Improving & Applying

Single-Cell RNA Sequencing

Christoph Ziegenhain

München 2017

1. Gutachter: Prof. Wolfgang Enard

2. Gutachter: Prof. Heinrich Leonhardt

Tag der Abgabe: 12.10.2017

Tag der mündlichen Prüfung: 13.12.2017

Statutory declaration and statement

(Eidestattliche Versicherung und Erklärung)

Eidesstattliche Erklärung

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

München, den 12.10.2017 Christoph Ziegenhain

(Unterschrift)

Erklärung

Hiermit erkläre ich,

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

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

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

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

München, den 12.10.2017 Christoph Ziegenhain

(Unterschrift)

Table of Contents

Abbreviations

ALL	acute lymphoblastic leukemia
AML	acute myeloid leukemia
bp	basepairs
DNA	deoxyribonucleic acid
ERCC	external RNA controls consortium
EST	expressed sequence tags
FACS	fluorescence-activated cell sorting
FISH	fluorescence in-situ hybridization
IVT	in-vitro transcription
mESC	mouse embryonic stem cell
MRD	minimal residual disease
nt	nucleotides
PCR	polymerase chain reaction
PDX	patient-derived xenograft
qPCR	quantitative polymerase chain reaction
RNA	ribonucleic acid
RT	reverse transcription
SAGE	serial analysis of gene expression
SBS	sequencing by synthesis
scRNA-seq	single-cell RNA sequencing
TF	transcription factor

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 (May): 25533. doi:10.1038/srep25533

II. Ebinger S*, Özdemir EZ*, **Ziegenhain C***, Tiedt S*, Alves CC*, Grunert M, Dworzak M, Lutz C, Horny HP, Sotlar K, Parekh S, Spiekermann K, Hiddemann W, Schepers A, Polzer B, Kirsch S, Hoffmann M, Knapp B, Hasenauer J, Pfeifer H, Panzer-Grümayer R, Enard W, Gires O, Jeremias I:
"Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia." (2016)
Cancer Cell, November. doi:10.1016/j.ccell.2016.11.002.

III. **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): 631–43.e4. doi:10.1016/j.molcel.2017.01.023

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

V. Parekh S*, **Ziegenhain C***, Vieth B, Enard W, Hellmann I:
"zUMIs: A fast and flexible pipeline to process RNA sequencing data with UMIs." (2017)
BioRxiv. doi:10.1101/153940

VI. Bagnoli JW*, **Ziegenhain C***, Janjic A*, Wange LE, Vieth B, Parekh S, Geuder J, Hellmann I, Enard W:
"mcSCRB-seq: sensitive and powerful single-cell RNA sequencing." (2017)
unsubmitted Manuscript

Other publications (not included in the thesis):

VII. Schneider M, Tigges B, Meggendorfer M, Helfer M, **Ziegenhain C**, Brack-Werner R:
"A New Model for Post-Integration Latency in Macroglial Cells to Study HIV-1 Reservoirs of the Brain." (2015)
AIDS 29 (10): 1147–59. doi: 10.1097/QAD.0000000000000691

VIII. Schreck C*, Istvanffy R*, **Ziegenhain C**, Sippenauer T, Ruf F, Henkel L, Gärtner F, Vieth B, Florian MC, Mende N, Taubenberger A, Prendergast A, Wagner A, Pagel C, Grziwok S, Götze KS, Guck J, Dean DC, Massberg S, Essers M, Waskow C, Geiger H, Schiemann M, Peschel C, Enard W, Oostendorp RAJ:
"Niche WNT5A Regulates the Actin Cytoskeleton during Regeneration of Hematopoietic Stem Cells." (2016)
The Journal of Experimental Medicine, December. doi:10.1016/jem.20151414.

IX. Witzel M, Petersheim D, Fan Y, Bahrami E, Racek T, Rohlfs M, Puchalka J, Mertes C, Gagneur J, **Ziegenhain C**, Enard W, Stray-Pederson A, Arkwright PD, Abboud MR, Pazhakh V, Lieschke GJ, Mundlos S, Krawitz PM, Dahlhoff M, Schneider MR, Wolf E, Horny HP, Schmidt H, Schäffer AA, Klein C:
"Chromatin-Remodeling Factor SMARCD2 Regulates Transcriptional Networks Controlling Differentiation of Neutrophil Granulocytes." (2017)
Nature Genetics, April. doi:10.1038/ng.3833.

X. Krendl C*, Shaposhnikov D*, Rishko V, Ori C, **Ziegenhain C**, Sass S, Simon L, Müller NS, Straub T, Brooks KE, Chavez SL, Enard W, Theis FJ, Drukker M:
"GATA2/3-TFAP2A/C transcription factor network couples human ES cell differentiation to trophectoderm with repression of pluripotency." (2017)
Proceedings of the National Academy of Sciences.

Declaration 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 me. I prepared the RNA-seq libraries used in the publication and helped in data processing. Swati Parekh, Ines Hellmann and Beate Vieth analyzed the data and performed power simulations. The manuscript was written by Ines Hellmann, Swati Parekh and Wolfgang Enard.

Comparative Analysis of Single-Cell RNA Sequencing Methods

I had the idea to this publication and planned the experiments with Wolfgang Enard. the single-cell RNA-seq protocols were established by me in our lab and I performed the library preparations for Smart-seq/C1, SCRB-seq, Drop-seq and CEL-seq2/C1. I processed all data with help from Swati Parekh. Data for all methods were analyzed by me and the power simulation framework was developed by Beate Vieth. Ines Hellmann guided computational work. The manuscript was written by Wolfgang Enard and me with valuable input from Ines Hellmann and Björn Reinius.

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

Beate Vieth and Ines Hellmann conceived the study after working on power simulations for “Comparative Analysis of Single-Cell RNA Sequencing Methods”. Beate Vieth developed and programmed *powsimR*. I tested the program and evaluated its performance relative to empirical scRNA-seq data. Beate Vieth, Ines Hellmann and Wolfgang Enard wrote the manuscript.

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

Swati Parekh and I had the idea to this work, designed and implemented the pipeline. Beate Vieth tested code and performed power simulations to evaluate intron mappings. Swati Parekh, Wolfgang Enard, Ines Hellmann and I wrote the manuscript.

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

_____

Swati Parekh

_____

Christoph Ziegenhain

mcSCRB-seq: sensitive and powerful single-cell RNA sequencing

Wolfgang Enard and I conceived the study as a conclusion of the results of *Comparative Analysis of Single-Cell RNA Sequencing Methods*. Optimization experiments and sequencing library preparations were done by Johannes Bagnoli, Aleksandar Janjic, Lucas Wange and me. Sequencing data was processed by Swati Parekh and me. Johannes Bagnoli, Aleksandar Janjic, Beate Vieth and I analyzed the data. Johannes Bagnoli, Aleksandar Janjic, Wolfgang Enard and I wrote the manuscript.

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

_____

Johannes W. Bagnoli

_____

Christoph Ziegenhain

_____

Aleksandar Janjic

Characterization of Rare, Dormant, and Therapy-Resistant Cells

in Acute Lymphoblastic Leukemia

This study was conceived and supervised by Irmela Jeremias. Sebastian Tiedt and Catarina Castro Alves established the mouse model for leukemia xenografts. Sarah Ebinger and Erbey Özdemir performed and analyzed the mouse work and experiments with ALL cells. I performed bulk and single-cell RNA sequencing library preparations from isolated cells. Sequencing data was processed and analysed by Swati Parekh and me with contributions from Erbey Özdemir. The manuscript was written by Irmela Jeremias with participation of Wolfgang Enard.

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

_____

Sarah Ebinger

_____

Erbey Ziya Özdemir

_____

Christoph Ziegenhain

_____

Sebastian Tiedt

_____

Catarina Castro Alves

Aims of this Work

Building on the emergence of several effective whole-transcriptome amplification techniques, single-cell RNA sequencing is a rapidly developing novel tool with transformative impact on biology (Regev et al. 2017). The aims of this work may be subdivided on questions related to improving the technology of single-cell sequencing and an exemplary application to a biomedical question.

First, it is unclear how much noise or bias whole-transcriptome amplification introduces into quantitative gene expression profiles measured by RNA-seq and whether these amplification artifacts may be identified and removed computationally. For this aim, low-input RNA-seq datasets were generated and analysed from several protocols, including one with unique molecular identifiers (UMIs). Further investigating RNA-seq technology, the second aim was to establish and compare various newly developed single-cell RNA sequencing methods. To this end, four methods (CEL-seq2/C1, Drop-seq, SCRB-seq and Smart-seq/C1) were established, data from mouse ES cells generated and together with data generated by collaborators with two additional methods (MARS-seq, Smart-seq2) assessed for their sensitivity, accuracy, precision, power to detect differential gene expression and cost-efficiency.

The computational frameworks for power simulations and data processing that emerged during this study were extended and wrapped up as the applications *powsimR* and *zUMIs*. Based on the strengths and weaknesses of the benchmarked scRNA-seq methods, we developed a the highly sensitive, powerful and cost-efficient *mcSCRB-seq* protocol.

Lastly, applying single-cell transcriptomics to a biomedical question, the aim was to characterize minimal residual disease cells from acute lymphoblastic leukemia. In order to achieve this, this clinically important cell type was isolated from PDX mouse models and subjected to low-input bulk RNA-sequencing, single-cell RNA sequencing as well as functional characterization.

Summary

The cell is the fundamental building block of life. With the advent of single-cell RNA sequencing (scRNA-seq), we can for the first time assess the transcriptome of many individual cells. This has profound implications for biological and medical questions and is especially important to characterize heterogeneous cell populations and rare cells. However, the technology is technically and computationally challenging as complementary DNA (cDNA) needs to be generated and amplified from minute amounts of mRNA and sequenceable libraries need to be efficiently generated from many cells. This requires to establish different protocols, identify important caveats, benchmark various methods and improve them if possible. To this end, we analysed amplification bias and its effect on detecting differentially expressed genes in several bulk and a single-cell RNA sequencing methods. We found that correcting for amplification bias is not possible computationally but improves the power of scRNA-seq considerably, though neglectable for bulk-RNA-seq. In the second study we compared six prominent scRNA-seq protocols as more and more single-cell RNA-sequencing are becoming available, but an independent benchmark of methods is lacking. By using the same mouse embryonic stem cells (mESCs) and exogenous mRNA spike-ins as common reference, we compared six important scRNA-seq protocols in their sensitivity, accuracy and precision to quantify mRNA levels. In agreement with our previous study, we find that the precision, i.e. the technical variance, of scRNA-seq methods is driven by amplification bias and drastically reduced when using unique molecular identifiers to remove amplification duplicates. To assess the combined effects of sensitivity and precision and to compare the cost-efficiency of methods we compared the power to detect differentially expressed genes among the tested scRNA-seq protocols using a novel simulation framework. We find that some methods are prohibitively inefficient and others show trade-offs depending on the number of cells per sample that need to be analysed. Our study also provides a framework for

benchmarking further improvements of scRNA-seq protocol and we published an improved version of our simulation framework *powsimR*. It uniquely recapitulates the specific characteristics of scRNA-seq data to enable streamlined simulations for benchmarking both wet lab protocols and analysis algorithms. Furthermore, we compile our experience in processing different types of scRNA-seq data, in particular with barcoded libraries and UMIs, and developed *zUMIs*, a fast and flexible scRNA-seq data processing software overcoming shortcomings of existing pipelines.

In addition, we used the in-depth characterization of scRNA-seq technology to optimize an already powerful scRNA-seq protocol even further. According to data generated from exogenous mRNA spike-ins, this new *mcSCRB-seq* protocol is currently the most sensitive scRNA-seq protocol available.

Single-cell resolution makes scRNA-seq uniquely suited for the understanding of complex diseases, such as leukemia. In acute lymphoblastic leukemia (ALL), rare chemotherapy-resistant cells persist as minimal residual disease (MRD) and may cause relapse. However, biological mechanisms of these relapse-inducing cells remain largely unclear because characterisation of this rare population was lacking so far. In order to contribute to the understanding of MRD, we leveraged scRNA-seq to study minimal residual disease cells from ALL. We obtained and characterised rare, chemotherapy-resistant cell populations from primary patients and patient cells grown in xenograft mouse models. We found that MRD cells are dormant and feature high expression of adhesion molecules in order to persist in the hematopoietic niche. Furthermore, we could show that there is plasticity between resting, resistant MRD cells and cycling, therapy-sensitive cells, indicating that patients could benefit from strategies that release MRD cells from the niche. Importantly, we show that our data derived from xenograft models closely resemble rare primary patient samples.

In conclusion, my work of the last years contributes towards the development of experimental and computational single-cell RNA sequencing methods enabling their widespread application to biomedical problems such as leukemia.

Introduction

Gene Expression

DNA is the essential biomolecule containing all genetic information being passed on from generation to generation (Avery et al. 1944). The central dogma of molecular biology (Crick 1958) describes the directional relationship of genetic information in organisms: self-replication of DNA, transcription of DNA into a transient messenger RNA (mRNA), which is in turn translated into amino-acid sequences by ribosomes. Cells in complex multicellular organisms have to fulfill a wide variety of functions. Thus, each cell type needs a defined set of proteins to function correctly. However, all nucleated cells of an organism contain the complete DNA sequence including all gene and non-coding sequences, termed genome (Winkler 1920). Thus, specific patterns of transcription of certain genes control the proteome and thereby a cell's identity. Importantly, transcription is not controlled in a binary (on/off) manner but rather mRNA amounts correlate with protein abundance (Vogel & Marcotte 2012; Edfors et al. 2016). The necessary fine regulation of expression levels is achieved by several major mechanisms: (1) Chromatin state, (2) DNA methylation, (3) transcription factors and (4) enhancers. The first mechanism, chromatin state, describes the status of the packaging of DNA into complex nucleoprotein structures (Voss & Hager 2014). Second, methylation of cytosine residues, most importantly to 5-methylcytosine, is an important epigenetic mark controlling the silencing of genes (Jones 2012). Thirdly, transcription factors are proteins with sequence-specific binding properties that direct the initiation of transcription at promoters upon sequence-specific binding to modulate gene expression (Vaquerizas et al. 2009). Lastly, enhancers are regulatory sequences that contain transcription factor binding site DNA motifs, leading to an increased transcription level when in proximity to the transcription start site.

