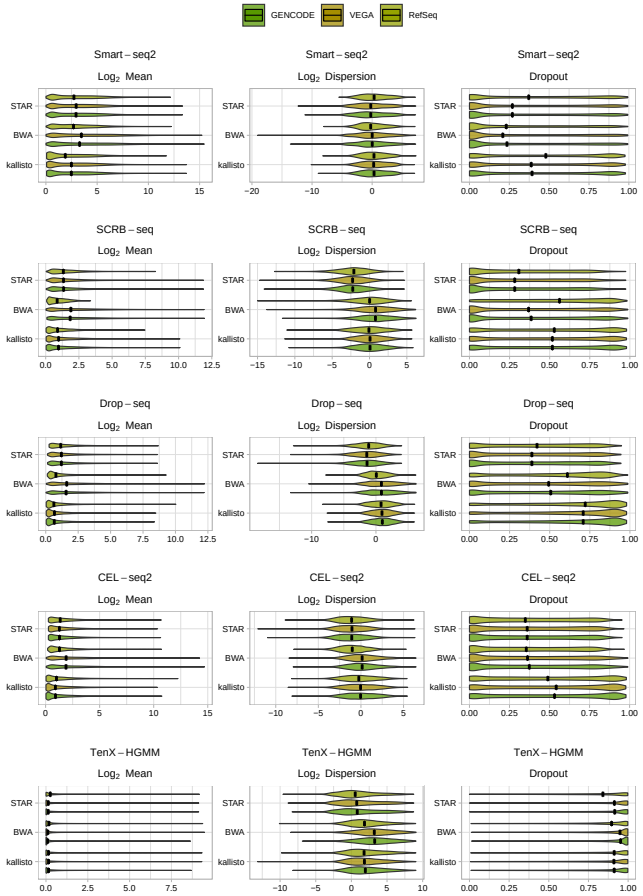
Supplementary Figure 4: Gene Detection Rates. Number of genes detected per Library Preparation Protocol stratified over Aligner and Annotation (i.e. at least 10% nonzero expression values).



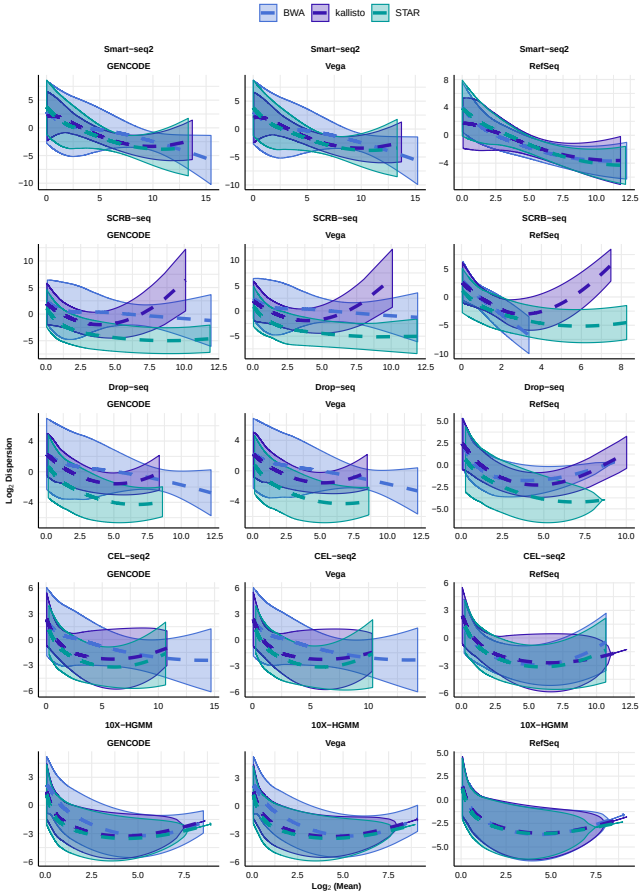
Supplementary Figure 5: Properties of genes detected by the different mapping strategies. Mappability is represented as 10^{M25}, where M25 is the lower quartile of the mappability scores[6] across the gene. All other gene properties were extracted from the corresponding annotation file. Genes detected by all three mappers tend to have higher expression levels. The only other consistent pattern is that genes only detected by kallisto have on average a lower expression and genes that escape detection by kallisto have slightly lower mappability. Box and whisker plot with centre line = median, bounds of box = 25th and 75th percentile, whiskers = 1.5 * interquartile range from the lower and upper bounds of the box.



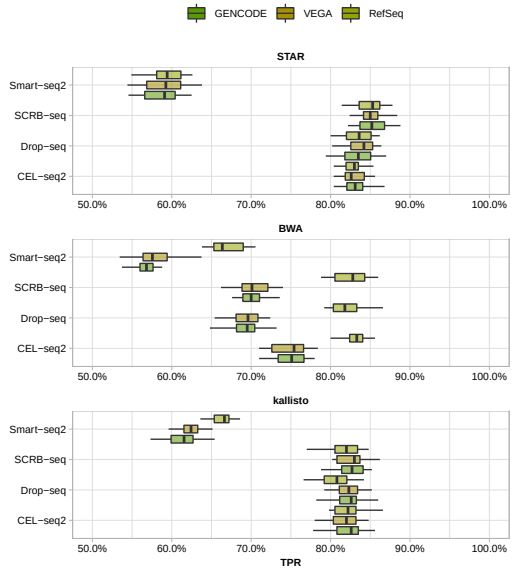
Supplementary Figure 6: Differences in genes detected with RefSeq and Gencode annotation. We used the matching between RefSeq and Gencode transcript annotations that is provided by Gencode and then summarise detection at the gene level. Comparing the genes found using Gencode vs. RefSeq annotation, we find that for the 3-prime method SCRB-seq both annotations yield approximately the same number of genes, but ∼1,000 appear annotation-specific. The Gencode annotation is more comprehensive, in that it contains more and often longer transcripts and indeed the Gencode-specific genes are longer and thus 3 mapping reads are easily lost. The RefSeq specific transcripts are more puzzling and the only distinguishing feature that we see is that they appear more GC-rich. For full-length data generated with Smart-seq2, Gencode detects 2,500 more genes than RefSeq and there are almost no RefSeq specific genes. Genes detected with Gencode only are longer and have on average more exons and transcripts. Box and whisker plot with centre line = median, bounds of box = 25th and 75th percentile, whiskers = 1.5 * interquartile range from the lower and upper bounds of the box.



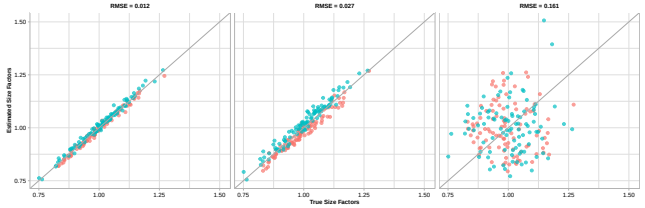
Supplementary Figure 7: Distribution Estimates. Estimated mean expression, dispersion and dropout rates per Library Preparation Protocol stratified over Aligner and Annotation. Black line indicates median value.



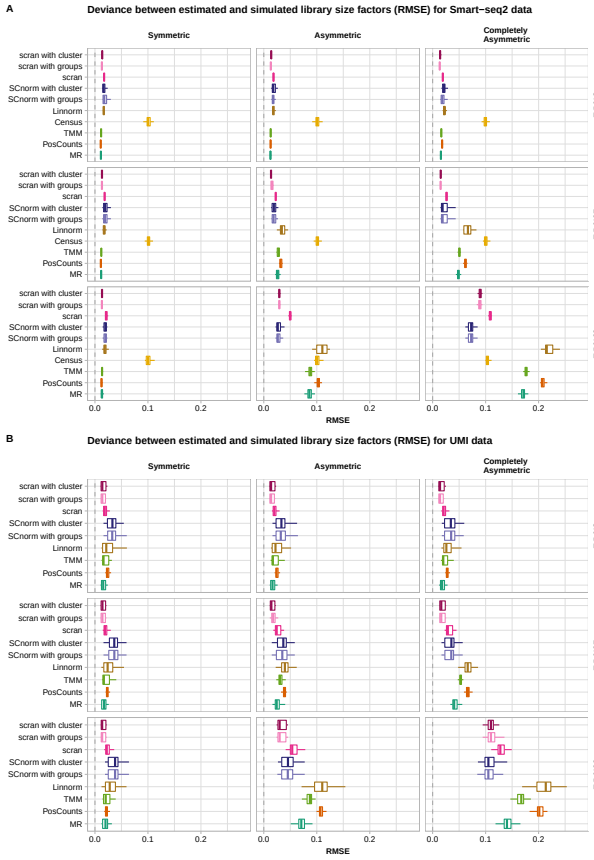
Supplementary Figure 8: Distributional Fittings for simulations. Mean-Dispersion Fitting Line applying a cubic smoothing spline with 95% variability bands per Library Preparation Protocol stratified over Aligner and Annotation.



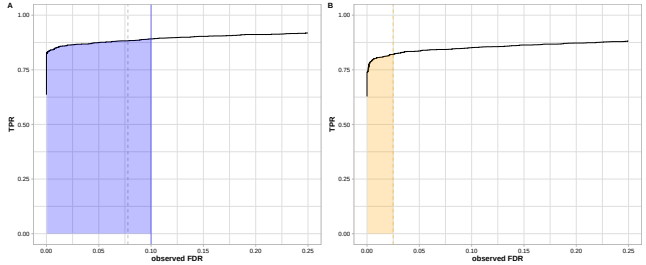
Supplementary Figure 9: The effect of quantification choices on detecting differential expression. The expression of 10,000 genes over 768 cells (384 cells per group) was simulated and 5% of the genes were differentially expressed following an asymmetric narrow gamma distribution. Any gene correctly called differentially expressed at FDR 10% contributed to the True Positive Rate (TPR). The TPR per library preparation method stratified over aligner and annotation is plotted. Box and whisker plot with centre line = median, bounds of box = 25th and 75th percentile, whiskers = 1.5 * interquartile range from the lower and upper bounds of the box.



Supplementary Figure 10: Illustration for RMSE evaluation of library size factors. The expression of 10,000 genes over 768 cells (384 cells per group (red and cyan points) were simulated and 20% of the genes were differentially expressed following an asymmetric narrow gamma distribution. To compare the estimated library size factors with the simulated library size factors, the factors were centred and scaled. The root mean squared error (RMSE) of a robust linear regression represents the deviation between estimated and simulated size factors.



Supplementary Figure 11: Deviation between simulated and estimated library size factors. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated and 5%, 20% or 60% of the genes were differentially expressed following a symmetric, an asymmetric or a completely asymmetric narrow gamma distribution. The root mean squared error (RMSE) of estimated scaling factors per normalisation method is plotted.Box and whisker plot with centre line = median, bounds of box = 25th and 75th percentile, whiskers = 1.5 * interquartile range from the lower and upper bounds of the box. **A** Smart-seq2 data **B** UMI data.



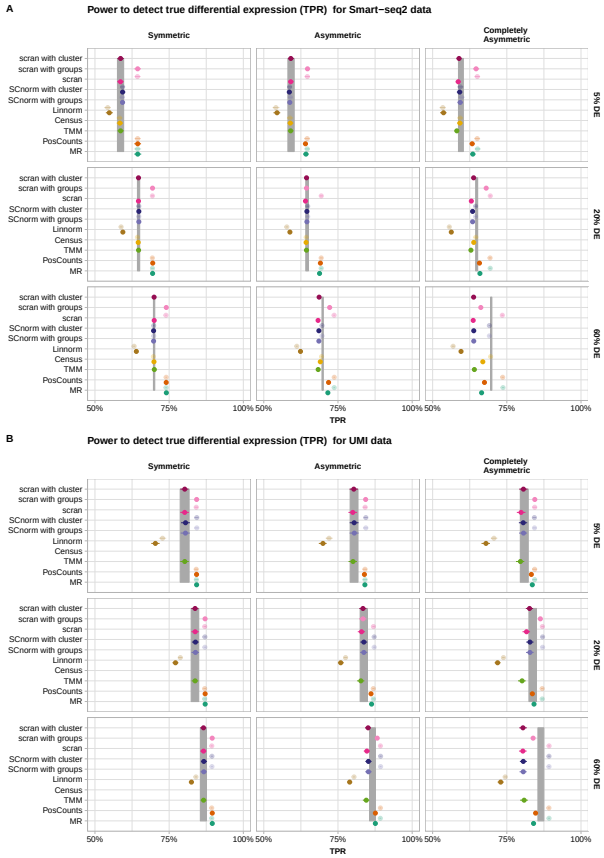
C

TP	FP	TN	FN	TPR	observed FDR	threshold FDR
1687	12	7988	313	0.84	0.0071	0.01
1716	23	7977	284	0.86	0.0132	0.02
1729	45	7955	271	0.86	0.0254	0.03
1735	60	7940	265	0.87	0.0334	0.04
1739	77	7923	261	0.87	0.0424	0.05
1747	83	7917	253	0.87	0.0454	0.06
1752	98	7902	248	0.88	0.053	0.07
1755	114	7886	245	0.88	0.061	0.08
1762	126	7874	238	0.88	0.0667	0.09
1765	149	7851	235	0.88	0.0778	0.1
1771	170	7830	229	0.89	0.0876	0.11
1779	190	7810	221	0.89	0.0965	0.12
1783	199	7801	217	0.89	0.1004	0.13
1788	214	7786	212	0.89	0.1069	0.14
1790	226	7774	210	0.9	0.1121	0.15

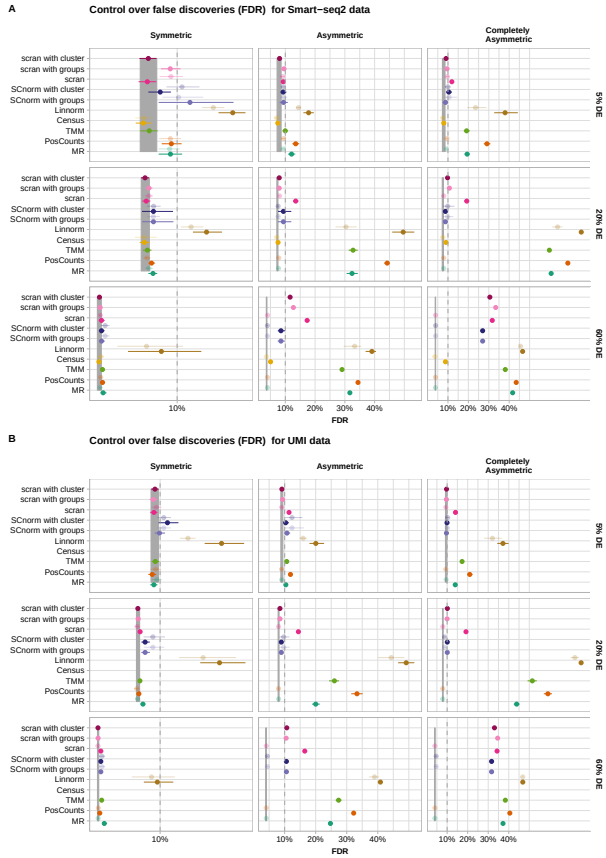
D

TP	FP	TN	FN	TPR	observed FDR	threshold FDR
1678	93	7907	322	0.84	0.0525	0.01
1688	148	7852	312	0.84	0.0806	0.02
1705	198	7802	295	0.85	0.104	0.03
1712	246	7754	288	0.86	0.1256	0.04
1726	296	7704	274	0.86	0.1464	0.05
1729	332	7668	271	0.86	0.1611	0.06
1735	371	7629	265	0.87	0.1762	0.07
1742	404	7596	258	0.87	0.1883	0.08
1745	437	7563	255	0.87	0.2003	0.09
1748	473	7527	252	0.87	0.213	0.1
1756	522	7478	244	0.88	0.2291	0.11
1758	562	7438	242	0.88	0.2422	0.12
1764	597	7403	236	0.88	0.2529	0.13
1767	636	7364	233	0.88	0.2647	0.14
1773	662	7338	227	0.89	0.2719	0.15