Taken together, these mechanisms provide the capacity to control gene expression in precise patterns that lead to individual cellular function, phenotype or development. Understanding

transcriptional regulation is of high importance, not only to understand biological processes, such as development, but also because misregulation of transcription is associated to a wide range of diseases, such as cancer (Vaquerizas et al. 2009).

mRNA Quantification

As the gene expression levels are important to regulate functions and development in cells, tissues and organisms, there is naturally a large interest in quantifying mRNA transcripts. Historically, mRNA quantification began with so-called ”Northern blots” in 1977 (Alwine et al. 1977), for which electrophoretically separated RNA molecules are transferred to paper membranes and detected by radioactively labelled probes. Later, fluorescently labelled DNA probes were used to quantify mRNA *in-situ* by hybridization (“fluorescence in-situ hybridization”, FISH) (Pachmann 1987). A third technique to quantify mRNA species relies on PCR (Mullis et al. 1986) after reverse transcription of mRNA into cDNA. Quantitative information of this “qPCR” technique is achieved by incorporation of a fluorescent dye and measuring fluorescent signals after each amplification round (Becker-André & Hahlbrock 1989; Weis et al. 1992). However, these methods only quantify specific mRNAs and cannot provide an unbiased, genome-wide survey of the transcriptome. Conversely, other techniques such as sequencing cloned cDNA, so-called “expressed sequence tags” (EST) (Marra et al. 1998; Adams et al. 1991), can survey the transcriptome but without quantitative information. The first methods obtaining global gene expression data with quantitative information were “serial analysis of gene expression” (SAGE) (Velculescu et al. 1995) and microarrays (Schena et al. 1995). In SAGE, short fragments of cDNA samples were produced by restriction digests and then concatenated for subsequent Sanger sequencing. The method could measure the expression of thousands of genes but suffered from the ambiguity of the often very short tags (eg. 9 bp) (Yamamoto et al. 2001).

DNA microarrays describe synthetic DNA oligonucleotide probes that are immobilized on a surface called microarray. Initially, oligonucleotides were synthesized and then spotted on the surface in a very fine grid, later on the synthesis of probes was done directly on the surface (Miller & Tang 2009). Importantly, the probes are designed to be complementary to cDNA and each "probe spot" thus corresponds to a specific gene sequence. Next, fluorescently labelled cDNA can be added to the microarray and will hybridize to complementary probes. A fluorescent signal will be produced wherever cDNA was bound, with light intensity proportional to the amount of cDNA hybridized. Because of their relatively easy application, low cost and good data quality, microarrays became a very popular method and are still utilized to date (Figure 1).

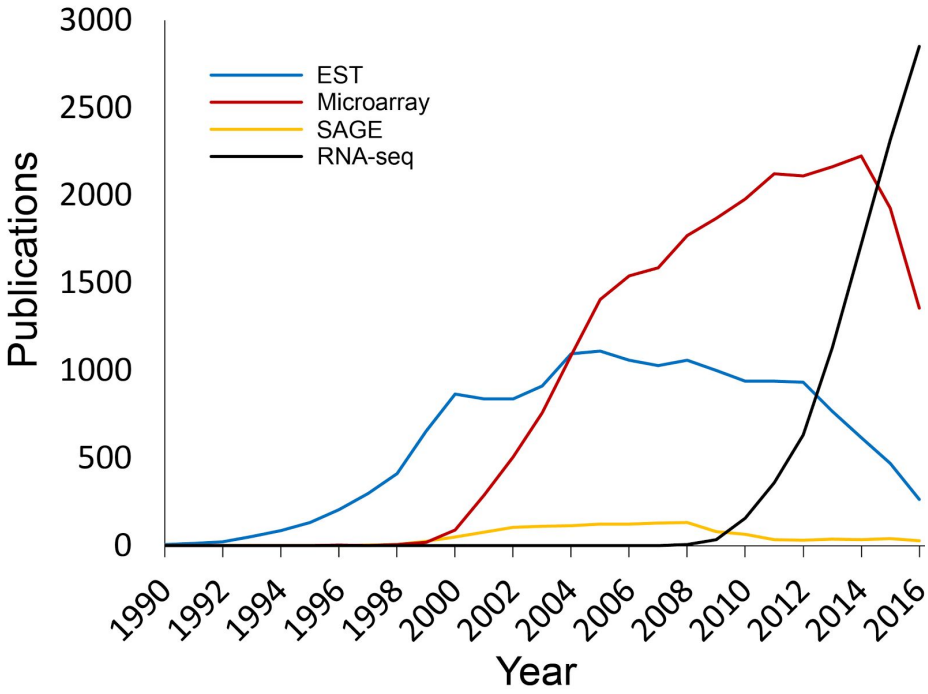


Figure 1: Popularity of transcriptomics methods. Adapted from Lowe et al. 2017 (Lowe et al. 2017). The graph shows the number of published articles per year using ESTs, Microarrays, SAGE or RNA-seq in Pubmed since 1990.

Further advances in the density of probes on the microarray and improved technology to detect fluorescence signals made the microarray even wider applicable, also for lowly expressed genes (Pozhitkov et al. 2007). However, microarrays have the major drawback that only known cDNA sequences can be interrogated, as the oligonucleotide probes have to be designed *a priori*. This prevents discovery of new genes or splicing isoforms and limits the applicability to model organisms with well-resolved transcriptomes, such as mouse and human. These limitations were later overcome in the transformative technique known as "RNA sequencing" (RNA-seq), which leverages the advent of next generation/high-throughput sequencing by sequencing large amounts of cDNA fragments to determine global gene expression levels in a given sample (Mortazavi et al. 2008).

High-Throughput DNA Sequencing

While determining the first human genome sequence (Lander et al. 2001) took a worldwide effort, 20 years to complete and approximately 3 billion dollars in funding, current technology permits the sequencing of a human genome at 30x coverage for only 1000 USD (Hayden 2014). This incredible cost reduction was caused by a rapid development of sequencing techniques (Kircher & Kelso 2010) that have considerably higher output than traditional Sanger sequencing (Sanger et al. 1977). Since 2005, several high-throughput sequencing techniques have been in the spotlight (Margulies et al. 2005; Valouev et al. 2008), but were superseded by a technology marketed by the company "Illumina", that currently captures most of the sequencing market worldwide (Zimmerman 2014). This sequencing method (Bentley et al. 2008) is a variant of the sequencing by synthesis strategy using cyclic reversible termination (Turcatti et al. 2008). DNA templates used in Illumina sequencing are prepared by adding immobilisation adapters to fragmented DNA. This enables the attachment of DNA templates to flow-cells (Fedurco et al. 2006). Initially, the attachment was done randomly, but later iterations of the technology utilize patterned nanowell flow-cells to increase template density

(Illumina 2015). After binding to the flow-cell surface, each template molecule is amplified in place to clusters consisting of 1000s of clonal copies by a process called “bridge amplification” (Fedurco et al. 2006). Here, the number of generated clusters will determine the number of sequenced reads. Next, sequencing-by-synthesis is performed by adding all four bases at the same time. The bases are modified to include a unique fluorophore and a cleavable chain terminator. Thus, only one base will be incorporated at a time. The fluorescence signal from each of the clonal template clusters can then be recorded. Subsequently, both the fluorophore and the chain terminator group is cleaved and the next cycle for sequencing can occur. After a predetermined amount of sequencing cycles are performed, base-calling can be done from fluorescence data (Kircher et al. 2009). Initially, read-lengths were limited to 26 bases (Kircher & Kelso 2010), but continuous technology improvements increased this to 600 bases (300 bases paired-end) for some Illumina machines (Genohub 2017; Goodwin et al. 2016). The major errors encountered are substitutions due to false base incorporation during the SBS reaction. Generally, sequencing errors occur more frequently towards the end of reads, as clusters are prone to phasing when some of the clonal molecules fail to cleave fluorophores or fail to incorporate a base (Kircher & Kelso 2010).

It should be noted that unlike the clonal amplification-based short read technologies discussed, two major technologies for true single-molecule real-time detection are currently in fast-paced technological development competition. Pacific Biosciences relies on the detection of DNA synthesis from a single polymerase enzyme fixed in place using fluorescently labelled dNTPs (Eid et al. 2009). The second technology does not rely on SBS. Oxford Nanopore Technologies sequencing strategy utilizes membrane proteins forming nanopores immobilized on an array (Deamer et al. 2016). Because each nucleotide of DNA has a slightly different molecular structure, an ionic current through the pore will change accordingly (Jain et al. 2016). Although challenges in the interpretation of signals are currently still large, this method has great potential because it gives the unique opportunity to directly detect

base-modifications, such as 5-mC (Simpson et al. 2017) or direct RNA sequencing without prior cDNA synthesis (Garalde et al. 2016).

Currently, Illumina sequencers generate large amounts of high-quality data for the lowest per-base price and are thus predominantly used in RNA sequencing studies, as relevant for this work.

RNA Sequencing (RNA-seq)

Overcoming the limitations of previous transcriptomics approaches, RNA sequencing (RNA-seq) has become the most widely used global gene expression analysis method to date (Figure 1). First published by several groups in 2008 (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Wilhelm et al. 2008; Marioni et al. 2008), RNA-seq leverages high-throughput DNA sequencing technologies to massively parallel sequence cDNA fragments produced from mRNA samples. This comprehensive profiling yields both quantitative and qualitative information on gene expression, without prior design or selection of probe sequences, removing microarrays' limitations for detection of RNA splice patterns and previously unannotated genes.

Although a plethora of library preparation methods exist for RNA sequencing (Levin et al. 2010), the general workflow can be summarised in few steps (Figure 2). First, RNA has to be extracted from the sample of interest. As more than 80% of a cell's RNA content is uninformative ribosomal RNA (rRNA), all protocols are actively depleting rRNA or enriching for mRNA (eg. selection of polyadenylated RNA species; "poly-A+ mRNA") (Choy et al. 2015). Next, mRNA is reverse transcribed into cDNA. Since most of the mRNA present in eukaryotic cells is much longer than the generated sequencing read length and clustering of DNA molecules on the sequencer's flow-cell is inefficient for molecules larger than ~1 kb, transcripts need to be fragmented prior to analysis.

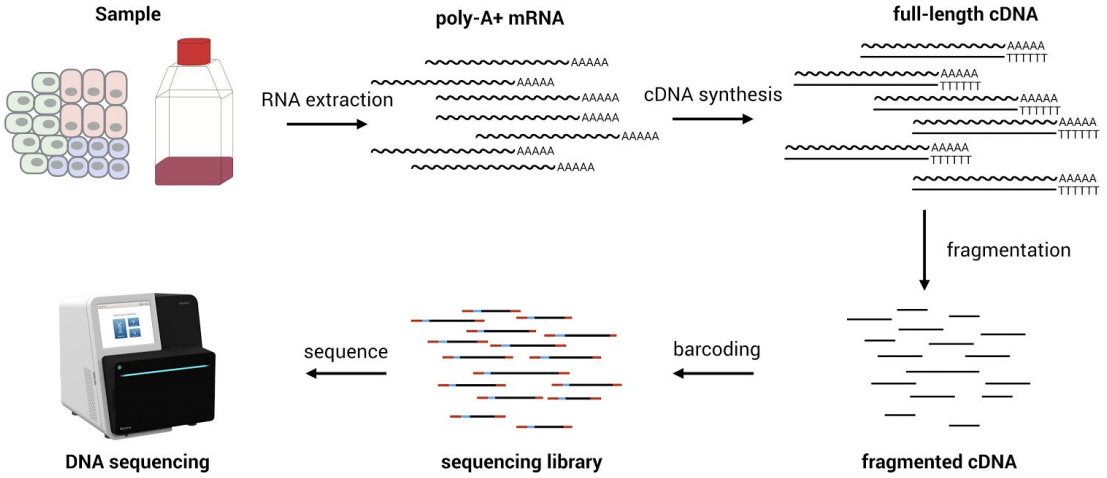


Figure 2: Experimental workflow for RNA sequencing. RNA sequencing can be performed from any biological samples after extraction of RNA. After reverse transcription, cDNA is fragmented and barcoded with multiplexing sequences and sequencing adapters. Finally, libraries are subjected to high-throughput sequencing.

This fragmentation step can be performed at the RNA level (heat or chemical fragmentation by RNA hydrolysis (Mortazavi et al. 2008)) or at the cDNA level by sonication (Head et al. 2014) or enzymatic processes (Adey et al. 2010; Picelli, Björklund, et al. 2014). Furthermore, sample-specific DNA barcodes may be added to facilitate multiplexing of many samples into single sequencing runs (Meyer & Kircher 2010; Kircher et al. 2012). Subsequently, final sequencing libraries may be amplified in a library PCR (van Dijk et al. 2014). Finally, sequencing libraries are loaded on high-throughput sequencing machines at defined concentrations (Marioni et al. 2008). For Illumina, massively parallel sequencing will typically yield millions of reads per sample in RNA sequencing experiments, with sequencing cost usually being the only limiting factor to read depth (Wang et al. 2009; Conesa et al. 2016).

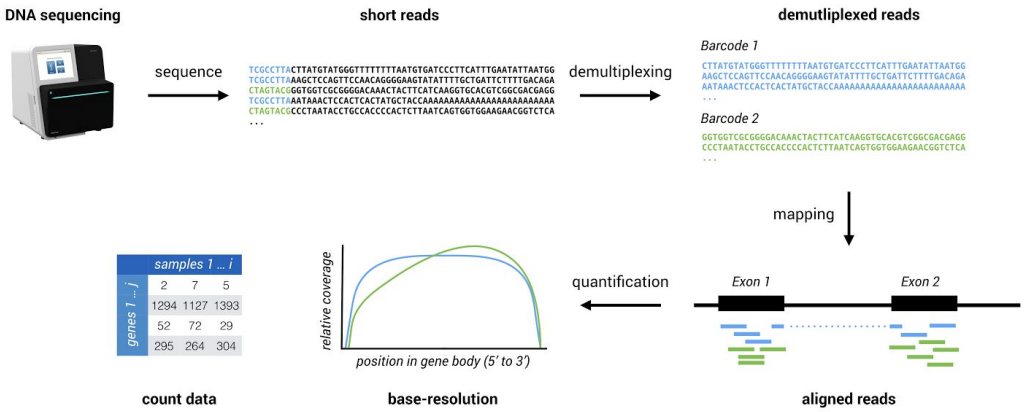


Figure 3: Computational workflow for RNA sequencing. Short reads are demultiplexed according to multiplexing barcode sequenced, mapped against reference genome and quantified.

For RNA-seq data analysis (Figure 3), generated reads from the sequencer will first need to be demultiplexed according to the sample-specific multiplexing barcode sequences used (Renaud et al. 2015). Next, reads will be aligned to the reference genome of the organism analysed to find out which genomic location they correspond to. This is a computationally complex task further complicated by the fact that most transcripts are spliced from several exons interspersed by long intronic sequences in the reference genome (Hayer et al. 2015). To overcome this, splice-aware alignment software has been developed to be able to confidently map short reads across annotated and newly discovered splice junctions (Engström et al. 2013; Baruzzo et al. 2017). After this processing, reads can be quantified in a per-base analysis over the gene body sequence or as gene-/isoform-level count data to obtain expression level estimates. Commonly, there are several quality-control checkpoints in the analysis of RNA sequencing data to ensure reliable high quality results (Conesa et al. 2016). Briefly, raw reads should be checked for their base-calling quality score distribution (eg. sequencing machine errors), overrepresentation of k-mers (eg. poly-A stretches or low-complexity libraries) and GC content distribution (Andrews 2010). Next, quality of the mapping step should be

controlled using tools such as RSeQC (Wang et al. 2012) or picard. In human and mouse experiments, most (> ~60-80%) of the reads in RNA-seq libraries should typically align to the reference genome. Apart from the fraction of mapped reads, GC bias, uniformity of the read distribution and sequence duplication levels can be indicative for data quality at this stage (Okonechnikov et al. 2016). At the stage of quantification, QC parameters are less obvious. Generally, gene expression values of replicates should be reproducible and thus show a high correlation with each other, especially for technical replicates. For biological replicates, acceptable correlation coefficients are dependant on the variance and heterogeneity of the studied biological system (Conesa et al. 2016).

After initial data processing and QC, higher level analysis can be performed using RNA sequencing data. This typically includes, but is not limited to: differential gene expression analysis (Soneson & Delorenzi 2013; Rapaport et al. 2013; Seyednasrollah et al. 2015), gene set enrichment analysis (Tarca et al. 2013; Bayerlová et al. 2015), co-expression and network analysis (Langfelder & Horvath 2008; Ballouz et al. 2015; van Dam et al. 2017).

In summary, RNA sequencing shows clear advantages over previously used transcriptomics methods and provides rich information in a cost-efficient manner.

Single-Cell RNA Sequencing

Conventional RNA sequencing methods generally require large amounts (100 - 1000 ng) of total RNA as input. However, this can be a limiting factor when samples consisting only of few cells need to be analysed. In order to access these low-input samples, whole transcriptome amplification methods have been developed (Bhargava et al. 2014). Nearly all of the presently used low-input protocols feature either exponential PCR amplification (Mullis et al. 1986), linear in-vitro transcription (Milligan et al. 1987) or multiple displacement amplification (Blanco et al. 1989). Shortly after RNA-seq became possible whole transcriptome amplification technologies were used to sequence the transcriptomes of individual cells (Tang et al. 2009).

Further development of this breakthrough technology (Nature Methods 2014) has already shown to be transformative to our understanding of biology (Shapiro et al. 2013; Sandberg 2014; Wagner et al. 2016). Single-cell RNA-seq opens the possibility to investigate global patterns of gene expression variability within cell types (Deng et al. 2014) or between groups of cells (Kolodziejczyk, Kim, Tsang, et al. 2015; Martinez-Jimenez et al. 2017). Furthermore, the increased resolution provided by scRNA-seq has allowed researchers to uncover previously unknown subpopulations in various compartments, such as the immune system (Villani et al. 2017; Jaitin et al. 2014), liver (Halpern et al. 2017), pancreas (Muraro et al. 2016; Grün et al. 2015; Baron et al. 2016), lung (Treutlein et al. 2014), retina (Macosko et al. 2015; Shekhar et al. 2016) and brain (Tasic et al. 2016; Fuzik et al. 2016; Usoskin et al. 2015; Zeisel et al. 2015; Lake et al. 2016; Habib et al. 2017). In addition, other studies have reconstructed and uncovered novel dynamics and heterogeneity in developmental processes of embryos (Biase et al. 2014; Yan et al. 2013), blood (Nestorowa et al. 2016; Moignard et al. 2015) and brain (La Manno et al. 2016) by applying single-cell transcriptomics. Importantly, scRNA-seq can not only be used for descriptive understanding of biology but is also applied to large-scale perturbation experiments providing mechanistic insight in molecular networks (Dixit et al. 2016; Jaitin et al. 2016; Xie et al. 2017; Datlinger et al. 2017). Lastly, single-cell transcriptomics is used to investigate heterogeneity in disease states, which is of high relevance for cancer evolution (Patel et al. 2014; Tirosh et al. 2016; Venteicher et al. 2017).

Although the exciting opportunities that single-cell RNA sequencing is providing are obvious, neither the experimental technology nor the computational analysis has converged to an optimum yet. In order to generate and interpret data adequately, it is therefore necessary to understand properties, power and limitations of scRNA-seq technologies.

Isolation of Single Cells

The first step of any scRNA-seq workflow is the isolation of single cells or nuclei. Cell isolation depends on many factors that need to be considered and can also be closely linked to the chosen scRNA-seq protocol, such as the number of available cells, cell viability and whether subsets of cells need to be enriched. While cell isolation can be relatively straightforward for suspension cells, other cases might require more complex experimental setups. For instance, manual microdissection has been useful to obtain very rare cells, as in studies of early development from zygote to blastocyst (Tang et al. 2009; Deng et al. 2014). In order to access solid tissues however, they need to be dissociated into single-cell suspensions. Yet, dissociations are prone to pitfalls (Poulin et al. 2016), depending on the tissue studied (Figure 4A). First, cells may be damaged or respond to enzymatic digestions of surface proteins leading to decreased capture (Huang et al. 2010). Second, long handling steps and incubation times can alter gene expression profiles prior to analysis (Alles et al. 2017; van den Brink et al. 2017). Third, readiness to dissociate is widely variable among cell types, leading to depletion of certain types or enrichment of others (Figure 4A). A well known example is brain tissue, where the dissociation of neurons from their strongly interconnected network is challenging (Poulin et al. 2016). Further, incomplete dissociation can lead to unwanted doublets or clumps that hinder true single-cell resolution (Figure 4A). Once samples are dissociated into single-cell suspensions, there are several possible ways to capture cells for sequencing (Figure 4B). Specifically designed microfluidic chips can capture 96 or 800 cells (Pollen et al. 2014; Wu et al. 2014) and process them in the Fluidigm C1 microfluidic controller. A second application of microfluidics technologies is the use of microdroplets. Here, cells are encapsulated in droplets of defined nanoliter size in a water-in-oil emulsion (Macosko et al. 2015; Klein et al. 2015; Zilionis et al. 2017; Zheng et al. 2017). Droplet capture of single-cells can be used for unbiased high-throughput capture of many single cells independent of their cell size. For droplet-based

methods, a major obstacle can be damaged cells which may break and release their RNA. This free RNA can subsequently leak into all droplets, generating a certain background noise and wasting sequencing coverage (Figure 4A).

In many experiments, preselection of certain cell types is desirable. Usually, this selection may be performed by FACS, in which case single cells may be directly sorted in individual wells of 96- or 384-well plates (Kolodziejczyk, Kim, Svensson, et al. 2015; Soumillon et al. 2014; Jaitin et al. 2014; Hashimshony et al. 2016). These multiwell plates typically contain lysis buffer to break the cell and release the RNA for subsequent processing. Furthermore, compatible FACS machines can allow "index sorting", which means fluorescence data for each deposited cell can be associated to the well position (Hayashi et al. 2010), providing additional data.

A relatively new alternate approach is the deposition of cells into microfabricated nanowells by limiting dilution (H. C. Fan et al. 2015; Gierahn et al. 2017; Hochgerner et al. 2017).

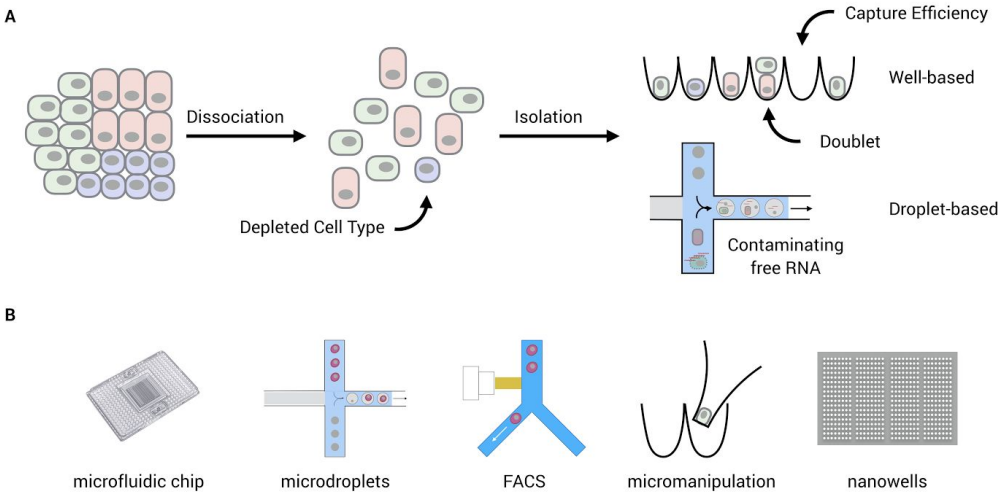


Figure 4: Isolation of single cells for sequencing. A: Illustration of typical issues during single-cell isolation. Dissociation can lead to depletion of certain cell types by damage and cell death. After incomplete dissociation, doublets may be problematic. In droplet-based cell isolation, contaminating free RNA may leak in microdroplets. B: Illustrations of major popular single cell isolation methods.

Nanowells methods are coming into focus as they offer high throughput without the need for microfluidic droplet setups.

Obviously, depending on the experimental parameters (eg. rare/abundant cells, unbiased/preselected cells, suspension/dissociated cells) tradeoffs of different techniques will need to be considered. In all cases, data should be carefully examined to estimate capture and doublet rates. For instance, in a study of pancreatic islet cells using the Fluidigm C1 system, the doublet rate was found to be 31% due to an issue with the microfluidic chip design by analysing mutually exclusive hormone-producing genes (Xin et al. 2016). In order to experimentally validate doublet rates Macosko et al. and Klein et al. first proposed co-isolation of mouse and human cells in one experiment. By mapping the transcriptomes, mixed-species samples can be identified readily and should make up a third of doublets (Macosko et al. 2015; Klein et al. 2015). After successful cell capture, single-cell RNA is obtained and processed for sequencing.

Generating scRNA-seq libraries

After cell capture, each protocol for single-cell RNA sequencing consists of three major steps: (1) reverse transcription of mRNA into cDNA, (2) pre-amplification of cDNA and (3) sequencing library preparation (Figure 5A). Reverse transcription is the essential first step in scRNA-seq after cell lysis is completed. The conversion into cDNA is considered to be especially inefficient, and only an estimated fraction of 10-40 % of mRNA molecules are reverse transcribed (Grün et al. 2014; Islam et al. 2014). Several studies have reported that small reaction volumes in microfluidic machines may increase the efficiency of this step (Wu et al. 2014; Streets et al. 2014; Hashimshony et al. 2016). Still, it is important to systematically optimize reaction conditions of this step, as was shown for the Smart-seq2 protocol (Picelli et al. 2013).

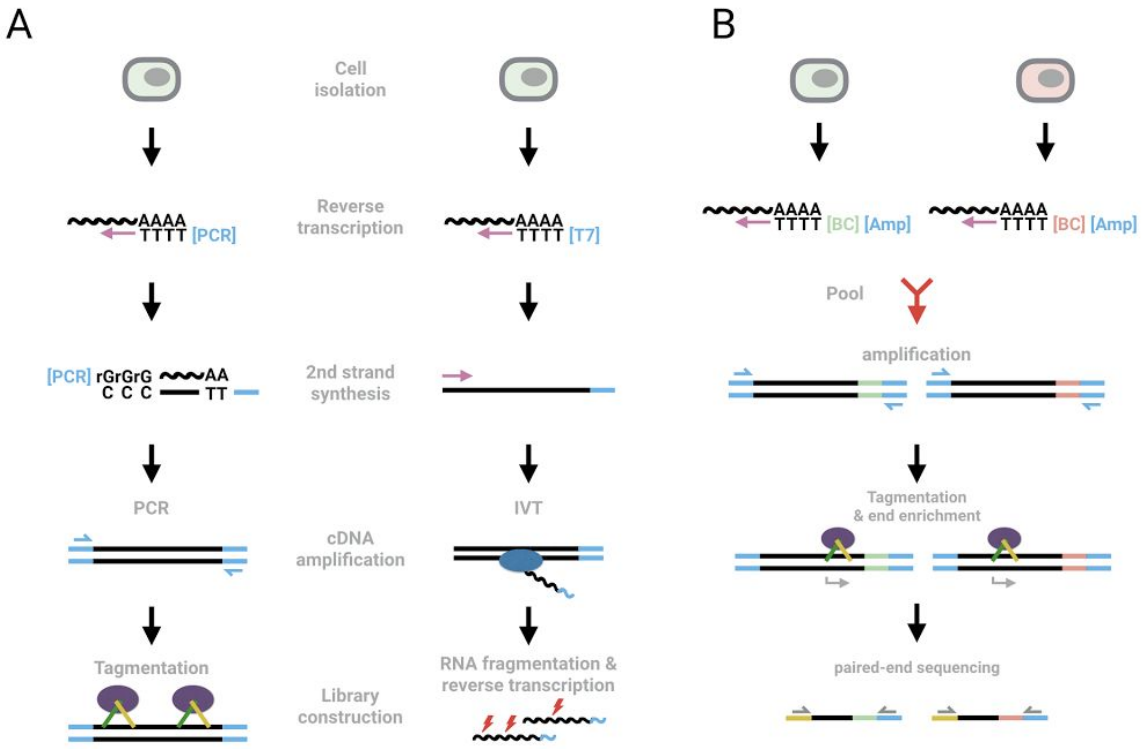


Figure 5: Preparation of scRNA-seq libraries. A: Typical whole transcriptome amplification strategies are illustrated (left: PCR, right: IVT). B: Illustration of early cell barcoding enabling massively parallel scRNA-seq.