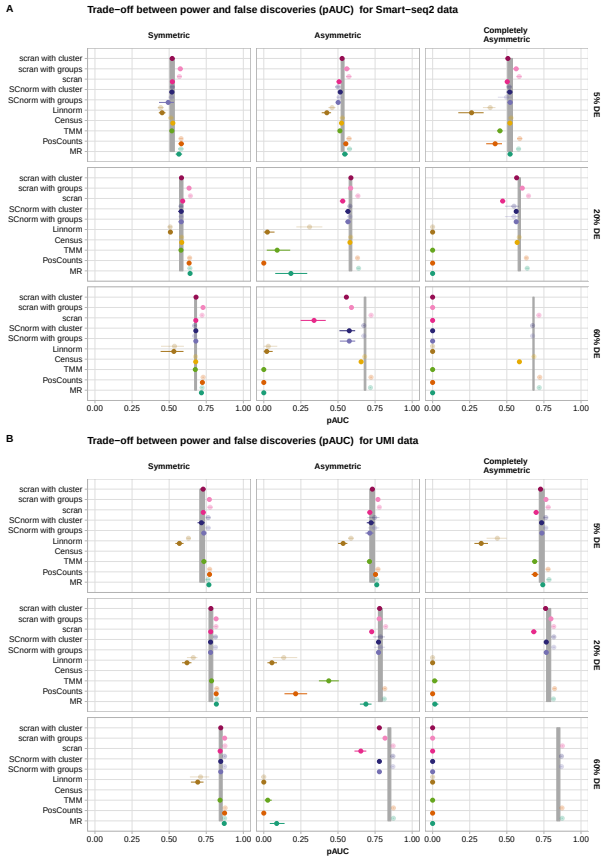
Supplementary Figure 12: Illustration for pAUC calculation used as an evaluation metric for performance. The expression of 10,000 genes estimated from SCRB-seq data over 768 cells (384 cells per group were simulated and 20% of the genes were differentially expressed following an asymmetric narrow gamma distribution. The power to detect DE (TPR) versus observed FDR based on DE-testing with limma-trend using scran with clustering (**A**) or Median-Ratio (**B**) is plotted[7]. The dashed line indicates the level at the which the observed stays below the nominal level of 10% which defines the right side boundary for the partial area under this curve[8]. The corresponding proportion and rates for a selection of nominal FDR thresholds are given in **C, D**. Since we do not want to punish conservative FDR control of methods, the test results using scran normalisation for example is extended to observed FDR of 10% whereas median ratio only ensures FDR control at a lower observed FDR.



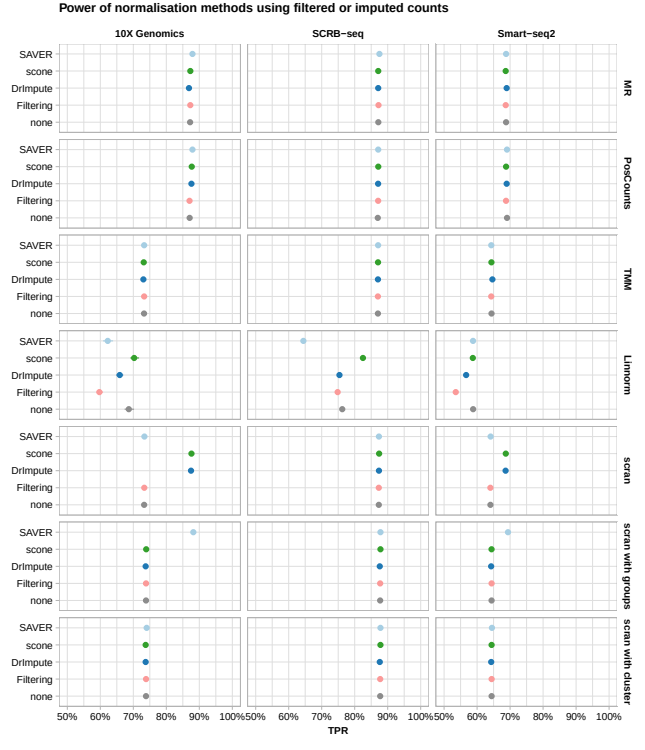
Supplementary Figure 13: Power to detect true differential expression per normalisation method. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated and 5%, 20% or 60% of the genes were differentially expressed following a symmetric, an asymmetric or a completely asymmetric narrow gamma distribution. The power to detect DE (mean TPR ± s.d.) based on DE-testing with limma-trend per normalisation method is plotted. The lighter shade indicates the usage of spike-ins for normalisation. The grey ribbon indicates the TPR given simulated size factors (mean TPR ± s.d.).
A Smart-seq2 data **B** UMI data.



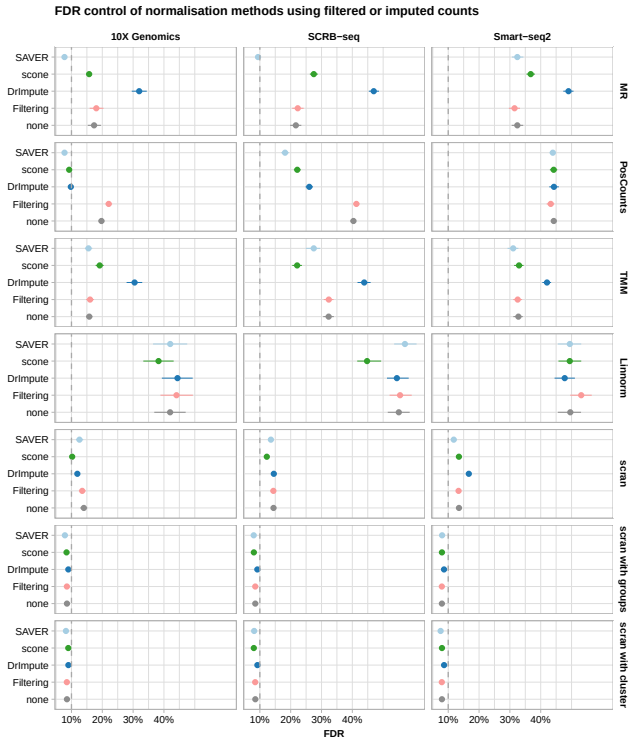
Supplementary Figure 14: Control over false discoveries per normalisation method. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated and 5%, 20% or 60% of the genes were differentially expressed following a symmetric, an asymmetric or a completely asymmetric narrow gamma distribution. The BH-FDR control based on DE-testing with limma-trend per normalisation method is plotted (mean FDR ± s.d.). The lighter shade indicates the usage of spike-ins for normalisation. The dashed line indicates the nominal FDR level of 10%. The grey ribbon indicates the FDR given simulated size factors (mean FDR ± s.d.).
A Smart-seq2 data **B** UMI data.



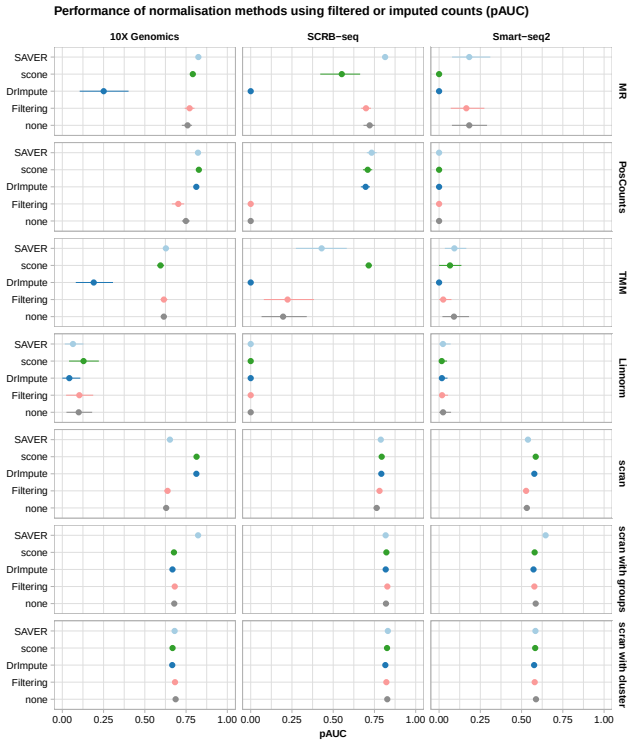
Supplementary Figure 15: Trade-off between power and false discoveries per normalisation method. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated and 5%, 20% or 60% of the genes were differentially expressed following a symmetric, an asymmetric or a completely asymmetric narrow gamma distribution. The discriminatory ability determined by the partial area under the curve (pAUC) based on DE-testing with limma-trend for normalisation per DE-setup is plotted (mean pAUC ± s.d.). The lighter shade indicates the usage of spike-ins for normalisation. The grey ribbon indicates the pAUC given simulated size factors (mean pAUC ± s.d.).
A Smart-seq2 data **B** UMI data.



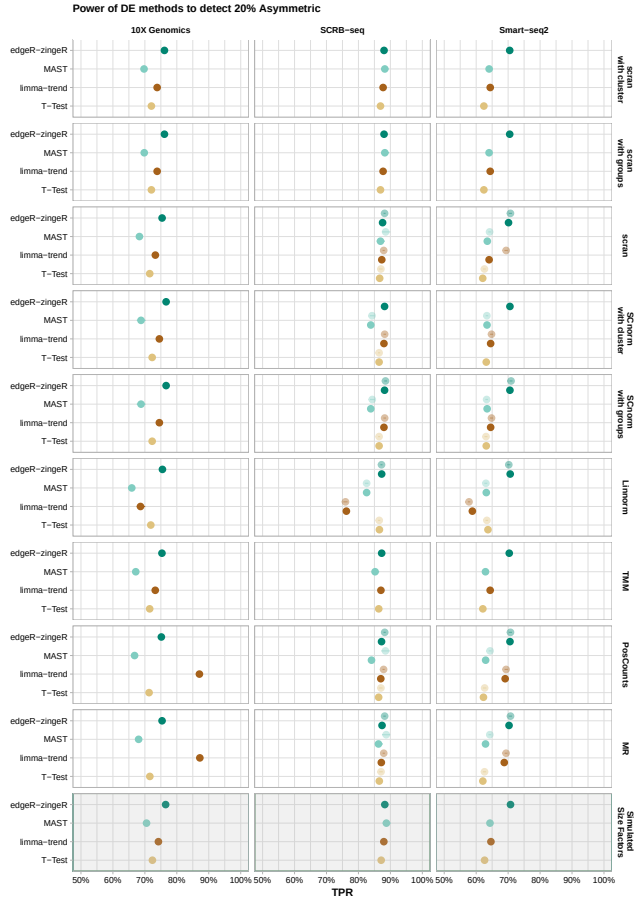
Supplementary Figure 16: Power of normalisation method using filtered or imputed counts. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated and 20% of the genes were differentially expressed following an asymmetric narrow gamma distribution. The power to detect DE (TPR) based on DE-testing with limma-trend per count preprocessing approach stratified over library preparation protocol and normalisation method is plotted (mean TPR ± s.d.).



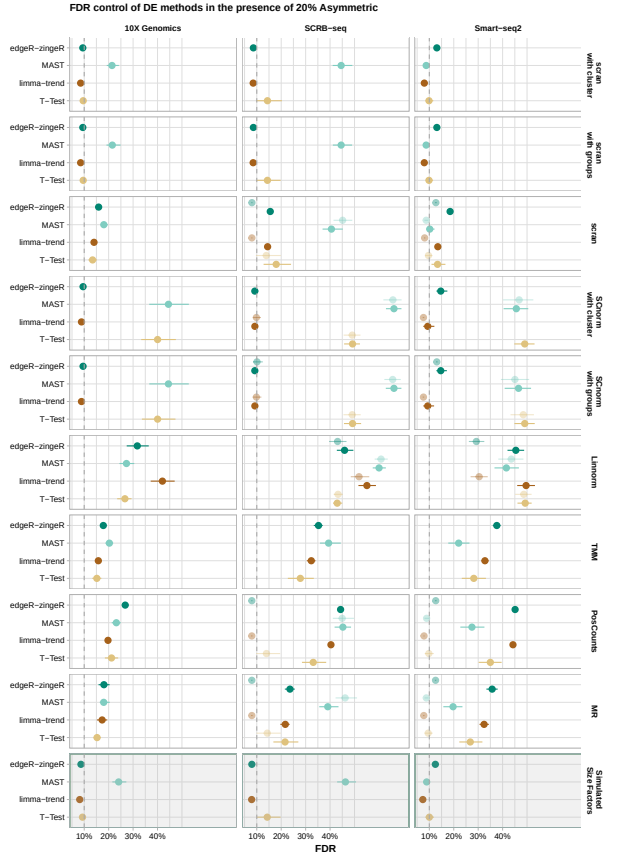
Supplementary Figure 17: FDR control of normalisation method using filtered or imputed counts. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated and 20% of the genes were differentially expressed following an asymmetric narrow gamma distribution. The BH-FDR control based on DE-testing with limma-trend per count preprocessing approach stratified over library preparation protocol and normalisation method is plotted (mean FDR ± s.d.). The dashed line indicates the nominal FDR level of 10%.