In order to select for the polyadenylated mRNA fraction of the cellular RNA content, virtually all protocols utilize oligo-dT primers to initiate the reverse transcription reaction, with exceptions of protocols specifically aiming to sequence the total RNA transcriptome (X. Fan et al. 2015; Sheng et al. 2017). Recently developed protocols feature cell-specific barcodes incorporated during the reverse transcription step (Figure 5B) to increase throughput of scRNA-seq methods (Islam et al. 2011; Hashimshony et al. 2012; Soumillon et al. 2014; Jaitin et al. 2014; Macosko et al. 2015; Klein et al. 2015; Zheng et al. 2017; Hochgerner et al. 2017). Being able to pool reactions as early as possible is associated with a large drop in reagent costs and labor time. Because the cell-barcode is usually located in the primer sequence (eg. oligo-dT primer), sequencing is restricted to the end of the transcript (5' end or 3' end) to be able to

associate cDNA fragments with barcode information. However, the increasing throughput of protocols and the association of barcode information to cDNA sequence poses challenges to bioinformatics pipelines for processing such data. After reverse transcription, the minute amounts of cDNA are pre-amplified. Popularly, scRNA-seq protocols use PCR or IVT to achieve this. In the case of PCR, most of the current protocols (Ramsköld et al. 2012; Picelli et al. 2013; Soumillon et al. 2014; Macosko et al. 2015; Rosenberg et al. 2017) rely on the template-switching mechanism (Zhu et al. 2001; Zajac et al. 2013) to place a known primer sequence at the 5' end of the transcript, in addition to a PCR handle being present at the 3' end within the oligo-dT primer (Figure 5A). Much is already known on sequence-specific biases in general sequencing library PCR amplification (Aird et al. 2011). Since PCR amplification is exponential, any sequence-dependent bias (eg due to GC content, length) can propagate and potentially distort expression profiles (Kolodziejczyk, Kim, Svensson, et al. 2015). Thus, the number of amplification cycles should be carefully optimized in PCR-based scRNA-seq methods (Picelli, Faridani, et al. 2014).

In contrast, the most common alternative pre-amplification method, in-vitro transcription (IVT), is a linear amplification technique and has been used in several important scRNA-seq protocols (Hashimshony et al. 2012; Jaitin et al. 2014; Klein et al. 2015; Hashimshony et al. 2016; Zilionis et al. 2017). Here, amplification biases should be less pronounced than in PCR-based methods. However, IVT-based methods need a second reverse transcription reaction, which increases 3' bias (Kolodziejczyk, Kim, Svensson, et al. 2015). Thus, all presently used IVT-based methods selectively sequence 3' ends.

Finally, after pre-amplification of single-cell material, library preparation needs to be done to make cDNA compatible with high-throughput sequencing. Since most of the protocols use the Illumina platform for sequencing, the Illumina Nextera kit (Adey et al. 2010; Picelli, Björklund, et al. 2014) is a popular choice for fragmentation and adapter incorporation.

In scRNA-seq, the inefficient capture and amplification of the low starting amounts of mRNA leads to high technical noise, especially for lowly expressed genes (Brennecke et al. 2013; Islam et al. 2014; Grün et al. 2014; Kim et al. 2015). In order to remove amplification noise, many protocols include unique molecular identifiers (UMIs; Figure 6). By incorporation of a random sequence during reverse transcription, each initial cDNA molecule will most likely have a unique sequence, given sufficient UMI length. During amplification, several copies of each cDNA with its specific UMI sequence will be made. After sequencing of cDNA fragments together with their UMI, counting of initial cDNA molecules is possible by comparing the number of unique UMI sequences per gene and cell (Figure 6).

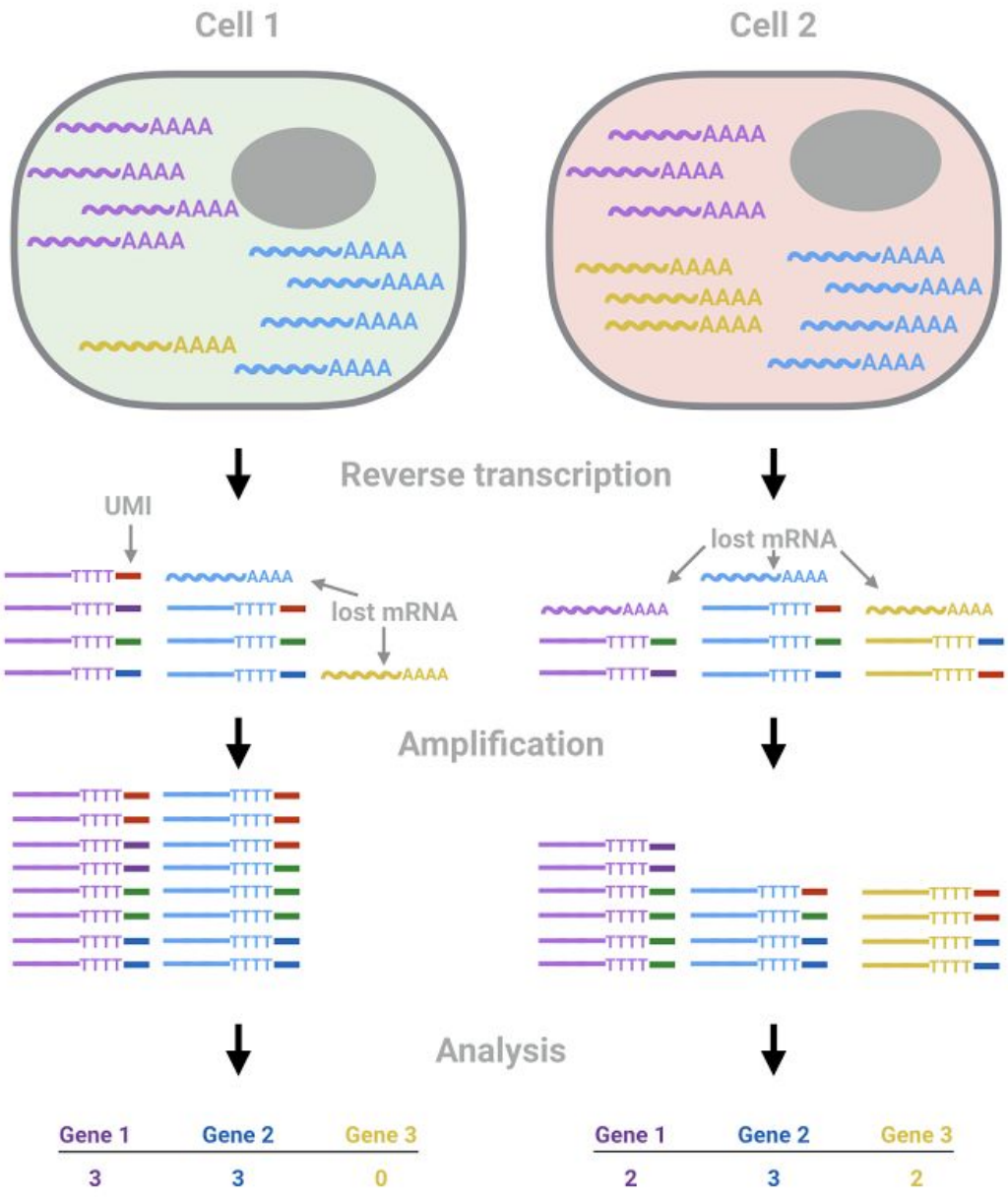


Figure 6: Illustration of unique molecular identifiers. Shown are two schematic cells with transcripts from three genes. During reverse transcription, unique molecular identifiers are incorporated. After amplification and sequencing, reads coming from unique molecules can be distinguished by their UMI sequence, thus removing amplification bias.

With the rapid pace of scRNA-seq method development, there is a high need for independent assessment of protocol performance. Ideally, a single-cell RNA sequencing protocol would be (1) compatible with several cell isolation procedures, (2) highly sensitive to capture mRNA molecules, (3) accurately represent absolute expression levels and (4) be precise. First, being flexible to several cell inputs gives a scRNA-seq method broad applicability. Second, high sensitivity to capture and detect mRNA molecules is beneficial to observe gene expression comprehensively including lowly expressed genes and prevent dropout events (Figure 7A). Third, accuracy describes how closely gene expression measurements correspond to actual mRNA concentrations in the cell (Figure 7B). Fourth, precision, describes the technical variation of gene expression measurements, a parameter that is largely driven by amplification noise in RNA sequencing data. In practice, not only the outright performance of a single-cell protocol will decide the choice of method. Rather, practical considerations, such as batch size, available sample and equipment are important considerations. Still, measuring and comparing these technical parameters is an important question in the field. Furthermore, the technical performance of scRNA-seq methods should be seen relative to their cost to inform optimal design of future studies.

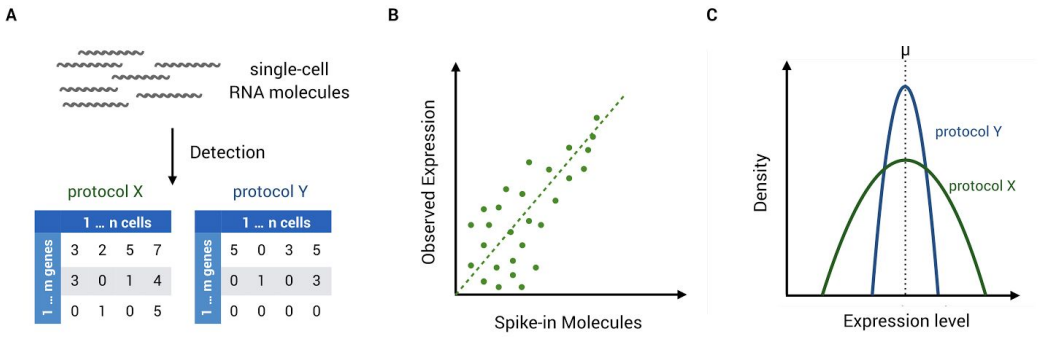


Figure 7: Technical parameters of scRNA-seq data. A: Sensitivity describes the fraction of single-cell RNA molecules that can be detected and quantified. B: Accuracy describes the correlation of observed expression levels to annotated spike-in mRNA concentrations. C: Precision describes the technical variation of gene expression measurements.

Experimental Design

As mentioned above, single-cell RNA sequencing data is subject to large amounts of technical noise (Brennecke et al. 2013; Grün et al. 2014) that can vary from batch to batch (Tung et al. 2017). Hence, technical nuisance factors need to be decoupled from biological factors of interest by appropriate experimental design (Hicks et al. 2015). Early barcoding and multiplexing can help reduce batch effects by handling and processing samples of different conditions in parallel as much as possible (Robles et al. 2012). A multiplexed workflow should include sequencing samples on mixed lanes (Auer & Doerge 2010). Because technical variation can never be excluded, it is recommend that cells from all studied conditions are sequenced together in multiple batches to be able to factorize and remove batch-associated variability in the following statistical analysis (Figure 8) (Hicks et al. 2015). Naturally, practical technological constraints, such as the number of input cell suspensions into a microfluidic device, may pose limits to the accommodation of several factors into a single batch (Lun & Marioni 2017). Batch effects can be included in the DE modelling as an extra covariate to separate biological and technical effects on gene expression measurements. Furthermore, algorithms that are able to align and integrate data from various batches or technologies are in active development in order to make most use of existing data (Butler & Satija 2017).

In summary, when generating scRNA-seq data, batch effects and unwanted variation should be controlled for in the experimental design as much as possible or removed computationally at the analysis stage.

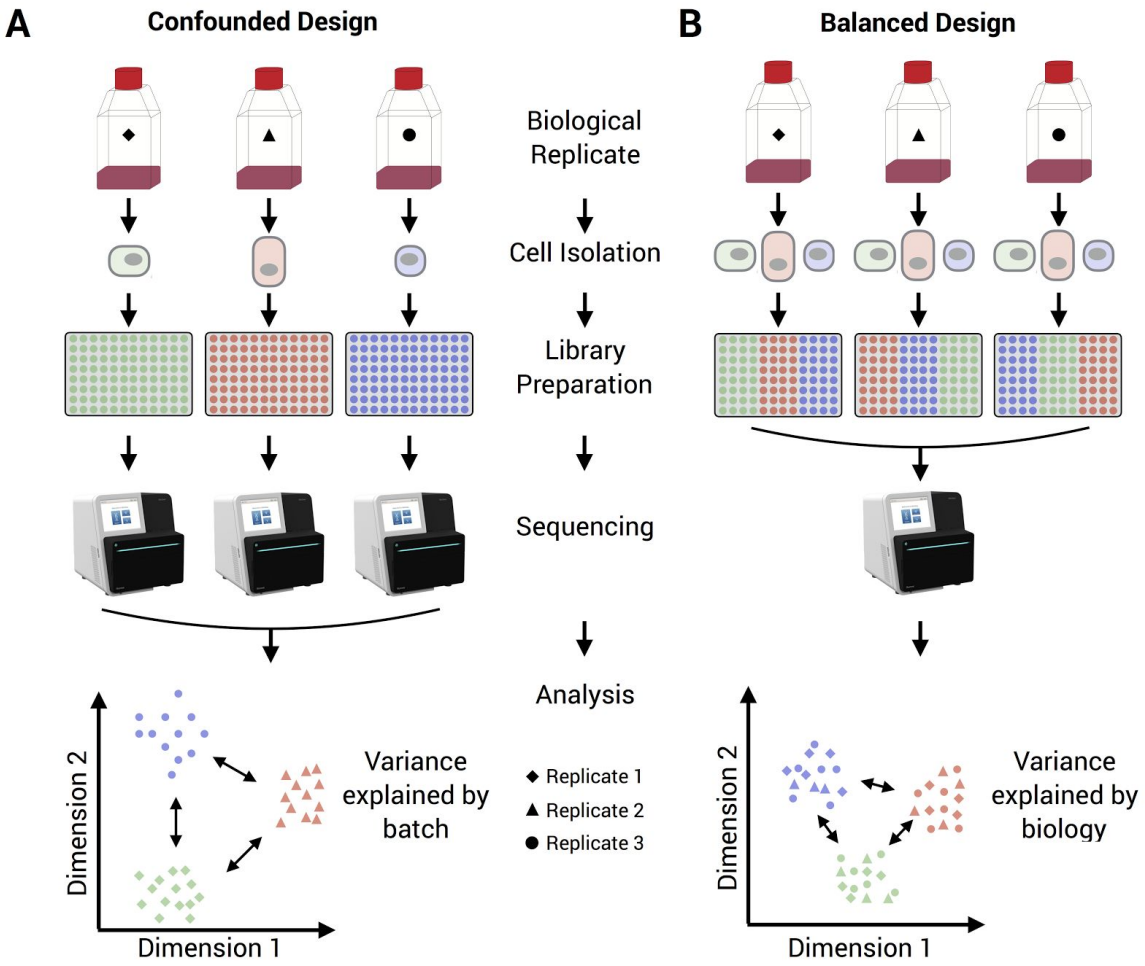


Figure 8: Design of scRNA-seq experiments. Shown is an exemplary study to detect differences between three cell types. The left side illustrates an experimental design, where cell identity is confounded with cell isolation, library preparation and sequencing batches. The right side illustrates an experiment where cell identity is balanced over technical batches, decoupling technical from biological variance.

Acute Lymphoblastic Leukemia

To illustrate the high practical pertinence of emerging scRNA-seq technologies, this work features an application in acute leukemia, which poses a number of relevant biomedical questions. Leukemia, the tenth most common cancer type (Yamamoto & Goodman 2008), describes a group of cancer diseases affecting the blood progenitor cells in the bone marrow. Acute leukemias are grouped according to the affected blood progenitor lineage into acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) (Sawyers et al. 1991). Although ALL occurs in both children and adults, over 60% of cases are pediatric, making ALL the most common cancer in children under 15 years of age (Inaba et al. 2013). With current treatment options, 90% of patients survive the disease (Hunger et al. 2012). Still, in some cases a minority of cancer cells survive therapy and cause relapse with poor prognosis (Gökbuget et al. 2012; van Dongen et al. 2015). Although of high relevance, the biological mechanisms of these relapse-inducing cells remain largely unclear. Relapse-inducing cells have self-renewal and tumor-initiating potential and thus can regrow the tumor after treatment, similar to cancer stem cells (Trumpp & Wiestler 2008). In addition, tumor-initiating cells are often resistant to chemotherapy (Clevers 2011). Because of resistance, patients with relapse have especially adverse prognosis (Nguyen et al. 2008). In ALL, resistance of tumor-initiating cells could be linked to dormancy, as chemotherapy targets proliferating cells. Dormant cells thus may persist as minimal residual disease (MRD) during treatment and give rise to relapses, and has indeed been described recently in ALL (Lutz et al. 2013). As MRD cells per definition occur at extremely rare frequencies (less than 1 ALL cancer cell in 10,000 normal cells), biological characterisation of this compartment is extremely difficult and information is limited so far (van Dongen et al. 2015). Obstacles in research on MRD cells are the rarity of primary patient material and the fact that ALL cell cultures are unsuitable to study MRD because of their continuous proliferation. Thus, a promising tool are patient-derived xenograft models, where

patient leukemia cells grow in immuno-deficient mice, closely mimicking human disease (Castro Alves et al. 2012).

Obtaining as comprehensive as possible characterisation from very rare cells is an ideal application for single-cell RNA sequencing because it enables the genome-wide quantification of gene expression from this rare cell population. Furthermore, potential heterogeneity from subclones (Inaba et al. 2013) may be detected with single-cell resolution.

Results

Improving Single-Cell RNA Sequencing Technology

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

SCIENTIFIC REPORTS

OPEN

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

Received: 25 January 2016
Accepted: 20 April 2016
Published: 09 May 2016

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

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

High throughput RNA sequencing methods (RNA-seq) are currently replacing microarrays as the method of choice for gene expression quantification[1–5]. For many applications RNA-seq technologies are required to become more sensitive, the goal being to detect rare transcripts in single cells. However, sensitivity, accuracy and precision of transcript quantification strongly depend on how the mRNA is converted into the cDNA that is eventually 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)

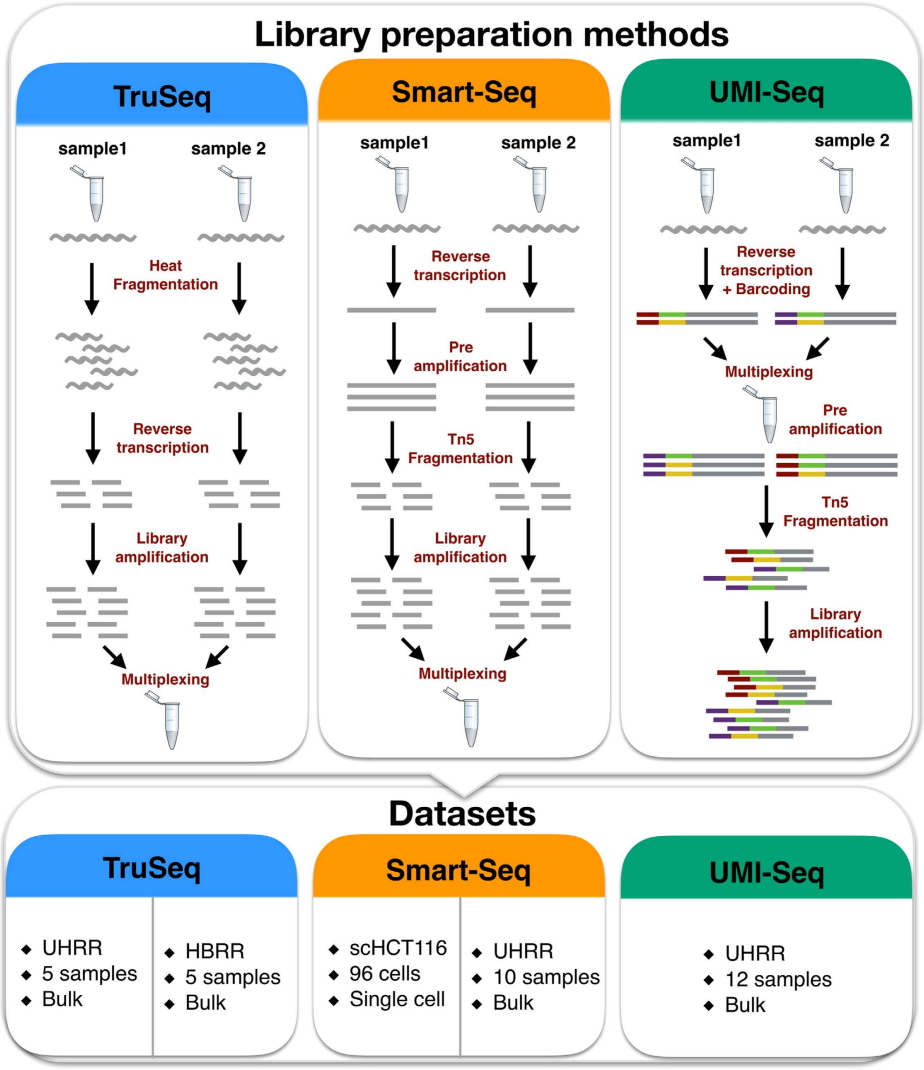


Figure 1. Schematic of library preparation protocols and datasets. The upper panel details the steps for the three sequencing library preparation methods analysed in this study. In the UMI-seq flow-chart red and purple tags represent the sample barcodes and the green and yellow tags the UMIs.

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

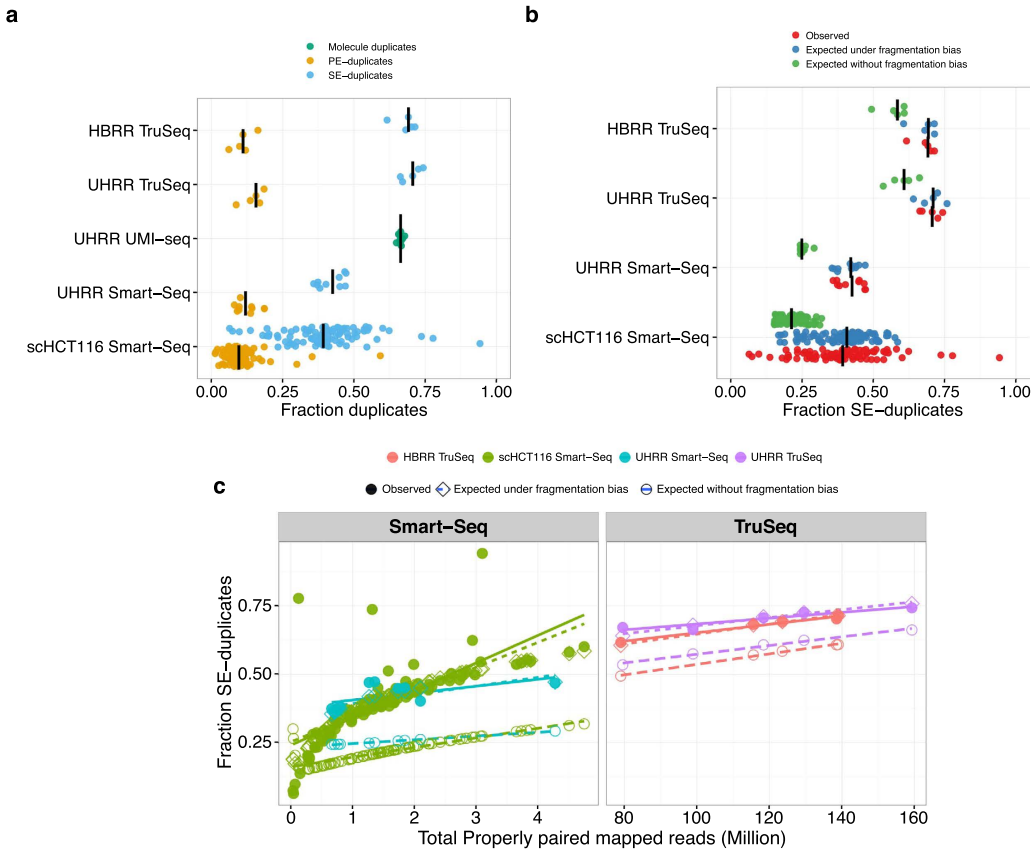


Figure 2. The Fraction of SE-duplicates increases with the total number of reads. In panel (**a**), we plot the fraction of computationally identified SE-duplicates (blue) and PE-duplicates (yellow) per sample. For the UMI-seq data, we identify duplicates only based on the experimental evidence provided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

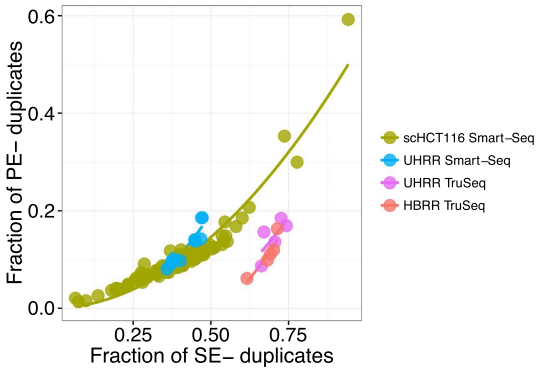


Figure 3. The relation between SE- and PE-duplicates. The relation between SE- and PE-duplicates is expected to follow a quadratic function, if the majority of duplicates are natural, i.e. due to fragmentation and sampling. Here, we show a quadratic fit for the different datasets (UHRR-TruSeq–purple,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

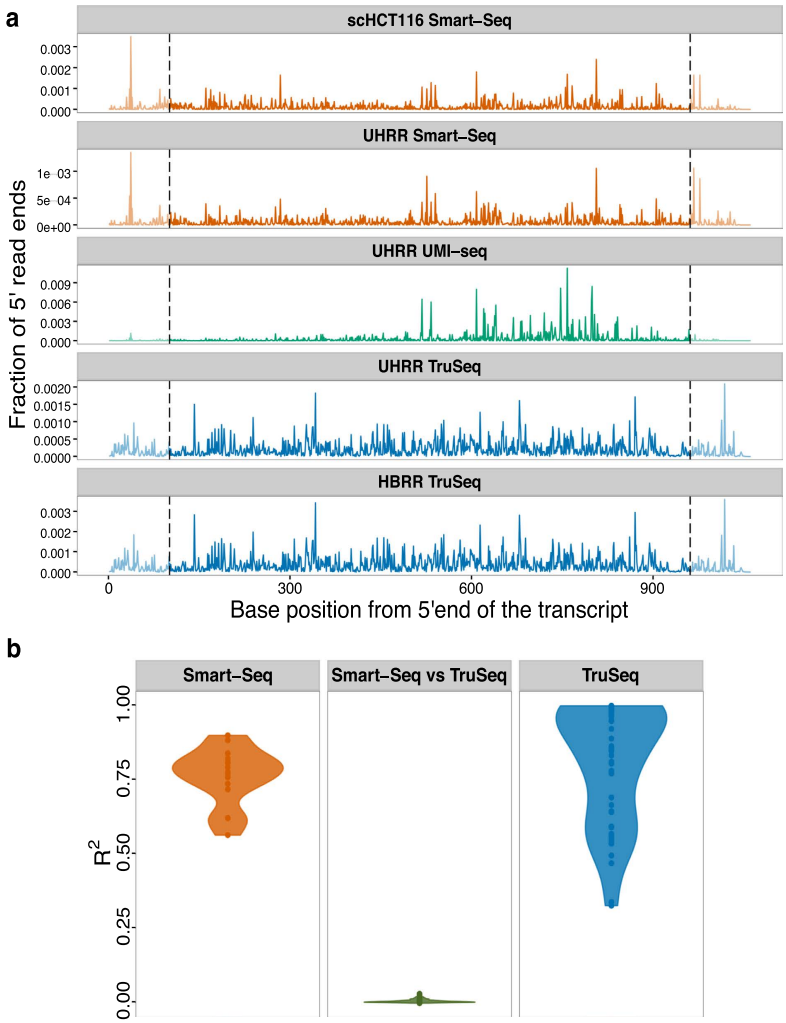


Figure 4. The fragmentation patterns of the ERCCs are highly reproducible for different samples prepared with the same RNA-seq library method. (**a**) Here, we plot the fraction of 5′ read ends per position of ERCC-00002. Because the TruSeq libries (blue) had read lengths of 100 bases, we do not consider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