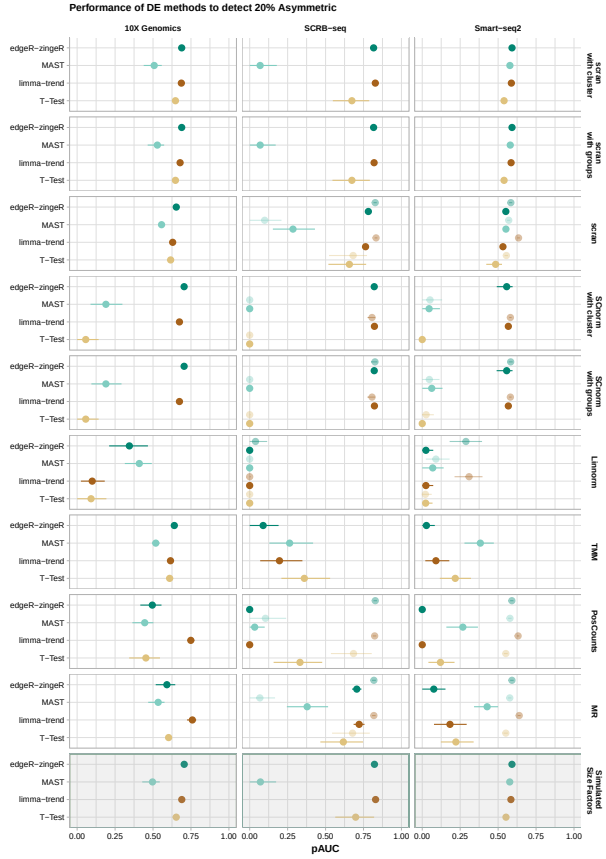
Supplementary Figure 18: Trade-off between power and false discoveries per normalisation method using filtered or imputed counts. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated and 20% of the genes were differentially expressed following an asymmetric narrow gamma distribution. The discriminatory ability determined by the partial area under the curve (pAUC) based on DE-testing with limma-trend per count preprocessing approach stratified over library preparation protocol and normalisation method is plotted (mean pAUC ± s.d.).



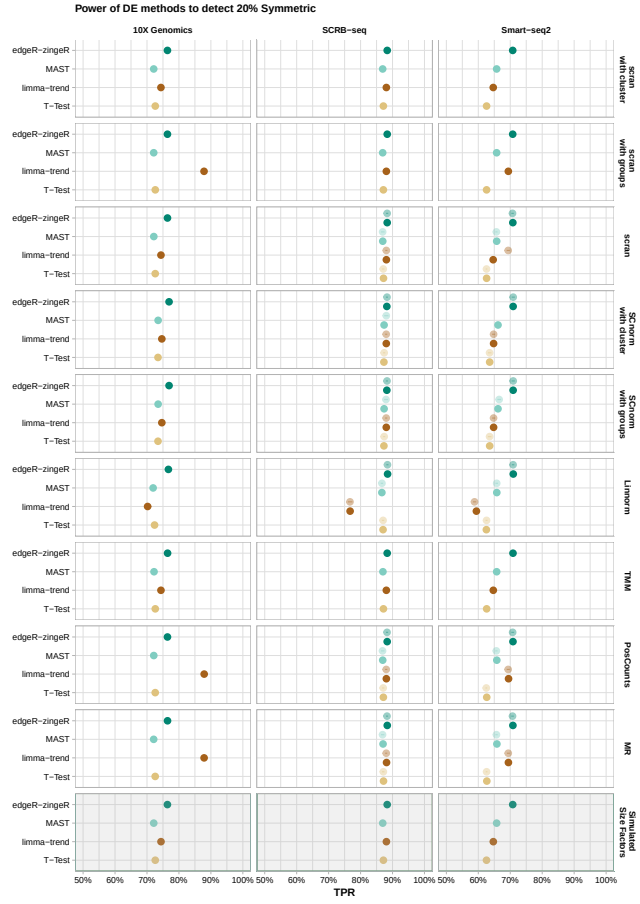
Supplementary Figure 19: Power of DE-tools for 20% Asymmetric. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation per protocol. 20% of the simulated genes are differentially expressed following an asymmetric narrow gamma distribution. Unfiltered counts were normalised using simulated library size factors or applying normalisation methods. Differential expression was tested using T-Test, limma-trend, MAST or edgeR-zingeR. The power to detect differential expression is plotted (mean TPR ± s.d.).



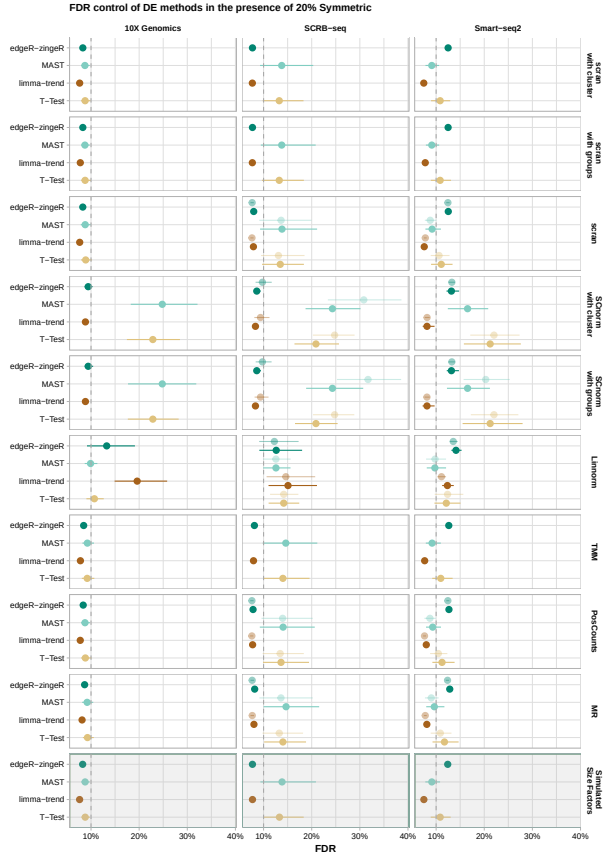
Supplementary Figure 20: FDR control of DE-tools for 20% Asymmetric. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation per protocol. 20% of the simulated genes are differentially expressed following an asymmetric narrow gamma distribution. Unfiltered counts were normalised using simulated library size factors or applying normalisation methods. Differential expression was tested using T-Test, limma-trend, MAST or edgeR-zingeR. The FDR control of the DE-methods is plotted (mean FDR ± s.d.). The dashed line indicates the nominal FDR level of 10%.



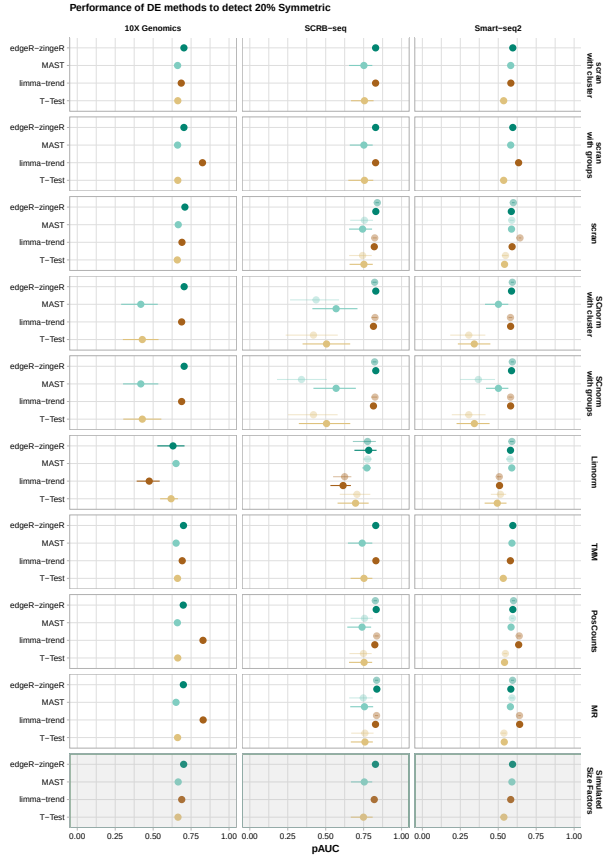
Supplementary Figure 21: The trade-off between power and false discoveries of DE-tools for 20% Asymmetric. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation per protocol. 20% of the simulated genes are differentially expressed following an asymmetric narrow gamma distribution. Unfiltered counts were normalised using simulated library size factors or applying normalisation methods. Differential expression was tested using T-Test, limma-trend, MAST or edgeR-zingeR. The discriminatory ability determined by the partial area under the curve (pAUC) based on the TPR-FDR curve is plotted (mean pAUC ± s.d.).



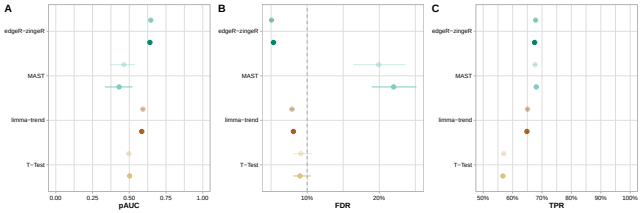
Supplementary Figure 22: Power of DE-tools for 20% Symmetric. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation per protocol. 20% of the simulated genes are differentially expressed following a symmetric narrow gamma distribution. Unfiltered counts were normalised using simulated library size factors or applying normalisation methods. Differential expression was tested using T-Test, limma-trend, MAST or edgeR-zingeR. The power to detect differential expression is plotted (mean TPR ± s.d.).



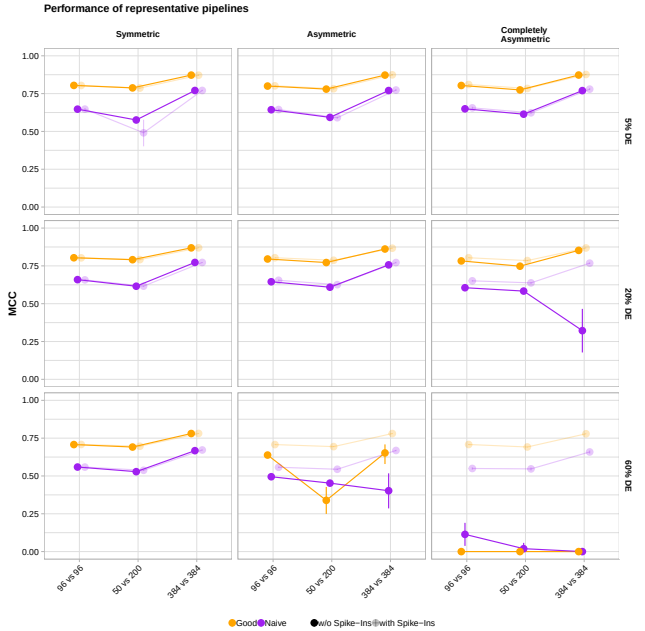
Supplementary Figure 23: FDR control of DE-tools for 20% Symmetric. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation per protocol. 20% of the simulated genes are differentially expressed following a symmetric narrow gamma distribution. Unfiltered counts were normalised using simulated library size factors or applying normalisation methods. Differential expression was tested using T-Test, limma-trend, MAST or edgeR-zingeR. The FDR control of the DE-methods is plotted (mean FDR ± s.d.). The dashed line indicates the nominal FDR level of 10%.



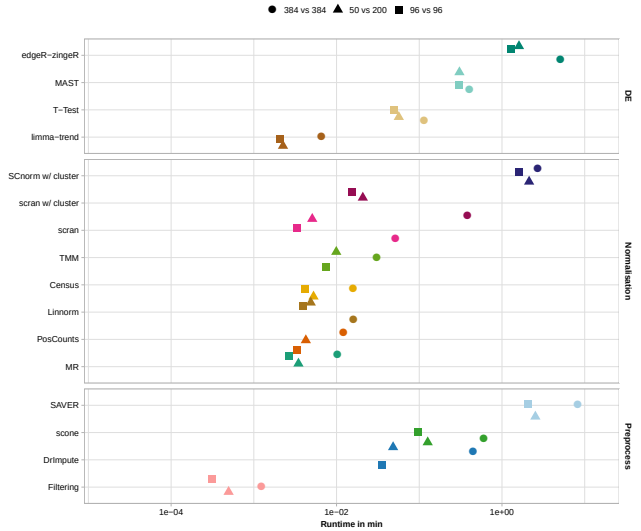
Supplementary Figure 24: The trade-off between power and false discoveries of DE-tools for 20% Symmetric. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation per protocol. 20% of the simulated genes are differentially expressed following a symmetric narrow gamma distribution. Unfiltered counts were normalised using simulated library size factors or applying normalisation methods. Differential expression was tested using T-Test, limma-trend, MAST or edgeR-zingeR. The discriminatory ability determined by the partial area under the curve (pAUC) based on the TPR-FDR curve is plotted (mean pAUC ± s.d.).



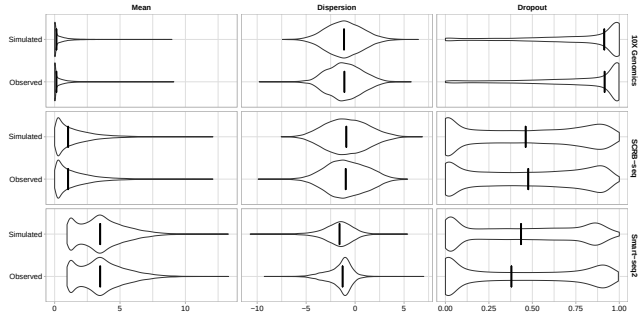
Supplementary Figure 25: Performance of DE-tools using Census normalisation. The expression of 10,000 genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation of Smart-seq2 data. 20% of the simulated genes are differentially expressed following an asymmetric narrow gamma distribution. Unfiltered counts were normalised using Census method. Differential expression was tested using T-Test, limma-trend, MAST or edgeR-zingeR. The lighter shade indicates the usage of spike-ins for normalisation.
A) The discriminatory ability determined by the partial area under the curve (pAUC) based on the TPR-FDR curve is plotted (mean pAUC ± s.d.). B) FDR control (mean FDR ± s.d.). The dashed line indicates the nominal FDR level of 10%. C) The power (TPR) to detect differential expression (mean TPR ± s.d.).



Supplementary Figure 26: Performance of pipelines. The expression of 10000 genes in 184, 250 or 768 cells were simulated and 5%, 20% or 60% of the genes were differentially expressed following a symmetric, asymmetric or completely asymmetric narrow gamma distribution. This simulation setup was applied to one good pipeline (SCRB-seq + STAR + GENCODE + no preprocessing + scran + limma-trend) and one naive pipeline (SCRB-seq + STAR + GENCODE + no preprocessing + MR + T-Test). For each analysis set, the Matthew Correlation Coefficient was averaged over 20 simulations (mean MCC ± s.d.) and rescaled to [0,1] interval.



Supplementary Figure 27: Computational Run Time. The real CPU-time per pipeline step and method stratified over sample size.



Supplementary Figure 28: Comparison of simulated and observed parameters. The marginal distribution of the observed and simulated log2 mean, log2 dispersion and gene dropout rate for Smart-seq2, SCRB-seq and 10X Genomics HGMM scRNA-seq data. For Smart-seq2 the mean and dispersion value excluding zeroes is plotted since a ZINB distribution is assumed. Black line indicates the median value.

Supplementary Tables

Protocol	Description	Cell Processing	Full Length	UMI	Number of cells	ERCC Spike-ins	Raw reads per cell
CEL-seq2	J1 mESC cultured in 2i/LIF medium (two batches)	Fluidigm C1	-	+	48 + 48	+	1 million
Drop-seq	J1 mESC cultured in 2i/LIF medium (two batches)	Droplets	-	+	45 + 34	-	1 million
SCRB-seq	J1 mESC cultured in 2i/LIF medium (two batches)	FACS	-	+	44 + 49	+	1 million
Smart-seq2	J1 mESC cultured in 2i/LIF medium (two batches)	FACS	+	-	40 + 45	+	1 million
10X Genomics	NH3T3 mouse cells (originally 1:1 mixture of human and mouse cells with a total of 1k cells)	10X Genomics Chromium	-	+	473	-	~60 thousand
10X Genomics	Peripheral blood mononuclear cells (PBMC)	10X Genomics Chromium	-	+	1022	-	~54 thousand

Supplementary Table 1: Description of single cell RNA-sequencing data sets.

Name	Version	Number of Transcripts	Number of Genes	Number of Exons per Transcript	Number of Transcripts per Gene	Transcript Length	Gene Length
GENCODE	M15	131195	52636	6	2	1690	2348
Vega	68	110696	41175	6	3	1718	2671
RefSeq (curated)	85	34890	-	9	-	2852	-

Supplementary Table 2: Description of *Mus musculus* transcript annotations.

Name	Command	Version	Ref.
BWA	`bwa index -p [annotation_transcriptome_ercc.fa]`	0.7.12	[9]
kallisto	`kallisto index -i [annotation_transcriptome_ercc.fa]`	0.43.1	[5]
BWA	`bwa aln -t [bwa-index] [protocol_cDNA_reads.fastq] > [protocol.reads.sai]`	0.7.12	[9]
BWA	`bwa samse [reads.sai] [protocol_cDNA_reads.fastq] > [aligned.sam]`	0.7.12	[9]
kallisto &Smart-seq2	`kallisto pseudo -i [kallisto-index] -o [outputpath] -b [protocol_cDNA_reads.fastq] --single -l [Mean_FragmentLength_Protocol] -d [SD_FragmentLength_Protocol]`	0.43.1	[5]
kallisto & UMI	`kallisto pseudo -i [kallisto-index] -o [outputpath] -b [protocol_cDNA_reads.fastq] --single --umi -l [Mean_FragmentLength_Protocol] -d [SD_FragmentLength_Protocol]`	0.43.1	[5]
STAR zUMIs	`STAR --runThreadN 12 --runMode genomeGenerate --genomeDir /index/star/[annotation] --genomeFastaFiles [mm10_genome_annotation_ercc.fa] --sjdbGTFfile [mm10_genome_annotation_ercc.gtf] --sjdbOverhang 44`	2.5.3a	[10]
zUMIs	`bash <path-to-zUMIs>/zUMIs-master.sh -y parameters.yaml ter.sh -f [protocol_barcode_reads.fq.gz] -r [protocol_cDNA_reads.fq.gz] -n [protocol-batch] -g [star-index] -a [mm10_genome_annotation_ercc.gtf] -c [cell-barcode-range] -m [umi-barcode-range] -l [read-length] -b [expected-cell-barcodes.txt] -o [outputpath] -d 1000000 -R no -S yes -s 0 -i [zUMIs-pipeline-path]`	0.0.3	[11]

Supplementary Table 3: Alignment and assignment commands for expression quantification.

Pipeline Step	Method Name	Version	Reference
Preprocessing	Gene Dropout Filtering	-	-
Preprocessing	DrImpute	1.0	[12]
Preprocessing	scone	1.6.1	[13]
Preprocessing	SAVER	1.1.1	[14]
Normalisation	MR in DESeq2	1.22.2	[15]
Normalisation	PosCounts in DESeq2	1.22.2	[15]
Normalisation	TMM in edgeR	3.24.3	[16]
Normalisation	Census in monocle	2.10.1	[17]
Normalisation	Linnorm	2.6.1	[18]
Normalisation	SCnorm	1.4.3	[19]
Normalisation	scran	1.10.1	[20]
DE-tool	T-Test in stats	3.5.3	[21]
DE-tool	limma-trend	3.38.3	[22]
DE-tool	MAST	1.8.2	[23]
DE-tool	edgeR-zingeR	0.1.0	[24]

Supplementary Table 4: Description of implemented methods in powsimR.

References

1. Ziegenhain, C. *et al.* Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell* **65**, 631–643 (2017).

2. Kolodziejczyk, A. A. *et al.* Single cell RNA-Sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).

3. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).

4. Genomics, X. hgmm_1k - datasets - single cell gene expression - official 10x genomics support. https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm_1k (2018). Accessed: 2019-3-15.

5. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

6. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120–e133 (2018).

7. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).

8. Soneson, C. & Robinson, M. D. iCOBRA: open, reproducible, standardized and live method benchmarking. *Nat. Methods* **13**, 283 (2016).

9. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

10. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

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

12. Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. & Garry, D. J. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* **19**, 220–230 (2018).

13. Cole, M. B. *et al.* Performance assessment and selection of normalization procedures for Single-Cell RNA-Seq. *Cell Syst* **8**, 315–328 (2019).

14. Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**, 539–542 (2018).

15. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550–571 (2014).

16. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25–R34 (2010).

17. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with census. *Nat. Methods* **14**, 309–315 (2017).

18. Yip, S. H., Wang, P., Kocher, J.-P. A., Sham, P. C. & Wang, J. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.* **45**, e179–e191 (2017).

19. Bacher, R. *et al.* SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* **14**, 584–586 (2017).

20. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75–89 (2016).

21. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2019). URL https://www.R-project.org/.

22. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29–R46 (2014).

23. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 1–13 (2015).

24. Van den Berge, K. *et al.* Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* **19**, 24–41 (2018).

3 | Discussion

The recent development of single-cell RNA sequencing protocols enables the quantification of genome-wide expression profiles of thousands to millions of single cells opening up possibilities to systematically characterize cells and the underlying developmental and regulatory mechanisms[213,88,214,215,86]. Since the first publication on single-cell transcriptomics by Tang and coworkers in 2009[44], hundreds of single-cell RNA-seq data sets from a variety of sources have been released, profiling gene expression of sorted cells[216,52], tumors[217,50], whole dissociated organs[58,102] and even complete organisms[130,131]. Currently, it is also the main tool to systematically characterize cells in several Atlas Projects like the Human Cell Atlas[46,218]. Furthermore, scRNA-seq provides a powerful tool to reconstruct developmental patterns by sampling cells during differentiation processes[219,220,45] and to characterize intratumoral heterogeneity[221].

As scRNA-seq is increasingly used by many labs but still very novel, best practices are not yet defined, neither for wet lab protocols nor for computational analysis tools. Quantitative and independent investigations of scRNA-seq workflows is urgently needed to guide important decisions on experimental design choices. I contributed to these efforts during my PhD by developing realistic simulations for single-cell RNA-sequencing experiments, powsimR, and showed that statistical power analysis is an ideal framework not only to determine the best experimental design in terms of sample size plans but is also integral for evaluation and comparison of all steps in a RNA-seq workflow.

Establishing realistic simulations for statistical power analysis of scRNA-seq experiments (Manuscript III)

An integral first step in designing a simulation framework based on probability distributions is finding the most appropriate sampling distribution for count data. Thus we evaluated the fit of five distributions, namely the negative binomial (NB), the zero inflated negative binomial (ZINB), Poisson (P), zero-inflated Poisson (ZIP) and the Beta-Poisson (BP) for a total of 8 published studies that utilized 9 different RNA-seq library preparation methods [154]. Similar to previous studies [63,99], we found that even though single-cell data are very sparse with low mean expression values compared to bulk data, the majority of gene expression distributions is still best modeled by a negative binomial distribution. The zero-inflated negative binomial is only the best fitting distribution for at most a quarter of the genes, if amplification noise cannot be removed by experimental means [63,210]. We thus refrain from using a mixture distribution, however for some of the protocols that do not utilize UMIs, such as e.g. Smart-Seq2, the ZINB might provide a better fit and should be used as a sampling distribution in the simulations of scRNA-seq experiments. In any case, both distributions, NB and ZINB, are implemented in powsimR. Given the chosen distributions, we estimate the observed mean, dispersion and zero, also called gene dropout, expression values per gene. We then explicitly model the observed mean-dispersion relationship by a locally weighted polynomial regression fit to simulate expression values closely mimicking the observed mean-variance distribution as well as the gene dropout rates.

This simulation framework is the core of powsimR on which we then build a flexible tool to assess statistical power and sample size requirements for differential expression analysis of RNA-seq experiments. Firstly, we include all integral steps of a typical DE analysis pipeline: We have implemented 8 preprocessing and imputation, 9 normalisation methods and 14 DE-tools. Secondly, we offer a flexible framework to define effect sizes, sample size and sequencing depth designs, and possible batch effects. Thirdly, the test results are integrated to calculate error matrices for evaluation, where the True Positive Rate (TPR) and False Discovery Rate (FDR) are the most informative in relation to experimental design

question of RNA-seq experiments. Particularly, these matrices can be conditional on mean, dispersion and dropout values so that for example sample size designs can also be evaluated in terms of power to detect a significant biological differences for lowly expressed or highly variable expressed genes. In addition to providing a framework to design sample size plans ensuring sufficient statistical power, powsimR can also inform about the power to detect a biological signal in a data set at hand. This is particularly useful for posterior power analyses to compare conducted experiments in order to rule out lack of power as the reason for incongruities in DE genes. Furthermore, the quality of synthetic data generated by powsimR has been independently validated[222] and the simulation framework has also been used to conduct a comprehensive evaluation of differential testing algorithms[153]. powsimR's simulation and statistical power analysis capabilities formed an integral part of our analyses in the following research studies which also contributed significantly to its ongoing further development and extension.

Power of scRNA-seq library preparation protocols is determined by capture efficiency and amplification noise (Manuscript I & II)

We used a preliminary version of powsimR in two other studies[223,210]. In Parekh et al. 2016, we investigated the impact of whole transcriptome amplification by PCR on the sensitivity, accuracy and precision of gene expression quantification. PCR amplification is an essential step in single-cell RNA-sequencing because single cells contain only very small amounts of mRNA that results in many duplicates. To that end, we analysed bulk and single-cell data sets generated with three library preparation protocols which differ in amount of starting material, fragmentation method, number and occurrence of amplification cycles as well as cellular and molecular barcoding abilities. At that time, one strategy to deal with amplification bias was to identify so called PCR duplicates computationally, based on their 5' end mapping position[224,225]. Based on this definition, we identified read duplicates as a common phenomenon, especially in the common single-end sequencing read design which can be explained by a random sampling model. However, we could show that this

strategy for RNA-seq data introduces more problems than it solves, because fragmentation in RNA-seq libraries is non random and thus the same sequence fragment could also have originated from distinct RNA molecules and is not a result of biased amplification. These natural duplicates have also previously been observed in high fractions in other data sets [226]. We determined that the chance to incorrectly remove natural duplicates instead of PCR duplicates by mapping position increases with higher expression levels of the gene and deeper sequencing depths. In addition, and fragmentation bias of the library preparation method. Given these findings, we evaluated the impact of PCR duplicates and their computational removal on the accuracy of transcript quantification using ERCC spike-ins and could prove that in no case does the removal of read duplicates improve relative abundance estimation of gene expression.

Most importantly, we investigated how much noise or bias PCR amplification introduces, how this affects the power to detect differential expression and whether duplicate removal has a positive or negative effect on error rates. To this end, we used a preliminary version of powsimR to simulate differential expression using the observed mean-variance relationship of count data with and without duplicates. We could prove that tagging sequences originating from the same RNA molecule by unique molecular identifiers (UMIs) did significantly lower the technical variance compared to computational removal. This decrease in variance contributed to the larger statistical power of UMI libraries while ensuring control over false detections. On the other hand, the computational removal of PCR-duplicates resulted in a decrease in power and an FDR exceeding the nominal level. In conclusion, we advise against the pure computational removal of read duplicates due to the associated loss of natural duplicates arising from sampling of real independent molecules. Instead, we recommend the early pooling and tagging of RNA molecules by UMIs.

Following up on this, we conducted an in-depth comparison of six popular single-cell RNA-seq library preparation protocols in Ziegenhain et al. 2017 (Smart-seq [55], Smart-seq2 [227], CEL-seq2 [64], SCRB-seq [65], Drop-seq [58] and MARS-seq [228]). Four of the methods use UMIs of which one protocol utilizes in vitro transcription for linear amplification (IVT) instead of exponential amplification by polymerase chain reaction (PCR). In their current implementation, three protocols are well-plate based while two were run on a microfluidic

system and one utilizes droplets. The protocols were applied to identically processed mouse embryonic stem cells in two batches supplemented with ERCC spike-ins[38], providing an ideal benchmarking data set to quantitatively measure and compare sensitivity, accuracy and precision.

The capacity of a protocol to observe the transcriptome as completely as possible, i.e. its sensitivity, is especially important in scRNA-seq experiments. Here, the limiting step is mainly the reverse transcription and second strand synthesis reactions of minute starting RNA material. Currently, the estimated efficiency to capture mRNA molecules is between 10 to 50%[85,63,80]. While it was found that reaction volumes in the nanoliter range, as implemented in microfluidic devices, can improve sensitivity[64,80,229,55], we found that the currently most sensitive protocols are still plate-based, particularly Smart-seq2, in our comparison as well as other benchmarks[120,115,229]. Another important aspect of scRNA-seq protocols is the accuracy of gene expression estimates. Exogenous RNA molecules (ERCC,SIRV)[37,38,230] spanning various known concentration ranges are spiked into single-cell lysates, can be used for accuracy estimation. Similar to another study[229], we found scRNA-seq methods are accurate and thus quantify expression levels well. However, spike-ins have their limitations and in how far the accuracy estimation based on spike-ins can be translated to cellular gene expression is still under debate. Lastly, precision describes the variability of measured gene expression estimates. As discussed earlier, considerable amplification from the minute starting material is needed in single-cell RNA-seq experiments, introducing amplification noise as we showed previously[223]. This technical variation exceeding the expectation of Poisson sampling of molecules is Extra-Poisson variability. As we showed previously, incorporating unique molecular identifiers (UMIs) makes it possible to distinguish read duplicates from natural duplicate molecules and thus technical noise generated during amplification can be removed. However, we found here that the variability is more pronounced in PCR-based methods with exponential amplification than in linear IVT amplification.

Nevertheless, accuracy and precision are not independent parameters, both are strongly coupled to sensitivity. For accuracy measured by spike-ins, the correlation coefficient is largely dependent on the number of ERCC transcripts detected in the cell, and thus is strongly linked to sensitivity. Similarly, accuracy is not strongly affected by sequencing depth[229], as

long as transcripts remain detected. Likewise, precision is not an independent entity. For full-length scRNA-seq protocols, where UMIs cannot yet be easily accommodated to remove amplification noise, more sensitive protocols like Smart-seq2 [227] show significantly lower Extra-Poisson Variability as the larger initial cDNA complexity requires less amplification. On the other hand, IVT amplified libraries showed the highest sensitivity at lower sequencing depth among the UMI methods.