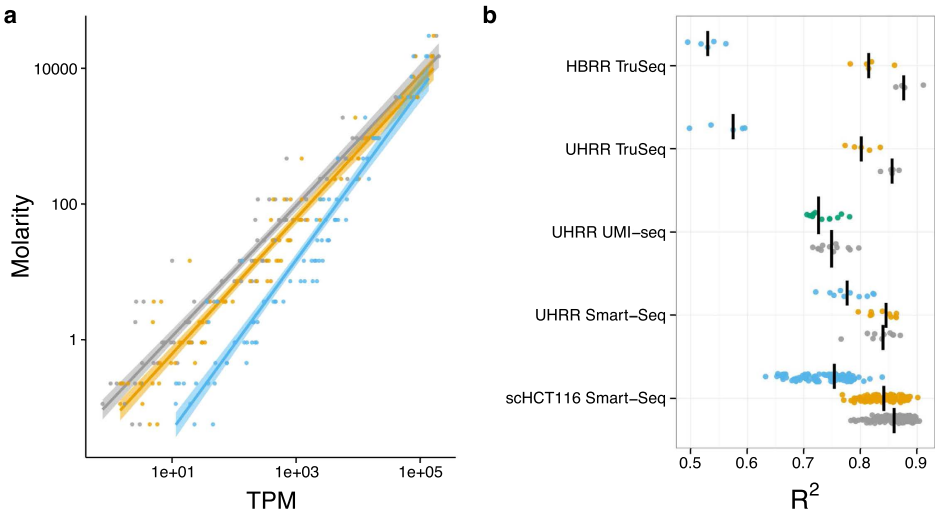


Figure 5. Removing duplicates does not improve the accuracy of expression quantification as measured using the ERCC spike-ins. Expression levels as quantified in transcripts per million reads (TPM) are a good predictor of the concentrations of the ERCC spike-ins. The log-linear fit of TPM vs. Molarity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

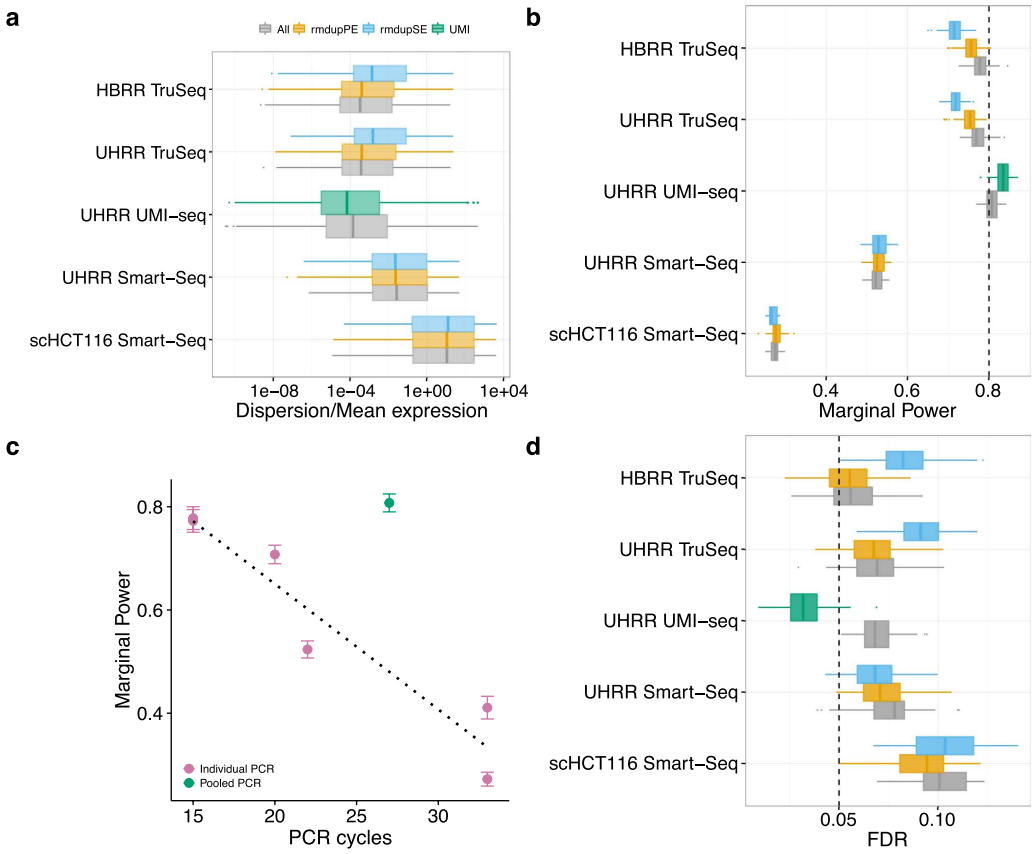


Figure 6. Duplicate removal has little influence on the power and FDR to detect DE-genes in comparison to the library preparation method. We estimated the distributions of mean expression and dispersion across genes for each dataset using DESeq2 after downsampling the datasets to 3 or 1 million reads.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

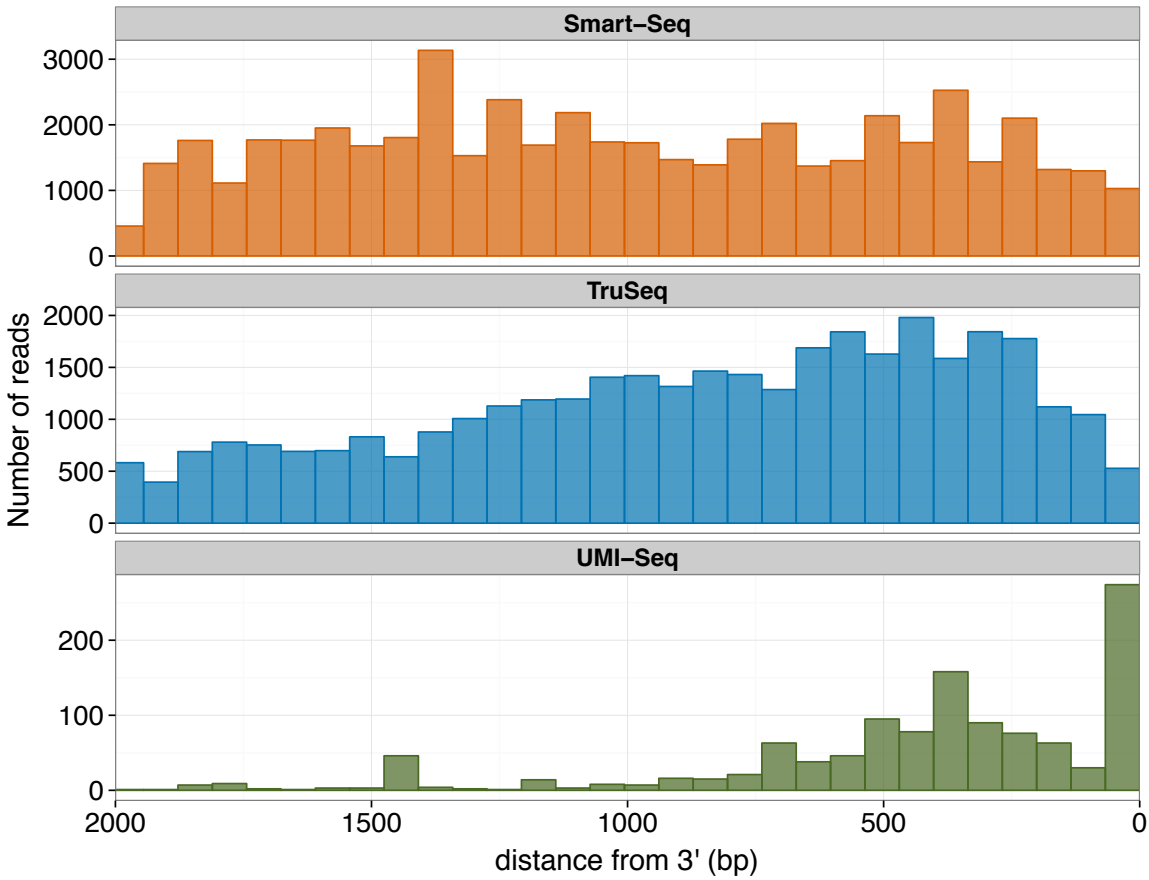


Figure S1: 3' bias in fragmentation site is prominent in UMI-seq. The histogram showing distance of the fragmentation site from 3' end of the gene measured from ERCC spike-ins of length $\sim 2kb$. Colors represent library preparation methods, 'blue' - Smart-Seq, 'orange' - TruSeq, 'green' - UMI-seq.

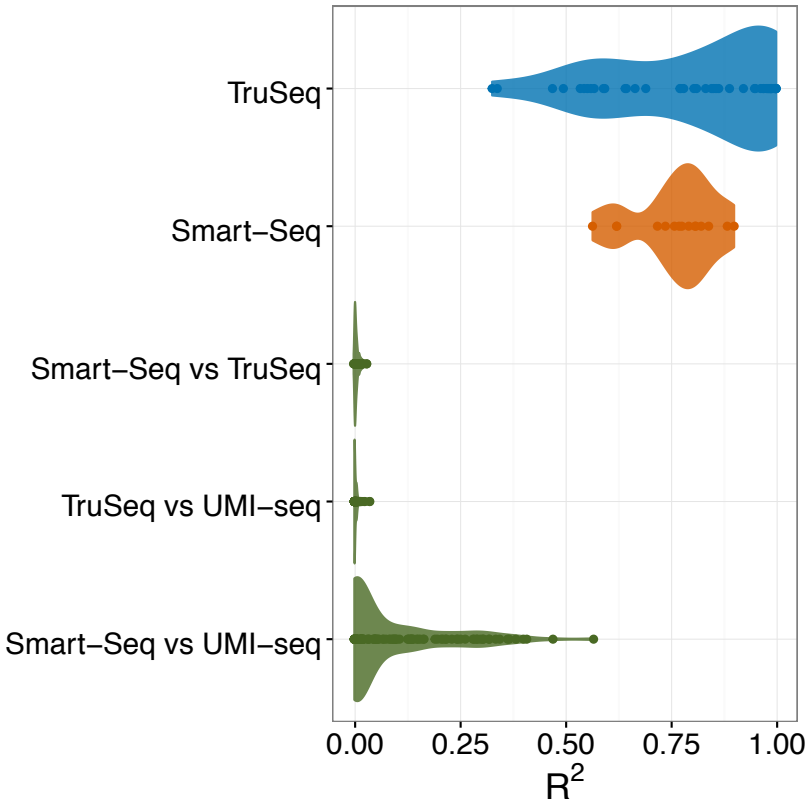


Figure S2: The fragmentation patterns of the most 3' 600bp of ERCCs are relatively reproducible between Smart-Seq and UMI-seq. Violin plots of the adjusted R^2 from a linear model between fraction of 5'read ends from different samples. The adjusted R^2 are calculated considering full length for Smart-Seq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

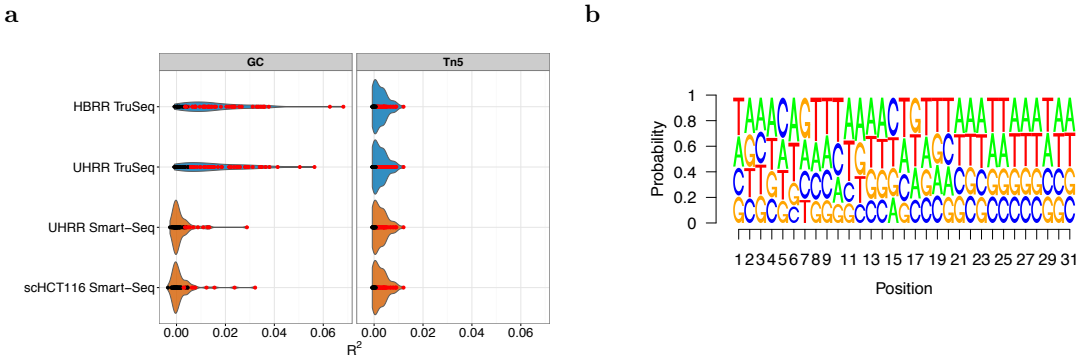


Figure S3: Fragmentation does not appear to have a cutting site preference. Colors of the violin plots represent library preparation methods, 'blue - Smart-Seq, 'orange' - TruSeq and dots are colored by the significance of the fit where 'red' - pvalue ≤ 0.05 and 'black' - pvalue > 0.05. **a)** The left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