In summary, given our findings concerning these parameters individually as well as their interdependencies, no single best library preparation protocol exists. Thus, as the more simple descriptive statistics have limitations in comparing performance, we chose simulations in order to investigate the combined effects of sensitivity and precision on the power of each method to detect a meaning biological expression difference [154]. As a first step, we selected a common subset of genes for estimation and simulation in powsimR. Of note, we included the presence of undetected genes in our simulations, so that the power simulations considered the full range of observed gene dropout rates and is not biased against more sensitive methods. Given the low technical variance of the methods, we chose to draw effect sizes from the observed distribution of fold changes with moderate differences between two microglial subpopulations as previously profiled by Zeisel et al. 2015 [102]. We found that the UMI well-plate based library preparation protocol SCRB-seq achieved 80% power with the smallest sample size of 64 cells per group. This is most likely due to the fairly high sensitivity, second to Smart-seq2, and reduced amplification noise due to the use of UMIs. In addition, we investigated in how far the power depends on sequencing depth. Interestingly, UMI protocols utilizing in vitro amplification were less affected by downsampling due to the lower technical noise in our comparison. Indeed, a recent benchmark of scRNA-seq protocols also applied downsampling and found that IVT amplified libraries retained a high sensitivity as well as high accuracy in cell type classification [115]. Sequencing costs are still substantial in large scale cell atlas projects (e.g. mouse cell atlas [129]), even with the strong decrease in sequencing costs [13]. Therefore, finding an optimal balance between replication and sequencing depth under sample availability and budget constraints has recently received considerable attention [188,231,179]. In this comparison, we could show that achieving uniform amplification contributes to information content maximization obtained by scRNA-seq.

In conclusion, we conducted the first comprehensive comparison of scRNA-seq library preparation protocols. With our power simulations, we were able to translate our findings into experimental costs for a given setup and in that way enumerate the cost efficiency of each method in a meaningful way. This enables researchers to make informed choices.

Quantification of intronic expression supplements power of UMI scRNA-seq methods (Manuscript IV)

The novelty of the recent scRNA-seq methods containing UMIs and cell barcodes also poses a challenge for the basic data processing to yield UMI count matrices from sequencing reads. At the time, no analysis pipeline existed that included all functionalities that we wished for in a fast manner. Therefore, we developed a fast and flexible pipeline for the analysis of scRNA-seq data, zUMIs[232]. zUMIs can handle data from all kinds of scRNA-seq protocols: Firstly, it can use known cell barcodes and also automatically identify the barcodes that are likely to be associated with intact cells. Specifically, we fit a k-dimensional multivariate normal distribution to the number of reads per cellular barcode and choose only the last peak with the largest average to automatically select the barcodes with the most number of associated reads. This straight-forward approach is able to differentiate viable cells from debris which is particularly important for droplet-based methods and microfluidic devices that contain unknown numbers of barcoded cell transcriptomes[94,70]. In addition, we evaluated several methods to reliably identify cellular barcodes and UMIs and found that sequencing quality (PHRED score) filtering is a fast and accurate approach. Secondly, users can flexibly define the locations and lengths of cell barcode, UMI and cDNA sequences in the input sequencing files. Lastly, zUMIs provides basic summary statistics for quality control as well as more specialised analyses. Particularly, zUMis is the only pipeline that offers a downsampling utility that allows users to assess whether the library has been sequenced to saturation. Specifically, we have implemented adaptive downsampling of overrepresented libraries that are within three absolute median deviations of all sequenced libraries[95]. Apart from library saturation analysis, downsampling of UMI data has been suggested as a normalisation

approach for relative expression comparisons across single cells[233].

Recently, a number of scRNA-seq library preparation protocols have been developed and applied to isolated single nuclei of cells[234,235,236,237]. Nuclei can usually be rapidly and easily isolated from lightly fixed, frozen tissues and archived without the extended incubation and processing required for isolating single cells. On the other hand, the sequenced libraries of single nuclei contain a significant fraction of nascent, unspliced mRNAs. We were therefore interested in extending gene expression estimation with intronic mapping reads in zUMIs. We could show that including intronic reads in gene expression quantification achieved an increased resolution of identifiable clusters and an increased number of marker genes detected thereby improving the sensitivity and precision of gene expression estimation in scRNA-seq data sets. In order to evaluate the power gained by counting exon as well as intron mapping reads in gene expression quantification, we performed power simulations using powsimR. We could show that the power to detect differential expression differences of lowly expressed genes is higher when intronic mapping reads are considered.In addition, this increase in power does not come at the cost of an increased false detection rates. In conclusion, this makes exon plus intron counting worthwhile, especially from low coverage data enriched with nuclear nascent RNA transcripts. Furthermore, there is the exciting possibility with RNA velocity to reconstruct cell lineage and developmental trajectories[238]. For that, the abundance of unspliced and spliced RNA needs to be estimated from scRNA-seq data, and zUMIs is one of the few tools providing this functionality.

Library preparation and normalisation methods have the biggest impact on the performance of scRNA-seq studies (Manuscript V)

Many experimental protocols and computational analysis approaches exist for scRNA-seq. Often, computational pipelines used in scRNA-seq studies are pieced together without thorough quantitative evaluation. If such benchmarking efforts exists, they at best cover only separate and distinct analysis steps of a pipeline. As the performance measures differ among the comparisons they do not allow to quantify how the steps interact and how individual

steps impact the performance of an entire computational pipeline. Especially the impact of the combined choices of the separate analysis steps on overall pipeline performance has not been quantified yet. At a minimum, this includes choices of (1) library preparation protocols, (2) mapping and annotation, (3) preprocessing of the count matrix by filtering or imputation, and (4) normalisation to allow comparisons across cells. The resulting filtered and normalised count matrix is the most downstream standard output of any scRNA-seq pipeline. From here on out the types of analyses diverge, ranging from differential expression (DE) analysis, cluster analysis, classification of cells to the analysis of trajectories[86].

In order to achieve a fair and unbiased comparison of computational pipelines, simulations of realistic data sets are necessary. This is because the ground truth for real data sets is unknown and alternatives, such as concordance analyses are bound to favor similar methods and not necessarily better methods. To this end, we integrated frequently used methods for each analysis step into our simulation framework powsimR[154,210]. As the basis for realistic simulations, powsimR uses a raw count matrix to define the technical variance and model the mean-variance relationship of gene expression levels. By adding varying levels of differential expression, we can measure the sensitivity and specificity of each pipeline based on how faithfully DE-genes are recovered.

A number of publications already conducted detailed performance evaluations of RNA-seq to genome mapping[73,239], therefore we chose to rather evaluate BWA[32], STAR[31] and kallisto [74], three popular aligners that reflect the overall breadth in current approaches. Similarly, we tried to cover several preprocessing approaches including the imputation methods SAVER[164], DrImpute[169] and scone[240], while other imputation methods turned out to be prohibitively slow (e.g. scImpute[166]). Finally, the sparsity of scRNA-seq count matrices poses a formidable challenge to normalisation and several methods have been developed to tackle this problem, ranging from applying bulk methods such as TMM[104], DESeq's Median-Ratio (MR)[105] and adaptations thereof (PosCounts), converting relative RNA-seq expression levels into absolute transcript counts as in Census[84], pooling and deconvoluting with scran[99] to regression based methods implemented in SCnorm[106] and Linnorm[241]. Additionally, there are a number of studies evaluating the performance of differential gene expression analysis methods for scRNA-seq data[147,242,153]. Based on these comparisons, we chose four representative methods:

two methods usually applied to bulk data, T-Test and limma[243], and two methods developed for scRNA-seq, MAST[150] and zingeR[159].

One main assumption in traditional DE-analysis is that differences in expression are symmetric, preferably only a small fraction of genes of interest is DE while the expression of the majority of genes remains constant and hence that the total mRNA content is approximately the same in all samples[98]. This is no longer true when diverse cell types are considered, e.g. in the brain[102]. Thus, in contrast to other studies, we simulated varying numbers of DE-genes in conjunction with small to large differences in mRNA content, thus covering the entire spectrum of possible DE-settings.

Our realistic simulations in conjunction with the wide array of methods, allow us not only to quantify the performance within each pipeline step, but also to quantify interdependencies among the pipeline steps and their relative importance within the whole pipeline. We found that library preparation and normalisation methods have the biggest impact on the performance of scRNA-seq studies. More specifically, library protocols determine the ability to detect small symmetric differences. On the other hand, normalisation choices have the biggest effect when comparing asymmetric and very diverged expression profiles, where prior cell clustering or the usage of spike-ins improves performance. Interestingly, choices of imputation has little effect on performance, even for very sparse data. More markedly, it actually increased the false detection in our comparison. In line with previous findings[153], we could show that DE-tools specifically developed for scRNA-seq experiments were not superior to bulk tools. Rather, a good normalisation prior to testing alleviates the need for more complex approaches assuming zero-inflated distributions. Another recent benchmarking effort of scRNA-seq pipeline identified another interdependence, namely between normalisation and trajectory reconstruction as well as data integration methods[190]. Thus, normalisation determines the performance of many downstream analyses.

In addition, we investigated the effect on marker gene detection in a complex mixtures of cells using our recommended pipeline and could show that it performed better than the naive approach of using established bulk RNA-sequencing analysis techniques to single-cell RNA-seq expression data. Thus we could prove that our simulation and evaluation using DE testing results of pairwise comparisons can be translated to more complex tasks such as cell

type classification and marker gene detection in unknown cell mixtures.

Optimizing single-cell RNA sequencing

With our analyses we can provide sound recommendations for constructing an optimal scRNA-seq pipeline covering all essential steps from library preparation to differential expression analysis[223,232,154,155,210]. Nevertheless, recommendations should not be equated with defining a universally applicable pipeline that will fit any and all research questions.

Validation of new library preparation protocols by external sources such as smFISH are essential in method development and evaluation of new library preparation protocols, but only two protocols out of the eleven considered in a recent benchmark[115] included this [63,59]. Benchmarking efforts have recently extended the evaluation of library preparation methods to the capture efficiencies particularly of rare, but well-defined cell types using complex mixtures of cells[120,115], highlighting the importance of more upstream processes of the scRNA-seq pipeline. Attempts to improve the single cell capture rate can also been seen in the recent development of new well-plate technologies spanning low cost, easy installable methods like Microwell-seq[128] and Seq-Well[244] to commercial automated systems like Celsee [245] and ICELL8[246].

In any case, the increased number of cells profiled, particularly in cell atlas projects [46,129], covering millions of cells across tissues and subjects, necessitates the development of reproducible and reliable computational pipelines beginning with the preprocessing of sequencing reads to generate count matrices of gene expression[247]. For example, pseudoalignment quantification show promising trade-offs in terms of efficiency, accuracy and processing speed[248,155]. But the evaluation of these alternative methods is limited to comparing gene detection and accuracy of expression estimates. Therefore, read sequencing simulations (e.g. Flux simulator[249]) need to be extended to the current designs of library preparation protocols, including options for transposome mediated fragmentation, 3' prime coverage of transcripts and most importantly simulation of sequencing reads of cellular barcodes and UMIs used for tagging mRNA molecules and cells (Valtierra et al., unpublished). Generating these realistic read sequences could further help to systematically compare different alignment

approaches and in how far these methods are sensitive to differences in library generation. In any case, one possible limitation of pseudoalignments already evident is that if the gene expression of cells in species with subpar annotations are profiled, the k-mer read matching might be biased. Classical genome alignments to closely related species with good genome annotations should then be considered[250]. Another route, albeit associated with an increase in costs and still under active development, is the construction of reference transcriptomes for example by Nanopore native RNA sequencing[251,252]. Long-read sequencing technologies in general allow the identification of new isoforms and isoform features such as splice sites, transcription start and polyA sites, thereby helping to unambiguously annotate and quantify transcriptomes[253]. Nevertheless, given the overall decrease in sequencing costs, other types of measurements besides scRNA-seq can be useful supplements. Conventional bulk RNA-seq for example should be regarded as a reference for estimating the likelihood of dropout events and data smoothing approaches[254], and not only as an outdated averaging of gene expression profiles.

In summary, single-cell omics technologies are rapidly evolving and more widely applicable. Single-cell RNA-sequencing in particular is integral to characterizing cellular phenotypes. By developing realistic simulations for scRNA-seq experiments with powsimR, we have provided an excellent framework for benchmarking efforts, conducted detailed method comparisons and provided guidance in setting up experiments.

4 | Conclusion and Outlook

Single-cell RNA sequencing is clearly a transformative tool with wide applicability to biological and biomedical questions but researchers need more guidance to choose an appropriate library preparation method and computational analysis pipeline for their experiment. My contribution to this effort is in developing a realistic and versatile simulation framework for scRNA-seq. Using this framework, we were able to generate synthetic data closely resembling observed expression profiles that we used to conduct statistical power analysis and evaluate computational as well as library preparation methods for scRNA-seq. In that regard, I have made a significant contribution to the new field of single-cell RNA sequencing data analysis with my thesis work. The increase in throughput of scRNA-seq has been tremendous, making cell atlases projects realistic, feasible and fast endeavors. Given the mission statement of the Human Cell Atlas, namely profiling and mapping of cell types, the precise definition of a cell type and in particular how expression governs cellular identity and makes it unique to all other cells, is paramount. Therefore, we will extend our realistic simulation framework to complex cell mixtures to determine the signal-to-noise ratio needed to delineate distinct cell types and thereby contribute to the ongoing efforts of cell atlas projects. Very recently, many protocols have been developed to enable multiple measurements from the same individual cells, including methylation, chromatin state, protein expression, lineage tracing or spatial location[255,117]. These advancements offer exciting new possibilities to characterize individual cells at an even greater resolution but also pose great challenges with regard to multi-omics data integration[256]. I am confident that simulations of realistic synthetic multi-omics data will be integral to method development and systematic benchmarking efforts, and by that ultimately help us in our endeavors to better understand the cell.

Bibliography

1. F H Crick. On protein synthesis. *Symp. Soc. Exp. Biol.*, 12:138–163, 1958.

2. F Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

3. J D Watson and F H Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

4. Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Norton & Company, 6th revised edition. revised edition, 2014. ISBN 9780815344643.

5. Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.*, 20(4):207–220, 2019.

6. Assaf Zemach, Ivy E McDaniel, Pedro Silva, and Daniel Zilberman. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, 328(5980):916–919, 2010.

7. Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A Harvey Millar, James A Thomson, Bing Ren, and Joseph R Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, 2009.

8. Robin Andersson. Promoter or enhancer, what's the difference? deconstruction of established distinctions and presentation of a unifying model. *Bioessays*, 37(3):314–323, 2015.

9. Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, 2009.

10. A Schulze and J Downward. Navigating gene expression using microarrays–a technology review. *Nat. Cell Biol.*, 3(8):E190–5, 2001.

11. Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One*, 9(1):e78644, 2014.

12. Celine Everaert, Manuel Luypaert, Jesper L V Maag, Quek Xiu Cheng, Marcel E Dinger, Jan Hellemans, and Pieter Mestdagh. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci. Rep.*, 7(1):1559, 2017.

13. Wetterstrand KA. DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). `www.genome.gov/sequencingcostsdata`. Accessed: 2019-9-20.

14. Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Mol. Cell*, 58(4):586–597, 2015.

15. I Vomelová, Z Vanícková, and A Sedo. Methods of RNA purification. all ways (should) lead to rome. *Folia Biol.*, 55(6):243–251, 2009.

16. Arnold Berk, Chris A Kaiser, Harvey Lodish, Angelika Amon, Hidde Ploegh, Anthony Bretscher, Monty Krieger, and Kelsey C Martin. *Molecular Cell Biology*. WH Freeman, 8 edition, 2016. ISBN 9781464187445.

17. Alexander F Palazzo and Eliza S Lee. Non-coding RNA: what is functional and what is junk? *Front. Genet.*, 6:2, 2015.

18. Kimberly R Kukurba and Stephen B Montgomery. RNA sequencing and analysis. *Cold Spring Harb. Protoc.*, 2015(11):951–969, 2015.

19. Morgane Boone, Andries De Koker, and Nico Callewaert. Capturing the 'ome': the expanding molecular toolbox for RNA and DNA library construction. *Nucleic Acids Res.*, 46(6):2701–2721, 2018.

20. Shanrong Zhao, Ying Zhang, Ramya Gamini, Baohong Zhang, and David von Schack. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polya+ selection versus rRNA depletion. *Sci. Rep.*, 8(1):4781, 2018.

21. Illumina, Inc. An introduction to Next-Generation sequencing technology. 2017.

22. Malachi Griffith, Jason R Walker, Nicholas C Spies, Benjamin J Ainscough, and Obi L Griffith. Informatics for RNA sequencing: A web resource for analysis on the cloud. *PLoS Comput. Biol.*, 11(8):e1004393, 2015.

23. Andrew Adey, Hilary G Morrison, Asan, Xu Xun, Jacob O Kitzman, Emily H Turner, Bethany Stackhouse, Alexandra P MacKenzie, Nicholas C Caruccio, Xiuqing Zhang, and Jay Shendure. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.*, 11 (12):R119, 2010.

24. Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–628, 2008.

25. Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17(6):333–351, 2016.

26. H P J Buermans and J T den Dunnen. Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta*, 1842(10):1932–1941, 2014.

27. Barton E Slatko, Andrew F Gardner, and Frederick M Ausubel. Overview of Next-Generation sequencing technologies. *Curr. Protoc. Mol. Biol.*, 122(1):e59, April 2018. ISSN 1934-3639, 1934-3647. doi: 10.1002/cpmb.59.

28. Vijender Chaitankar, Gökhan Karakülah, Rinki Ratnapriya, Felipe O Giuste, Matthew J Brooks, and Anand Swaroop. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. *Prog. Retin. Eye Res.*, 55:1–31, 2016.

29. Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biol.*, 17:13, 2016.

30. Gabriel Renaud, Udo Stenzel, Tomislav Maricic, Victor Wiebe, and Janet Kelso. deML: robust demultiplexing of illumina sequences using a likelihood-based approach. *Bioinformatics*, 31(5):770–772, 2015.

31. 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, 2013.

32. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

33. Yang Liao, Gordon K Smyth, and Wei Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.*, 41(10):e108, May 2013. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkt214.

34. Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, January 2015. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btu638.

35. Xing Li, Asha Nair, Shengqin Wang, and Liguo Wang. Quality control of RNA-seq experiments. *Methods Mol. Biol.*, 1269:137–146, 2015.

36. Quanhu Sheng, Kasey Vickers, Shilin Zhao, Jing Wang, David C Samuels, Olivia Koues, Yu Shyr, and Yan Guo. Multi-perspective quality control of illumina RNA sequencing data analysis. *Brief. Funct. Genomics*, 16(4):194–204, 2017.

37. Shawn C Baker, Steven R Bauer, Richard P Beyer, James D Brenton, Bud Bromley, John Burrill, Helen Causton, Michael P Conley, Rosalie Elespuru, Michael Fero, Carole Foy, James Fuscoe, Xiaolian Gao, David Lee Gerhold, Patrick Gilles, Federico Goodsaid, Xu Guo, Joe Hackett, Richard D Hockett, Pranvera Ikonomi, Rafael A Irizarry, Ernest S Kawasaki, Tamma Kaysser-Kranich, Kathleen Kerr, Gretchen

Kiser, Walter H Koch, Kathy Y Lee, Chunmei Liu, Z Lewis Liu, Anne Lucas, Chitra F Manohar, Garry Miyada, Zora Modrusan, Helen Parkes, Raj K Puri, Laura Reid, Thomas B Ryder, Marc Salit, Raymond R Samaha, Uwe Scherf, Timothy J Sendera, Robert A Setterquist, Leming Shi, Richard Shippy, Jesus V Soriano, Elizabeth A Wagar, Janet A Warrington, Mickey Williams, Frederike Wilmer, Mike Wilson, Paul K Wolber, Xiaoning Wu, Renata Zadro, and External RNA Controls Consortium. The external RNA controls consortium: a progress report. *Nat. Methods*, 2(10):731–734, 2005.

38. Lichun Jiang, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.*, 21(9):1543–1551, 2011.

39. 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, 2014.

40. Patrick Deelen, Daria V Zhernakova, Mark de Haan, Marijke van der Sijde, Marc Jan Bonder, Juha Karjalainen, K Joeri van der Velde, Kristin M Abbott, Jingyuan Fu, Cisca Wijmenga, Richard J Sinke, Morris A Swertz, and Lude Franke. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.*, 7(1):30, 2015.

41. Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511–515, 2010.

42. Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, 2005.

43. Wei Zhao, Peter Langfelder, Tova Fuller, Jun Dong, Ai Li, and Steve Hovarth. Weighted gene coexpression network analysis: State of the art. *J. Biopharm. Stat.*, 20(2):281–300, 2010.

44. Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, 6(5):377–382, 2009.

45. Daniel E Wagner, Caleb Weinreb, Zach M Collins, James A Briggs, Sean G Megason, and Allon M Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392): 981–987, 2018.

46. Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten

Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. The human cell atlas. *Elife*, 6, 2017.

47. Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell RNA-seq in the last decade. 2017.

48. Valentine Svensson and Eduardo da Veiga Beltrame. A curated database reveals trends in single cell transcriptomics. 2019.

49. Jean-Francois Poulin, Bosiljka Tasic, Jens Hjerling-Leffler, Jeffrey M Trimarchi, and Rajeshwar Awatramani. Disentangling neural cell diversity using single-cell transcriptomics. *Nat. Neurosci.*, 19(9):1131–1141, 2016.

50. Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S Genshaft, Travis K Hughes, Carly G K Ziegler, Samuel W Kazer, Aleth Gaillard, Kellie E Kolb, Alexandra-Chloé Villani, Cory M Johannessen, Aleksandr Y Andreev, Eliezer M Van Allen, Monica Bertagnolli, Peter K Sorger, Ryan J Sullivan, Keith T Flaherty, Dennie T Frederick, Judit Jané-Valbuena, Charles H Yoon, Orit Rozenblatt-Rosen, Alex K Shalek, Aviv Regev, and Levi A Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, 2016.

51. Christoph Ziegenhain, Beate Vieth, Swati Parekh, Ines Hellmann, and Wolfgang Enard. Quantitative single-cell transcriptomics. *Brief. Funct. Genomics*, 17(4):220–232, 2018.

52. Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, Laura Jardine, David Dixon, Emily Stephenson, Emil Nilsson, Ida Grundberg, David McDonald, Andrew Filby, Weibo Li, Philip L De Jager, Orit Rozenblatt-Rosen, Andrew A Lane, Muzlifah Haniffa, Aviv Regev, and Nir Hacohen. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335), 2017.

53. Kishan Dholakia, Michael P MacDonald, Pavel Zemánek, and Tomás Cizmár. Cellular and colloidal separation using optical forces. *Methods Cell Biol.*, 82:467–495, 2007.

54. Robert Durruthy-Durruthy and Manisha Ray. Using fluidigm C1 to generate Single-Cell Full-Length cDNA libraries for mRNA sequencing. *Methods Mol. Biol.*, 1706:199–221, 2018.

55. Angela R Wu, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke, and Stephen R Quake. Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, 11(1):41–46, 2014.

56. Yurong Xin, Jinrang Kim, Min Ni, Yi Wei, Haruka Okamoto, Joseph Lee, Christina Adler, Katie Cavino, Andrew J Murphy, George D Yancopoulos, Hsin Chieh Lin, and Jesper Gromada. Use of the fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. *Proc. Natl. Acad. Sci. U. S. A.*, 113 (12):3293–3298, 2016.

57. Robert Salomon, Dominik Kaczorowski, Fatima Valdes-Mora, Robert E Nordon, Adrian Neild, Nona Farbehi, Nenad Bartonicek, and David Gallego-Ortega. Droplet-based single cell RNAseq tools: a practical guide. *Lab Chip*, 19(10):1706–1727, May 2019. ISSN 1473-0197, 1473-0189. doi: 10.1039/c8lc01239c.

58. Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

59. Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

60. Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.*, 12(1):44–73, 2017.

61. Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, 2017.

62. Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, and Marcus G Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, 10(11):1093–1095, 2013.

63. Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat. Methods*, 11(6):637–640, 2014.

64. Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, Yuval Dor, Aviv Regev, and Itai Yanai. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, 17:77, 2016.

65. 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*, page 003236, 2014.

66. Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, 9(1): 72–74, 2011.

67. Katsuyuki Shiroguchi, Tony Z Jia, Peter A Sims, and X Sunney Xie. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences*, 109(4):1347–1352, 2012.

68. Belinda Phipson, Luke Zappia, and Alicia Oshlack. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res.*, 6:595, 2017.

69. Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of Single-Cell RNA sequencing. *Mol. Cell*, 58(4):610–620, 2015.

70. Viktor Petukhov, Jimin Guo, Ninib Baryawno, Nicolas Severe, David T Scadden, Maria G Samsonova, and Peter V Kharchenko. dropest: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biol.*, 19(1):78, 2018.

71. Atefeh Lafzi, Catia Moutinho, Simone Picelli, and Holger Heyn. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.*, 13(12):2742–2757, 2018.

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

73. Giacomo Baruzzo, Katharina E Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A FitzGerald, and Gregory R Grant. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods*, 14(2):135–139, 2017.

74. Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527, 2016.

75. Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4):417–419, 2017.

76. Alexander Kanitz, Foivos Gypas, Andreas J Gruber, Andreas R Gruber, Georges Martin, and Mihaela Zavolan. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.*, 16:150, 2015.

77. Mingxiang Teng, Michael I Love, Carrie A Davis, Sarah Djebali, Alexander Dobin, Brenton R Graveley, Sheng Li, Christopher E Mason, Sara Olson, Dmitri Pervouchine, Cricket A Sloan, Xintao Wei, Lijun Zhan, and Rafael A Irizarry. A benchmark for RNA-seq quantification pipelines. *Genome Biol.*, 17:74, 2016.

78. Vivien Marx. How to deduplicate PCR. *Nat. Methods*, 14(5):473–476, April 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4268.

79. Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.*, 27(3):491–499, 2017.

80. Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11 (2):163–166, 2014.

81. Avi Srivastava, Laraib Malik, Tom Smith, Ian Sudbery, and Rob Patro. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol.*, 20(1):65, 2019.

82. Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, 17:63, 2016.

83. F Han and S J Lillard. In-situ sampling and separation of RNA from individual mammalian cells. *Anal. Chem.*, 72(17):4073–4079, 2000.

84. Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. Single-cell mRNA quantification and differential analysis with census. *Nat. Methods*, 14(3):309–315, 2017.

85. Johannes W Bagnoli, Christoph Ziegenhain, Aleksandar Janjic, Lucas E Wange, Beate Vieth, Swati Parekh, Johanna Geuder, Ines Hellmann, and Wolfgang Enard. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.*, 9(1):2937, 2018.

86. Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, 34(11):1145–1160, 2016.

87. Geng Chen, Baitang Ning, and Tieliu Shi. Single-Cell RNA-Seq technologies and related computational data analysis. *Front. Genet.*, 10:317, 2019.

88. Serena Liu and Cole Trapnell. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res.*, 5, 2016.

89. Sanjay M Prakadan, Alex K Shalek, and David A Weitz. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.*, 18(6):345–361, 2017.

90. Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, 16(3):133–145, 2015.

91. Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018.

92. Christopher S McGinnis, Lyndsay M Murrow, and Zev J Gartner. DoubletFinder: Doublet detection in Single-Cell RNA sequencing data using artificial nearest neighbors. *Cell Syst*, 8(4):329–337.e4, 2019.

93. Samuel L Wolock, Romain Lopez, and Allon M Klein. Scrublet: Computational identification of cell doublets in Single-Cell transcriptomic data. *Cell Syst*, 8(4):281–291.e9, 2019.

94. Aaron T L Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree, and John C Marioni. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.*, 20(1):63, 2019.

95. Davis J McCarthy, Kieran R Campbell, Aaron T L Lun, and Quin F Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, 2017.

96. Lan Jiang, Huidong Chen, Luca Pinello, and Guo-Cheng Yuan. GiniClust: detecting rare cell types from single-cell gene expression data with gini index. *Genome Biol.*, 17(1):144, 2016.

97. Tallulah S Andrews and Martin Hemberg. M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics*, 35(16):2865–2867, 2019.

98. Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.*, 19(5):776–792, 2018.

99. Aaron T L Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17:75, 2016.

100. Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Duoptdoit, and John C Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods*, 14(6):565–571, 2017.

101. Georgi K Marinov, Brian A Williams, Ken McCue, Gary P Schroth, Jason Gertz, Richard M Myers, and Barbara J Wold. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.*, 24(3):496–510, 2014.

102. Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Brain structure. cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015.

103. Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, 11(7):740–742, 2014.

104. Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11(3):R25, 2010.

105. Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106, 2010.

106. Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods*, 14(6):584–586, 2017.

107. Sandhya Prabhakaran, Elham Azizi, Ambrose Carr, and Dana Pe'er. Dirichlet process mixture model for correcting technical variation in Single-Cell gene expression data. In *International Conference on Machine Learning*, pages 1070–1079, 2016.

108. Bo Ding, Lina Zheng, Yun Zhu, Nan Li, Haiyang Jia, Rizi Ai, Andre Wildberg, and Wei Wang. Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics*, 31(13):2225–2227, 2015.

109. Shintaro Katayama, Virpi Töhönen, Sten Linnarsson, and Juha Kere. SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics*, 29(22):2943–2945, 2013.

110. Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Tomislav Ilicic, Sarah A Teichmann, and John C Marioni. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.*, 6:8687, 2015.