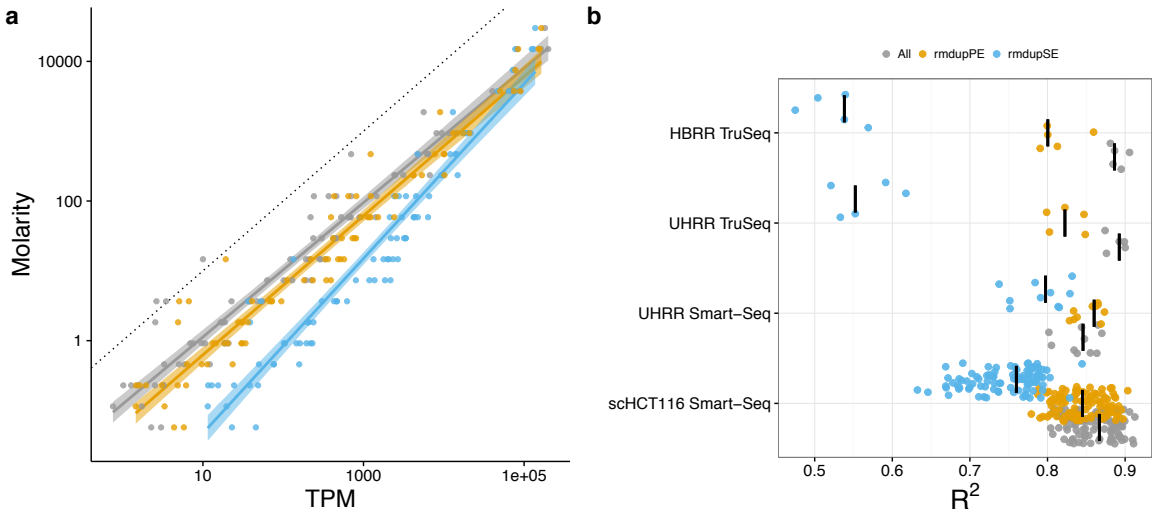


Figure S4: Removing duplicates does not improve the accuracy of expression quantification as measured using the ERCC spike-ins. Expression levels as quantified in transcripts per million reads (TPM) are considered to be good measure of ERCC spike ins. However, other factors like capture and sequencing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

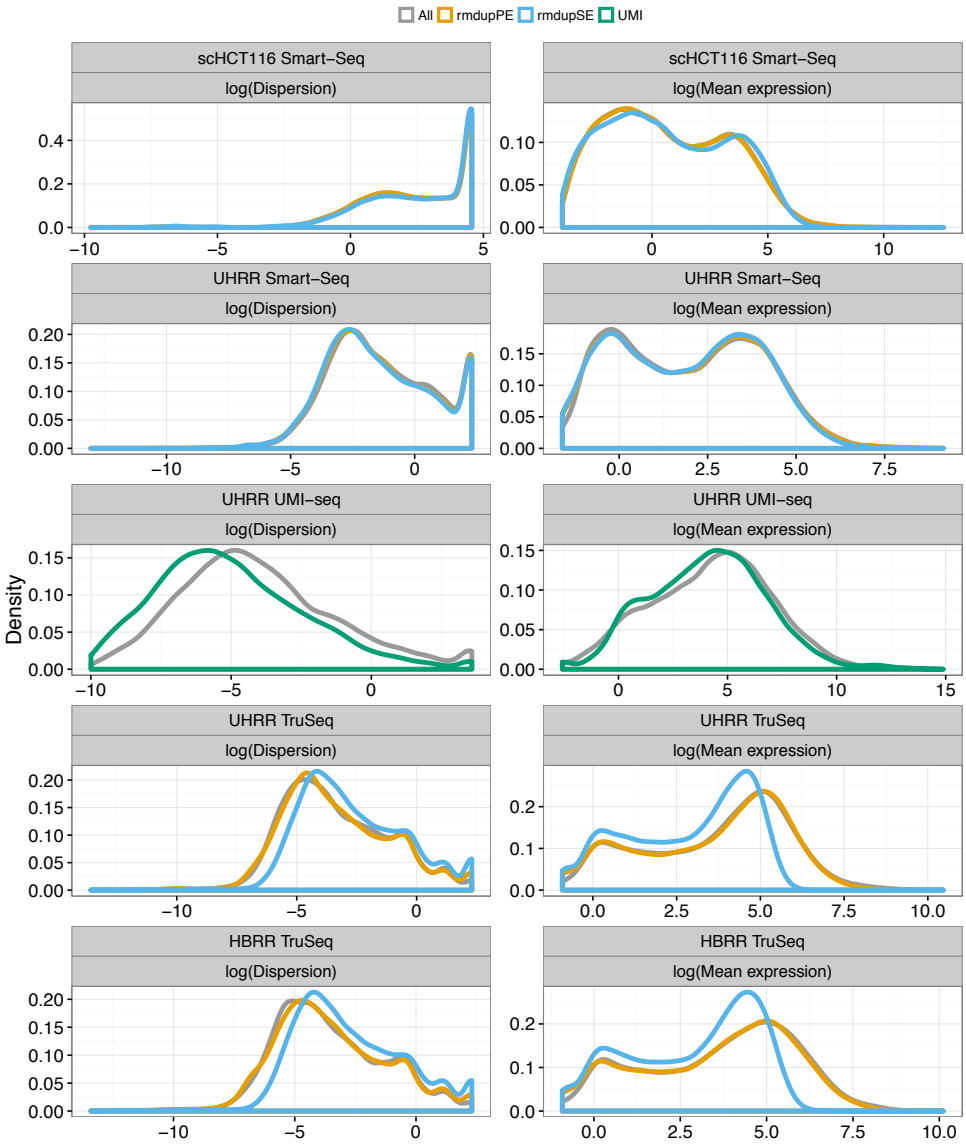


Figure S5: Empirical mean and dispersion distributions are used to estimate power to detect differential expression. The left panel shows density plot of log(dispersion) and the right panel the log(mean baseline expression) measured by DESeq2 for each study. Different duplicates treatments are represented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

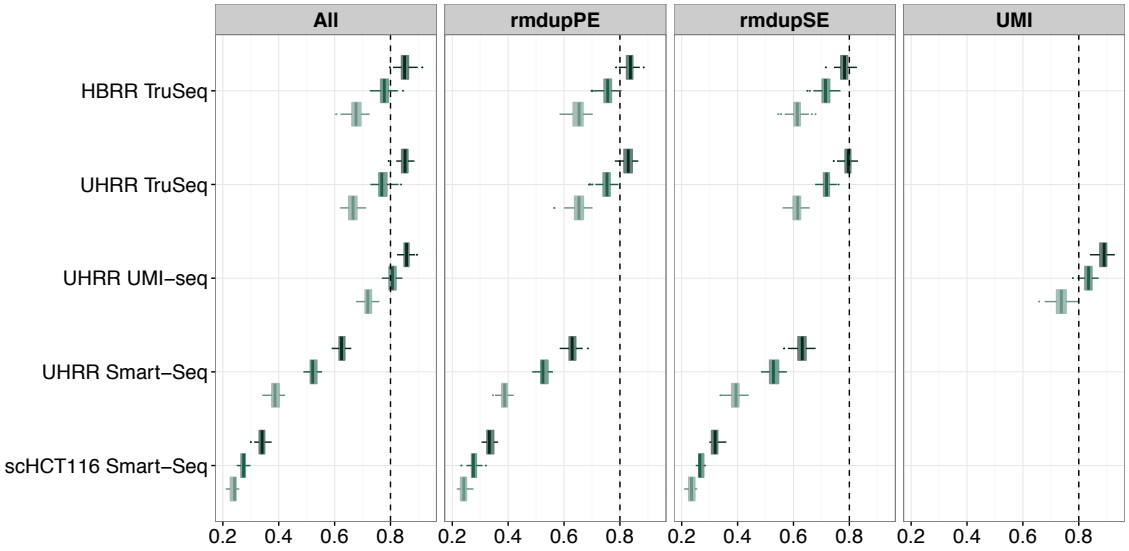


Figure S6: Power to detect differential expression increases with increased sample size. The boxplot shows marginal power to detect 0.5 log2foldchange at 5% nominal FDR for different sample sizes. Colors gradient from light to dark represent sample sizes 3,6 and 12 for the bulk and 30,45 and 90 for the 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 $\sim$ 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 $\leq$ 0.05

Table S3: Summary of power analysis

Study name	Sample size	Mean FDR	Marginal power	Avg # of TD	Avg # of FD	FDC	DupType	PCRcycles	Amount(ug)
HBRR TruSeq	3	0.06	0.68	239.63	16.28	0.07	All	15	1.00
HBRR TruSeq	3	0.06	0.65	232.52	16.35	0.07	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

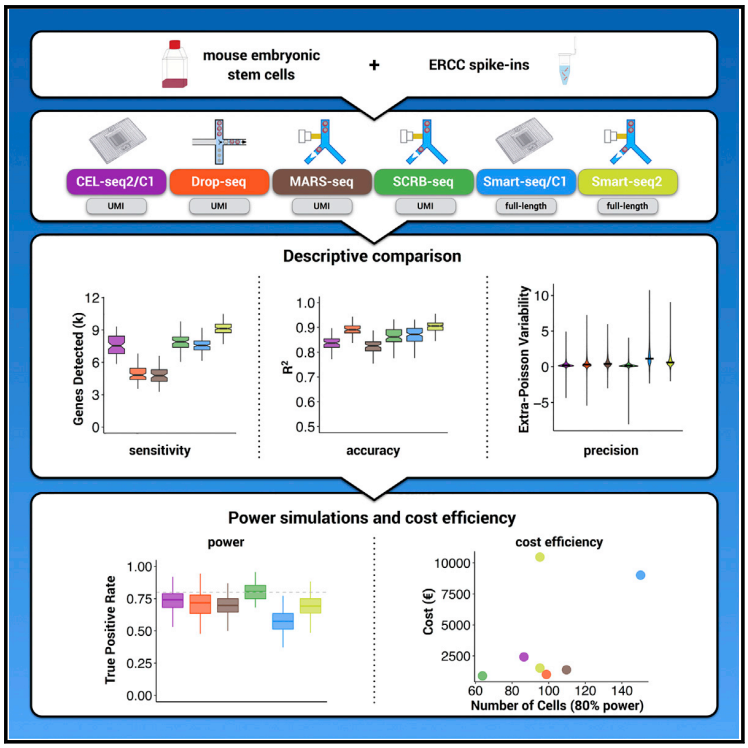
Comparative Analysis of Single-Cell RNA Sequencing Methods

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

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

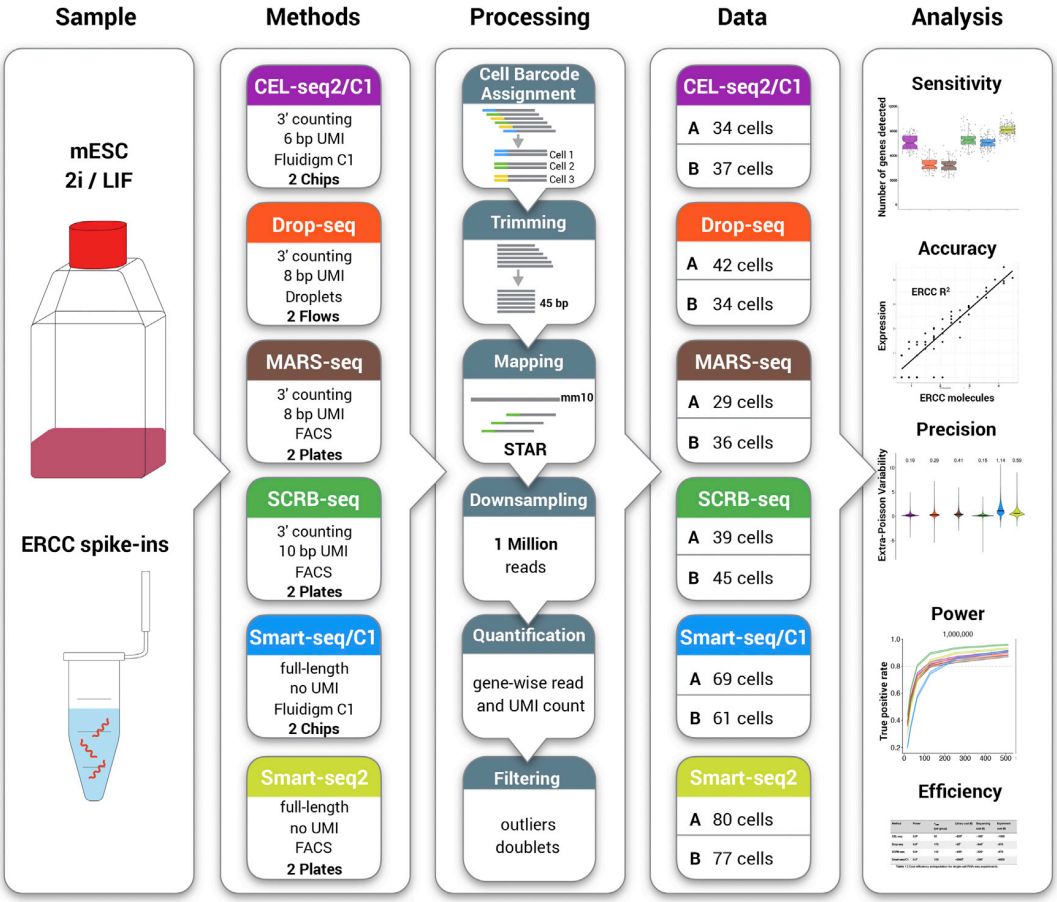


Figure 1. Schematic Overview of the Experimental and Computational Workflow

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.