111. Davide Risso, John Ngai, Terence P Speed, and Sandrine Duoptdoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, 32(9):896–902, 2014.

112. Catalina A Vallejos, John C Marioni, and Sylvia Richardson. BASiCS: Bayesian analysis of Single-Cell sequencing data. *PLoS Comput. Biol.*, 11(6):e1004333, 2015.

113. Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.*, 7:39921, 2017.

114. Aaron T L Lun, Fernando J Calero-Nieto, Liora Haim-Vilmovsky, Berthold Göttgens, and John C Marioni. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.*, 27(11):1795–1806, 2017.

115. Elisabetta Mereu, Atefeh Lafzi, Catia Moutinho, Christoph Ziegenhain, Davis J MacCarthy, Adrian Alvarez, Eduard Batlle, Sagar, Dominic Grün, Julia K Lau, Stéphane C Boutet, Chad Sanada, Aik Ooi, Robert C Jones, Kelly Kaihara, Chris Brampton, Yasha Talaga, Yohei Sasagawa, Kaori Tanaka, Tetsutaro Hayashi, Itoshi Nikaido, Cornelius Fischer, Sascha Sauer, Timo Trefzer, Christian Conrad, Xian Adiconis, Lan T Nguyen, Aviv Regev, Joshua Z Levin, Swati Parekh, Aleksandar Janjic, Lucas E Wange, Johannes W Bagnoli, Wolfgang Enard, Marta Gut, Rickard Sandberg, Ivo Gut, Oliver Stegle, and Holger Heyn. Benchmarking Single-Cell RNA sequencing protocols for cell atlas projects. 2019.

116. Andrew S Venteicher, Itay Tirosh, Christine Hebert, Keren Yizhak, Cyril Neftel, Mariella G Filbin, Volker Hovestadt, Leah E Escalante, Mckenzie L Shaw, Christopher Rodman, Shawn M Gillespie, Danielle Dionne, Christina C Luo, Hiranmayi Ravichandran, Ravindra Mylvaganam, Christopher Mount, Maristela L Onozato,

Brian V Nahed, Hiroaki Wakimoto, William T Curry, A John Iafrate, Miguel N Rivera, Matthew P Frosch, Todd R Golub, Priscilla K Brastianos, Gad Getz, Anoop P Patel, Michelle Monje, Daniel P Cahill, Orit Rozenblatt-Rosen, David N Louis, Bradley E Bernstein, Aviv Regev, and Mario L Suvà. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science*, 355(6332), March 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aai8478.

117. Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nat. Rev. Genet.*, 20(5):257–272, 2019.

118. W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

119. 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, 2015.

120. Jiarui Ding, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T Nguyen, John Y H Kwon, Boaz Barak, William Ge, Amanda J Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K Shalek, Alexandra-Chloé Villani, Aviv Regev, and Joshua Z Levin. Systematic comparative analysis of single cell RNA-sequencing methods. 2019.

121. Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, 36(5):421–427, 2018.

122. Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36(5):411–420, 2018.

123. Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nat. Biotechnol.*, 37(6):685–691, 2019.

124. Yingxin Lin, Shila Ghazanfar, Kevin Wang, Johann A Gagnon-Bartsch, Kitty K Lo, Xianbin Su, Ze-Guang Han, John T Ormerod, Terence P Speed, Pengyi Yang, and Jean Y H Yang. scmerge: Integration of multiple single-cell transcriptomics datasets leveraging stable expression and pseudo-replication. 2018.

125. Jong-Eun Park, Krzysztof Polański, Kerstin Meyer, and Sarah A Teichmann. Fast batch alignment of single cell transcriptomes unifies multiple mouse cell atlases into an integrated landscape. 2018.

126. Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods*, 16(1):43–49, 2019.

127. Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, 15(6):e8746, 2019.

128. Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Zimin Zhou, Haide Chen, Fang Ye, Daosheng Huang, Yang Xu, Wentao Huang, Mengmeng Jiang, Xinyi Jiang, Jie Mao, Yao Chen, Chenyu Lu, Jin Xie, Qun Fang, Yibin Wang, Rui Yue, Tiefeng Li, He Huang, Stuart H Orkin, Guo-Cheng Yuan, Ming Chen, and Guoji Guo. Mapping the mouse cell atlas by Microwell-Seq. *Cell*, 172(5):1091–1107.e17, 2018.

129. Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018.

130. Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, Andrew Adey, Robert H Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357 (6352):661–667, 2017.

131. Mireya Plass, Jordi Solana, F Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391):eaaq1723, 2018.

132. Luke Zappia, Belinda Phipson, and Alicia Oshlack. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.*, 14(6):e1006245, 2018.

133. Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, 14(5):483–486, 2017.

134. Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J T Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.*, 20(1):194, 2019.

135. Taiyun Kim, Irene Rui Chen, Yingxin Lin, Andy Yi-Yang Wang, Jean Yee Hwa Yang, and Pengyi Yang. Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief. Bioinform.*, 2018.

136. Luke Zappia and Alicia Oshlack. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience*, 7(7), 2018.

137. José Alquicira-Hernández, Anuja Sathe, Hanlee P Ji, Quan Nguyen, and Joseph E Powell. scpred: Cell type prediction at single-cell resolution. 2018.

138. Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods*, 15(5):359–362, 2018.

139. Rui Hou, Elena Denisenko, and Alistair R R Forrest. scmatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics*, 2019.

140. Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, Atul J Butte, and Mallar Bhattacharya. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.*, 20(2):163–172, 2019.

141. Robrecht Cannoodt, Wouter Saelens, and Yvan Saeys. Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.*, 46(11):2496–2506, 2016.

142. Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, 37(5):547–554, 2019.

143. Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

144. Gu Mi, Yanming Di, and Daniel W Schafer. Goodness-of-fit tests and model diagnostics for negative binomial regression of RNA sequencing data. *PLoS One*, 10(3):e0119254, 2015.

145. Anne Schwabe, Katja N Rybakova, and Frank J Bruggeman. Transcription stochasticity of complex gene regulation models. *Biophys. J.*, 103(6):1152–1161, 2012.

146. Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, 105(45):17256–17261, 2008.

147. Alessandra Dal Molin, Giacomo Baruzzo, and Barbara Di Camillo. Single-Cell RNA-Sequencing: Assessment of differential expression analysis methods. *Front. Genet.*, 8:62, 2017.

148. Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (D3E)–a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, 17:110, 2016.

149. Trung Nghia Vu, Quin F Wills, Krishna R Kalari, Nifang Niu, Liewei Wang, Mattias Rantalainen, and Yudi Pawitan. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*, 32(14):2128–2135, 2016.

150. Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, Peter S Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, 16(1):1–13, 2015.

151. Nils Eling, Arianne C Richard, Sylvia Richardson, John C Marioni, and Catalina A Vallejos. Correcting the Mean-Variance dependency for differential variability testing using Single-Cell RNA sequencing data. *Cell Syst*, 7(3):284–294.e12, 2018.

152. Jingshu Wang, Mo Huang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, John Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 115(28):E6437–E6446, 2018.

153. Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, 15(4):255–261, 2018.

154. Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, 33(21):3486–3488, 2017.

155. Beate Vieth, Swati Parekh, Christoph Ziegenhain, Wolfgang Enard, and Ines Hellmann. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.*, 10(1):4667, 2019.

156. Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, 16(1):241, 2015.

157. Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Duoptdoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, 9(1):284, 2018.

158. Zhun Miao, Ke Deng, Xiaowo Wang, and Xuegong Zhang. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, 34(18):3223–3224, 2018.

159. Koen Van den Berge, Fanny Perraudeau, Charlotte Soneson, Michael I Love, Davide Risso, Jean-Philippe Vert, Mark D Robinson, Sandrine Duoptdoit, and Lieven Clement. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.*, 19(1):24, 2018.

160. Lisa Amrhein, Kumar Harsha, and Christiane Fuchs. A mechanistic model for the negative binomial distribution of single-cell mRNA counts. 2019.

161. Valentine Svensson. Droplet scRNA-seq is not zero-inflated. 2019.

162. F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single cell RNA-Seq based on a multinomial model. 2019.

163. Lihua Zhang and Shihua Zhang. Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2018.

164. Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods*, 15(7):539–542, 2018.

165. Wenhao Tang, Francois Bertaux, Philipp Thomas, Claire Stefanelli, Malika Saint, Samuel Blaise Marguerat, and Vahid Shahrezaei. baynorm: Bayesian gene expression recovery, imputation and normalisation for single cell RNA-sequencing data. 2018.

166. Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell RNA-seq data. *Nat. Commun.*, 9(1):997, 2018.

167. David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Recovering gene interactions from Single-Cell data using data diffusion. *Cell*, 174(3):716–729.e27, 2018.

168. Jonathan Ronen and Altuna Akalin. *netSmooth*: Network-smoothing based imputation for single cell RNA-seq. *F1000Res.*, 7:8, 2018.

169. Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J Garry. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, 19(1):220, 2018.

170. Tallulah S Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Res.*, 7: 1740, 2018.

171. Jingshu Wang, Divyansh Agarwal, Mo Huang, Gang Hu, Zilu Zhou, Chengzhong Ye, and Nancy R Zhang. Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods*, 16(9):875–878, 2019.

172. Lingxue Zhu, Jing Lei, Bernie Devlin, and Kathryn Roeder. A UNIFIED STATISTICAL FRAMEWORK FOR SINGLE CELL AND BULK RNA SEQUENCING DATA. *Ann. Appl. Stat.*, 12(1):609–632, 2018.

173. Elham Azizi, Sandhya Prabhakaran, Ambrose Carr, and Dana Pe'er. Bayesian inference for single-cell clustering and imputing. *Genomics and Computational Biology*, 3(1):46, 2017.

174. Ronald A Fisher. *The Design of Experiments*. Macmillan Pub Co, 9 edition edition, 1971. ISBN 9780028446905.

175. Gary A Churchill. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.*, 32 Suppl: 490–495, 2002.

176. Paul L Auer and R W Doerge. Statistical design and analysis of RNA sequencing data. *Genetics*, 185(2): 405–416, 2010.

177. Andrea Sboner, Xinmeng Jasmine Mu, Dov Greenbaum, Raymond K Auerbach, and Mark B Gerstein. The real cost of sequencing: higher than you think! *Genome Biol.*, 12(8):125, 2011.

178. Yuwen Liu, Jie Zhou, and Kevin P White. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–304, 2014.

179. Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. 2019.

180. Douglas C Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, 2017. ISBN 9781119113478.

181. Jonathan Alles, Nikos Karaiskos, Samantha D Praktiknjo, Stefanie Grosswendt, Philipp Wahle, Pierre-Louis Ruffault, Salah Ayoub, Luisa Schreyer, Anastasiya Boltengagen, Carmen Birchmeier, Robert Zinzen, Christine Kocks, and Nikolaus Rajewsky. Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.*, 15(1):44, 2017.

182. Amy Guillaumet-Adkins, Gustavo Rodríguez-Esteban, Elisabetta Mereu, Maria Mendez-Lago, Diego A Jaitin, Alberto Villanueva, August Vidal, Alex Martinez-Marti, Enriqueta Felip, Ana Vivancos, Hadas Keren-Shaul, Simon Heath, Marta Gut, Ido Amit, Ivo Gut, and Holger Heyn. Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol.*, 18(1):45, 2017.

183. Jase Gehring, Jong Hwee Park, Sisi Chen, Matthew Thomson, and Lior Pachter. Highly multiplexed Single-Cell RNA-seq for defining cell population and transcriptional spaces. 2018.

184. Christopher S McGinnis, David M Patterson, Juliane Winkler, Daniel N Conrad, Marco Y Hein, Vasudha Srivastava, Jennifer L Hu, Lyndsay M Murrow, Jonathan S Weissman, Zena Werb, Eric D Chow, and Zev J Gartner. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods*, 16(7):619–626, 2019.

185. Dongju Shin, Wookjae Lee, Ji Hyun Lee, and Duhee Bang. Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. *Sci Adv*, 5(5):eaav2249, 2019.

186. Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nat. Protoc.*, 7(5): 813–828, 2012.

187. Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, 11(10):733–739, 2010.

188. Jeanette Baran-Gale, Tamir Chandra, and Kristina Kirschner. Experimental design for single-cell RNA sequencing. *Brief. Funct. Genomics*, 17(4):233–239, 2018.

189. Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, 14(9):R95, 2013.

190. Luyi Tian, Xueyi Dong, Saskia Freytag, Kim-Anh Lê Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S Weber, Azadeh Seidi, Jafar S Jabbari, Shalin H Naik, and Matthew E Ritchie. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods*, 16 (6):479–487, 2019.

191. Alyssa C Frazee, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, page btv272, 2015.

192. Charlotte Soneson. compcoder–an R package for benchmarking differential expression methods for RNA-seq data. *Bioinformatics*, 30(17):2517–2518, 2014.

193. Sam Benidt and Dan Nettleton. SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. *Bioinformatics*, 31(13):2131–2140, 2015.

194. Guillem Rigaill, Sandrine Balzergue, Véronique Brunaud, Eddy Blondet, Andrea Rau, Odile Rogier, José Caius, Cathy Maugis-Rabusseau, Ludivine Soubigou-Taconnat, Sébastien Aubourg, Claire Lurin, Marie-Laure Martin-Magniette, and Etienne Delannoy. Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis. *Brief. Bioinform.*, 19(1):65–76, 2018.

195. Claire R Williams, Alyssa Baccarella, Jay Z Parrish, and Charles C Kim. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*, 18 (1):38, 2017.

196. Susan Holmes and Wolfgang Huber. *Modern Statistics for Modern Biology*. Cambridge University Press, 2018. ISBN 9781108427029.

197. Youngchao Ge, Sandrine Duoptdoit, and Terence P Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003.

198. Sandrine Duoptdoit, Juliet Popper Shaffer, and Jennifer C Boldrick. Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, 18(1):71–103, 2003.

199. John D Storey. A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.*, 64(3): 479–498, 2002.

200. Kouros Owzar, William T Barry, and Sin-Ho Jung. Statistical considerations for analysis of microarray experiments. *Clin. Transl. Sci.*, 4(6):466–477, 2011.

201. Stanley B Pounds. Estimation and control of multiple testing error rates for microarray studies. *Brief. Bioinform.*, 7(1):25–36, 2006.

202. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300, 1995.

203. P H Westfall, S S Young, and S Paul Wright. On adjusting P-Values for multiplicity. *Biometrics*, 49(3): 941–945, 1993.

204. Sin-Ho Jung, Heejung Bang, and Stanley Young. Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, 6(1):157–169, January 2005. ISSN 1465-4644. doi: 10.1093/biostatistics/kxh026.

205. Alicia Poplawski and Harald Binder. Feasibility of sample size calculation for RNA-seq studies. *Brief. Bioinform.*, 19(4):713–720, 2018.

206. Hao Wu, Chi Wang, and Zhijin Wu. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics*, 31(2):233–241, 2015.

207. Michele A Busby, Chip Stewart, Chase A Miller, Krzysztof R Grzeda, and Gabor T Marth. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics*, 29(5): 656–657, 2013.

208. Alice Carter, Kate Tilling, and Marcus R Munafò. A systematic review of sample size and power in leading neuroscience journals. 2017.

209. Marcia McNutt. Journals unite for reproducibility. *Science*, 346(6210):679, 2014.

210. 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, 2017.

211. Charlotte Soneson and Mark D Robinson. iCOBRA: open, reproducible, standardized and live method benchmarking. *Nat. Methods*, 13(4):283, 2016.

212. Illumina | Bio-Rad. `https://info.bio-rad.com/ww-ddseq.html?WT.mc_id=170714020573&WT.srch=1&WT.knsh_id=0b4827f0-7e05-4a40-97b4-312e9824e32f&gclid=CjwKCAjw_MnmBRAoEiwAPRRWW2gSrNDNuEwU2JTU14kwosl0qbTod7wr5GH8cPHQl8iPrnQHOuSj_BoC4A8QAvD_BwE`. Accessed: 2019-5-8.

213. Pavithra Kumar, Yuqi Tan, and Patrick Cahan. Understanding development and stem cells using single cell-based analyses of gene expression. *Development*, 144(1):17–32, 2017.

214. Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, 14(9):618–630, 2013.

215. Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome Res.*, 25(10):1491–1498, 2015.

216. Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, 33(2):155–160, 2015.

217. Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, Zhengyan Kan, Wonshik Han, and Woong-Yang Park. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.*, 8:15081, 2017.

218. Orit Rozenblatt-Rosen, Michael J T Stubbington, Aviv Regev, and Sarah A Teichmann. The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453, 2017.

219. James A Briggs, Caleb Weinreb, Daniel E Wagner, Sean Megason, Leonid Peshkin, Marc W Kirschner, and Allon M Klein. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, 360(6392):eaar5780, 2018.

220. Hannah Hochgerner, Amit Zeisel, Peter Lönnerberg, and Sten Linnarsson. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat. Neurosci.*, 21(2):290–299, 2018.

221. Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, David N Louis, Orit Rozenblatt-Rosen, Mario L Suvà, Aviv Regev, and Bradley E Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

222. Charlotte Soneson and Mark D Robinson. Towards unified quality verification of synthetic count data with countsimQC. *Bioinformatics*, 34(4):691–692, 2018.

223. 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, 2016.

224. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

225. Picard toolkit. `http://broadinstitute.github.io/picard/`, 2019.

226. Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter A C 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

227. Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, 9:171, 2014.

228. Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, and Ido Amit. Massively parallel Single-Cell RNA-Seq for Marker-Free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

229. Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods*, 14(4):381–387, 2017.

230. L Paul, P Kubala, G Horner, M Ante, I Hollaender, and others. SIRVs: Spike-In RNA variants as external isoform controls in RNA-sequencing. *bioRxiv*, 2016.

231. Wei Vivian Li and Jingyi Jessica Li. A statistical simulator scdesign for rational scRNA-seq experimental design. *Bioinformatics*, 35(14):i41–i50, 2019.

232. Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. zUMIs - a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience*, 7(6), 2018.

233. Dominic Grün and Alexander van Oudenaarden. Design and analysis of Single-Cell sequencing experiments. *Cell*, 163(4):799–810, 2015.

234. Naomi Habib, Yinqing Li, Matthias Heidenreich, Lukasz Swiech, Inbal Avraham-Davidi, John J Trombetta, Cynthia Hession, Feng Zhang, and Aviv Regev. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*, 353(6302):925–928, 2016.

235. Naomi Habib, Inbal Avraham-Davidi, Anindita Basu, Tyler Burks, Karthik Shekhar, Matan Hofree, Sourav R Choudhury, François Aguet, Ellen Gelfand, Kristin Ardlie, David A Weitz, Orit Rozenblatt-Rosen, Feng Zhang, and Aviv Regev. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods*, 14(10): 955–958, 2017.

236. Benjamin Lacar, Sara B Linker, Baptiste N Jaeger, Suguna Krishnaswami, Jerika Barron, Martijn Kelder, Sarah Parylak, Apuã Paquola, Pratap Venepally, Mark Novotny, Carolyn O'Connor, Conor Fitzpatrick, Jennifer Erwin, Jonathan Y Hsu, David Husband, Michael J McConnell, Roger Lasken, and Fred H Gage. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.*, 7:11022, 2016.

237. Blue B Lake, Rizi Ai, Gwendolyn E Kaeser, Neeraj S Salathia, Yun C Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Mostafa Ronaghi, Jian-Bing Fan, Wei Wang, Jerold Chun, and Kun Zhang. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, 352(6293): 1586–1590, 2016.

238. Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E Borm, Zehua

Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, 2018.

239. Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, The RGASP Consortium, Gunnar Rätsch, Nick Goldman, Tim J Hubbard, Jennifer Harrow, Roderic Guigó, and Paul Bertone. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, 10(12):1185–1191, 2013.

240. Michael B Cole, Davide Risso, Allon Wagner, David DeTomaso, John Ngai, Elizabeth Purdom, Sandrine Duoptdoit, and Nir Yosef. Performance assessment and selection of normalization procedures for Single-Cell RNA-Seq. *Cell Syst*, 8(4):315–328.e8, 2019.

241. Shun H Yip, Panwen Wang, Jean-Pierre A Kocher, Pak Chung Sham, and Junwen Wang. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.*, 45(22):e179, 2017.

242. Maria K Jaakkola, Fatemeh Seyednasrollah, Arfa Mehmood, and Laura L Elo. Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief. Bioinform.*, 18(5):735–743, 2017.

243. Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, 15(2):R29, 2014.

244. Todd M Gierahn, Marc H Wadsworth, 2nd, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love, and Alex K Shalek. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods*, 14(4):395–398, 2017.

245. Celsee systems. `https://www.celsee.com/systems/`. Accessed: 2019-10-2.

246. ICELL8 system. `https://www.takarabio.com/learning-centers/automation-systems/icell8-introduction`. Accessed: 2019-10-2.

247. Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job van der Zwan, Martin Häring, Emelie Braun, Lars E Borm, Gioele La Manno, Simone Codeluppi, Alessandro Furlan, Kawai Lee, Nathan Skene, Kenneth D Harris, Jens Hjerling-Leffler, Ernest Arenas, Patrik Ernfors, Ulrika Marklund, and Sten Linnarsson. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014.e22, 2018.

248. Páll Melsted, A Sina Booeshaghi, Fan Gao, Eduardo Beltrame, Lambda Lu, Kristján Eldjárn Hjorleifsson, Jase Gehring, and Lior Pachter. Modular and efficient pre-processing of single-cell RNA-seq. 2019.

249. 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, 2012.

250. S Parekh, B Vieth, C Ziegenhain, W Enard, and I Hellmann. Strategies for quantitative RNA-seq analyses among closely related species. *bioRxiv*, 2018.

251. Daniel P Depledge, Kalanghad Puthankalam Srinivas, Tomohiko Sadaoka, Devin Bready, Yasuko Mori, Dimitris G Placantonakis, Ian Mohr, and Angus C Wilson. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat. Commun.*, 10(1):754, 2019.

252. Charlotte Soneson, Yao Yao, Anna Bratus-Neuenschwander, Andrea Patrignani, Mark D Robinson, and Shobbir Hussain. A comprehensive examination of nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.*, 10(1):3359, 2019.

253. Ashley Byrne, Charles Cole, Roger Volden, and Christopher Vollmers. Realizing the potential of full-length transcriptome sequencing. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 374(1786):20190097, 2019.

254. David Laehnemann, Johannes Köster, Ewa Szcureck, Davis McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Niko Beerenwinkel, Kieran R Campbell, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharin Jahn, Thamar Jessurun Lobo, Emma M Keizer, Indu Khatri, Szymon M Kiełbasa, Jan O Korbel, Alexey M Kozlov, Tzu-Hao Kuo, Boudewijn P F Lelieveldt, Ion I Manoptdoiu, John C Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Łukasz Rączkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J Theis, Huan Yang, Alex Zelikovsky, Alice C McHardy, Benjamin J Raphael, Sohrab P Shah, and Alexander Schönhuth. 12 grand challenges in single-cell data science. Technical Report e27885v1, PeerJ Preprints, 2019.

255. Youjin Hu, Qin An, Katherine Sheu, Brandon Trejo, Shuxin Fan, and Ying Guo. Single cell Multi-Omics technology: Methodology and application. *Front Cell Dev Biol*, 6:28, 2018.

256. M Colomé-Tatché and F J Theis. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, 7:54–59, 2018.

List of Figures

Acknowledgements

I got to meet so many interesting and great people during my PhD that I will for sure forget one or the other. So, my apologies in advance! With that out of the way, here we go:

First and foremost, I want to thank my PhD advisor Wolfgang Enard and my mentor Ines Hellmann. Both of you have created this incredibly awesome work group with such a positive atmosphere that I could not wish for more. Thank you so much for your constant support and advice. Wolfgang, I want to thank you for giving me so much freedom to explore my interests and Ines for helping me realize them. How you two think and conduct science has taught me what it really means to do science.

Of course, I have to thank the two other pillars of the holy trinity: Christoph for patiently answering all my questions concerning the mysteries of the wet lab and Swati for guiding me through all things related to bioinformatics. Both of you were the best in sharing ideas over coffee, group raging, hand holding while piping, eating yummy food and so much more. KK and Commander, without you two this would have not been so much fun. Naturally, thanks to everyone in the Enard lab: Lu, Johanna and Aleks for nice morning coffees; Daniel and Jojo for awesome company and beer; Philipp and Zane for your energy and infectious laughter; my fellow office mates Bria, Ilse and Mari for a great working atmosphere, even though the air is not sometimes; Karin, Ines B and Steffi for being the best support and lending a helping hand. Of course, all the great students, past and present, that I had the privilege to meet and teach on occasions: Michael, Volker, Gunnar, Alex, Lukas, Khalis, Chris, Zeynep, ... Thank you all for being great colleagues and friends.

I want to thank my family and friends: Papi, I want to thank you for your support, love and advice through all my studies and life in general. Tom, even though you are my only brother, you are the best big brother I could have wished for. Thanks for all your help, in and out of science and for kicking my butt to get stuff done. I owe a big thanks to mijn gekke meiden, Fleur, Iris and Nele, you guys have been the best. Last but not least, Rob, thank you for invading my home country just for me, loving me unconditionally and being my calm rock supporting me through all my struggles.

Curriculum Vitae

Beate Vieth

PhD Candidate

Education

Ludwig-Maximilians-University Munich / PhD
2013 - 2019, Munich, Germany
Advisor: Prof. Wolfgang Enard

Ghent Universiteit / Master of Science
2010 - 2012, Gent, Belgium
Applied Statistical Data Analysis

Katholieke Universiteit Leuven / Master of Science
2008 - 2010, Leuven, Belgium
Biology

Katholieke Universiteit Leuven / Bachelor of Science
2004 - 2008, Leuven, Belgium
Biology

Research Experience

Ludwig-Maximilians-University Munich / PhD research
2013 - 2019, Munich, Germany
Advisor: Prof. Wolfgang Enard
Anthropology and Human Genomics
Statistical Power Analysis for Single-Cell RNA-sequencing

Ghent Universiteit / Master thesis
2011 - 2012, Gent, Belgium
Advisor: Prof. Olivier Thas
Department of Data analysis and mathematical modelling
Sample size calculations and power analysis for microRNA arrays

Katholieke Universiteit Leuven / Master thesis

2009 - 2010, Leuven, Belgium

Advisor: Prof. Lutgarde Arckens
Laboratory of Neuroplasticity and Neuroproteomics
Characterisation of Cortical Plasticity in Mouse Visual Cortex after Monocular Enucleation during the Post-Critical Period

Swiss Federal Institute of Technology / Master studies

2008 - 2009, Zurich, Switzerland

Advisor: Prof. Nikolaus Amrhein
Exchange year; Focus on neurobiology and bioinformatics

Katholieke Universiteit Leuven / Bachelor thesis II

2008, Leuven, Belgium

Advisor: Prof. Jozef Vanden Broeck
Laboratory of Molecular Developmental Physiology and Signal Transduction
The Role of PSTI-homologues in proteolytic digestion of Locusta migratoria

Katholieke Universiteit Leuven / Bachelor thesis I

2008, Leuven, Belgium

Advisor: Prof. Robby Stoks
Laboratory of Aquatic Ecology and Evolutionary Biology
Effects of the Insecticide Endosulfan and Predation on the Fitness and Mortality of water boatmen (Corixidae)

Publications

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., & Hellmann, I. (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. Nature Communications, 10(1), 4667.

Schreiweis, C., Irinopoulou, T., **Vieth, B.**, … Groszer, M. (2019). Mice carrying a humanized Foxp2 knock-in allele show region-specific shifts of striatal Foxp2 expression levels. Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 118, 212–222.

Medvedeva, V. P., Rieger, M. A., **Vieth, B.**, … Groszer, M. (2019). Altered social behavior in mice carrying a cortical Foxp2 deletion. Human Molecular Genetics, 28(5), 701–717.

Ziegenhain, C.*, **Vieth, B.***, Parekh, S.*, Hellmann, I., & Enard, W. (2018). Quantitative single-cell transcriptomics. Briefings in Functional Genomics, 17(4), 220–232.

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

Parekh, S., **Vieth, B.**, Ziegenhain, C., Enard, W., & Hellmann, I. (2018). Strategies for quantitative RNA-seq analyses among closely related species. bioRxiv.

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

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

Ziegenhain, C., **Vieth, B.**, Parekh, S., … Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. Molecular Cell, 65(4), 631–643.e4.

Schreck, C., Istvánffy, R., Ziegenhain, C., … Oostendorp, R. A. J. (2017). Niche WNT5A regulates the actin cytoskeleton during regeneration of hematopoietic stem cells. The Journal of Experimental Medicine, 214(1), 165–181.

Parekh, S., Ziegenhain, C., **Vieth, B.**, Enard, W., & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. Scientific Reports, 6, 25533.

Wunderlich, S., Kircher, M., **Vieth, B.**, … Enard, W. (2014). Primate iPS cells as tools for evolutionary analyses. Stem Cell Research, 12(3), 622–629.

Presentations

Single Cell Genomics / Poster
Sep 2019, Stockholm, Sweden

A Systematic Evaluation of Single-Cell RNA-seq Pipelines

Single Cells: Technology to Biology / Poster
Feb 2019, Singapore

Systematic Evaluation of Single-Cell RNA-Sequencing Workflows

SFB1243: Cancer Evolution / Talk + Poster
Mar 2018, Munich, Germany

Statistical Power Analysis for Single-Cell RNA-Sequencing

SMBE / Poster
July 2015, Vienna, Austria

Evolution of gene expression patterns in primate stem cells

Teaching

Human Biology I / Bachelor
2014-2018, Munich, Germany

2 weeks lab course

Thesis Advisor / Master
2014 - 2017, Munich, Germany

Supervision and mentoring of several students for their MSc thesis