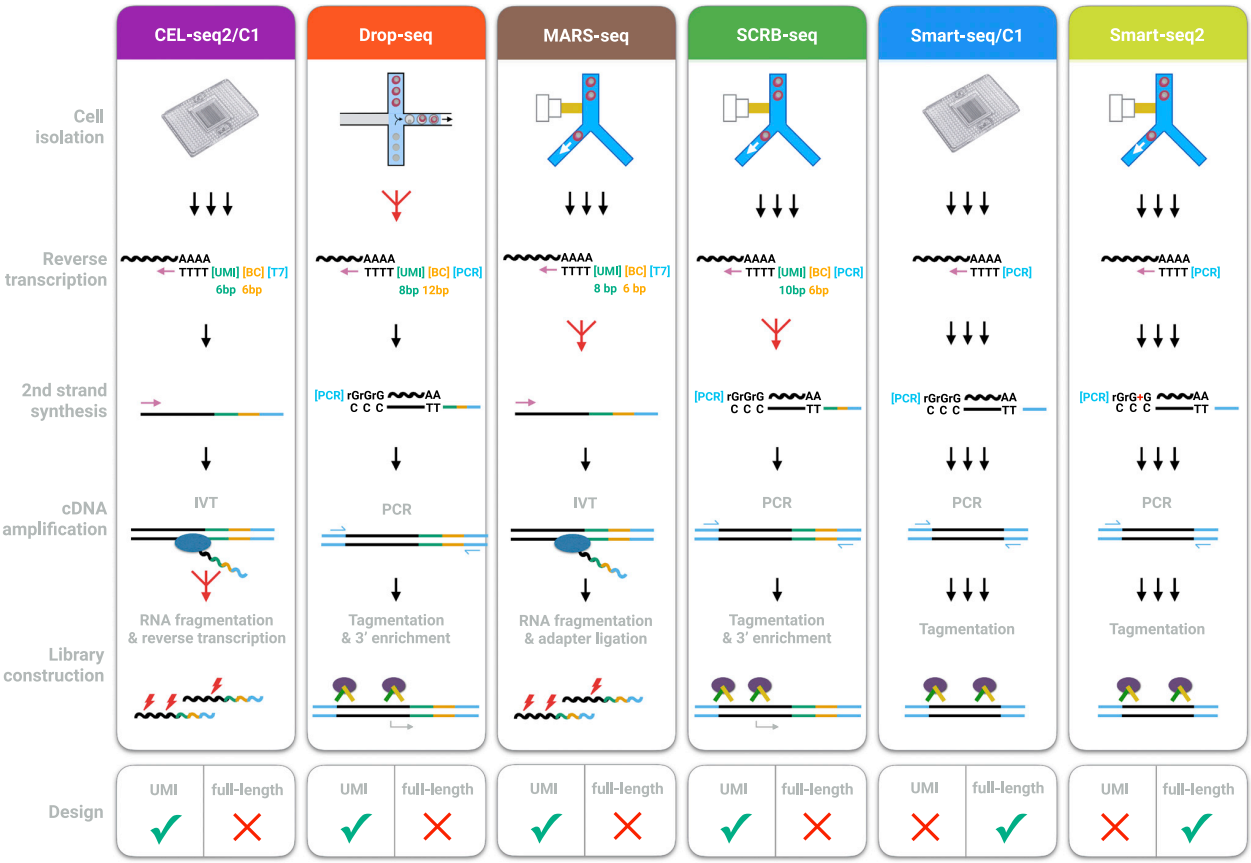


Figure 2. Schematic Overview of Library Preparation Steps
For details, see the text. See also Table S1.

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.

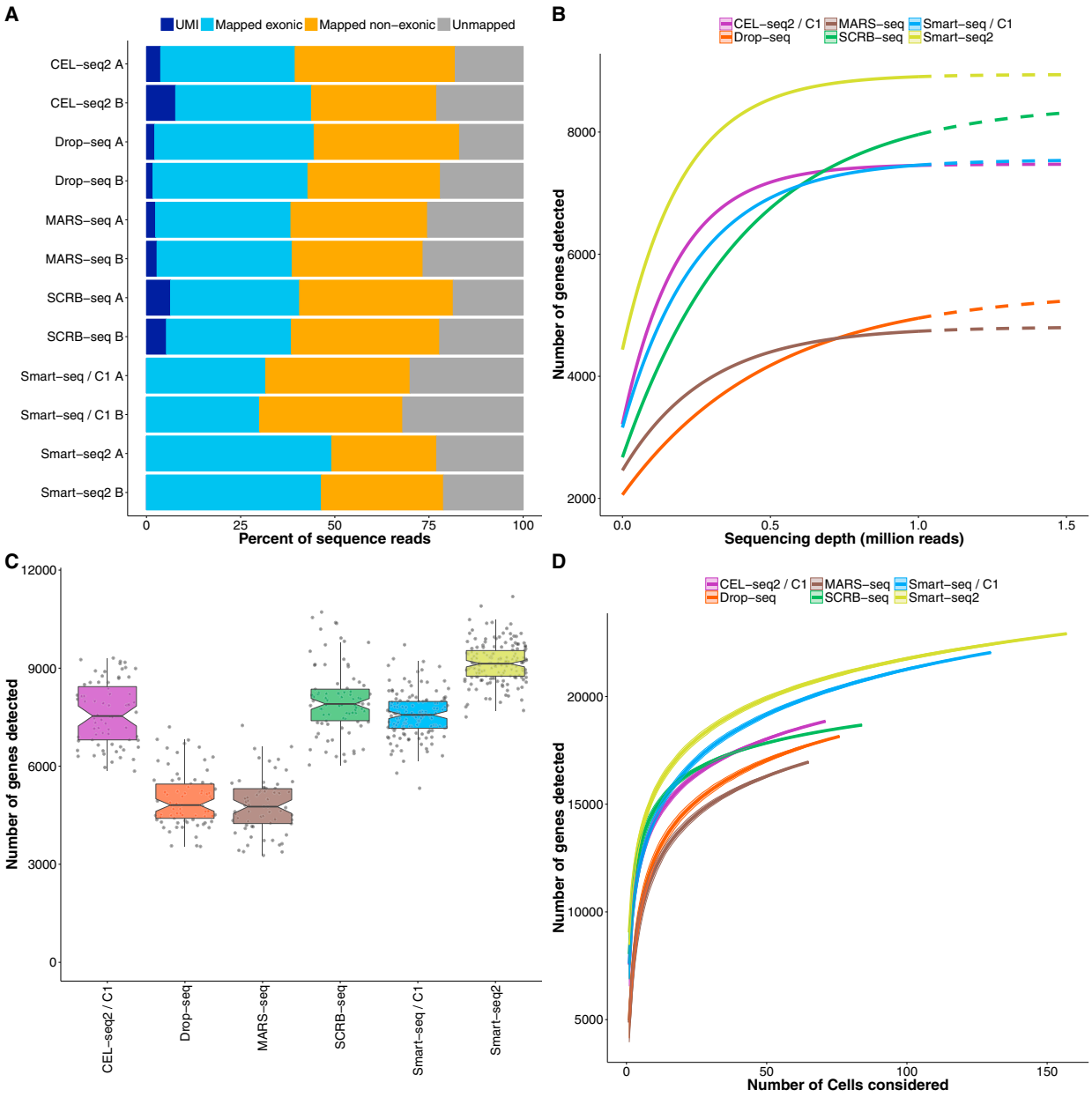


Figure 3. Sensitivity of scRNA-Seq Methods

(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

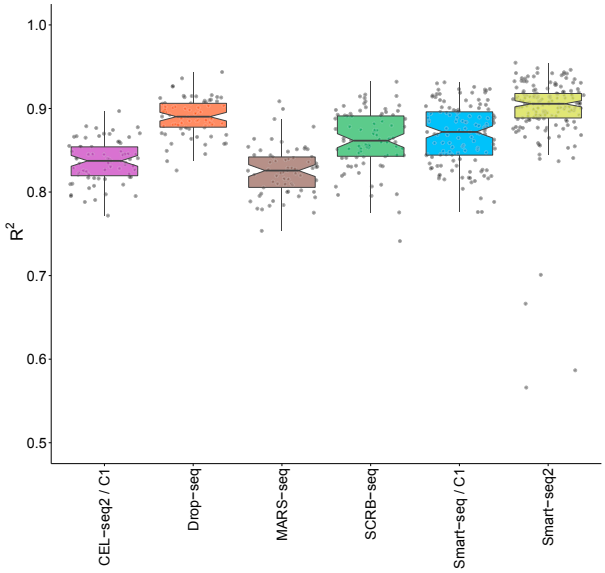


Figure 4. Accuracy of scRNA-Seq Methods

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

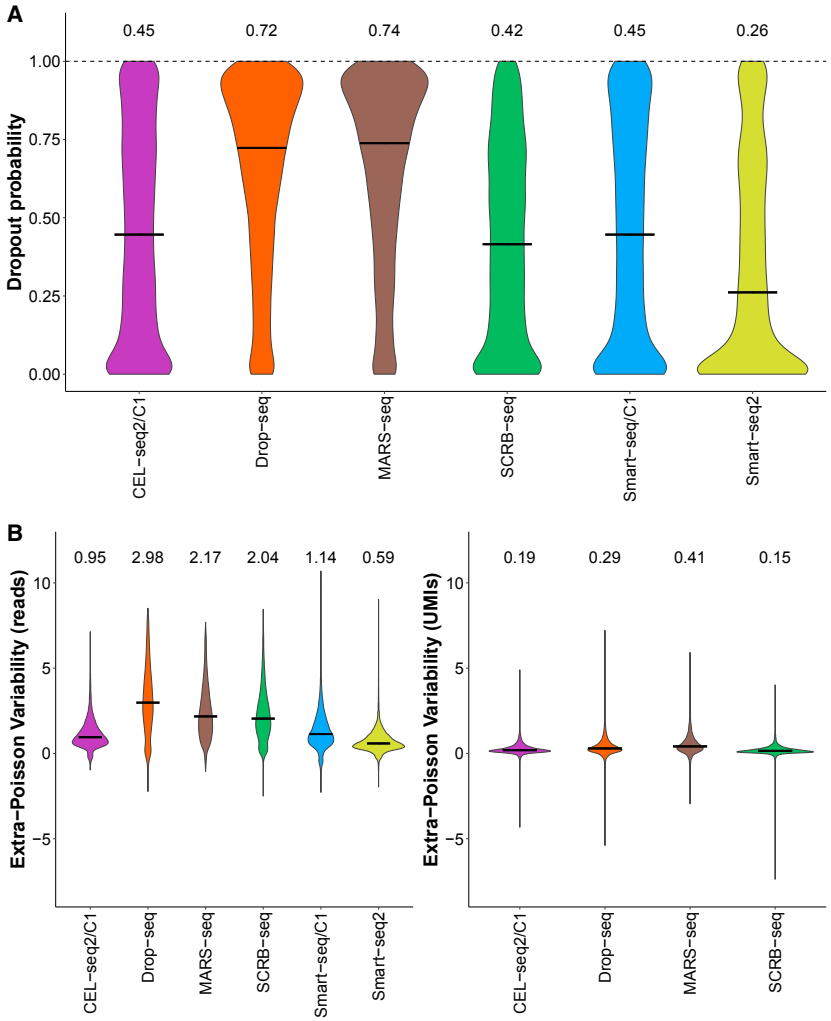


Figure 5. Precision of scRNA-Seq Methods

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.

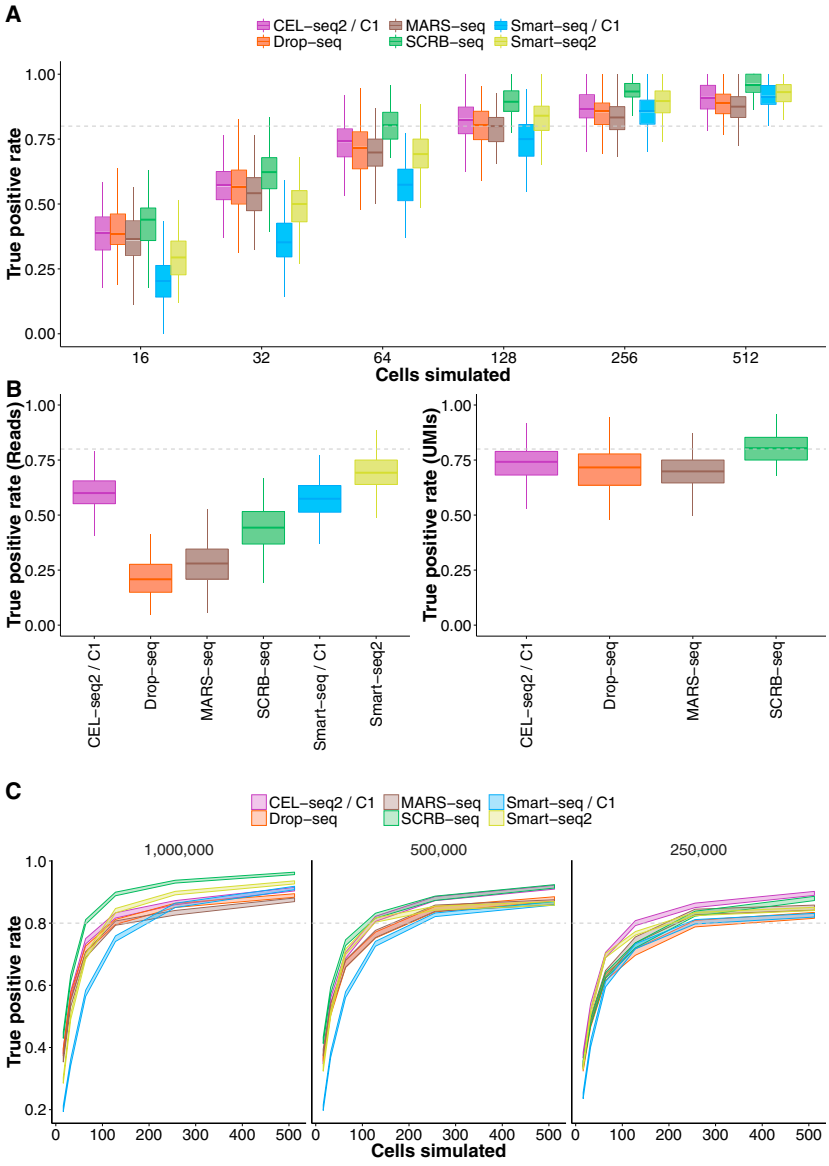


Figure 6. Power of scRNA-Seq Methods

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.

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.

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

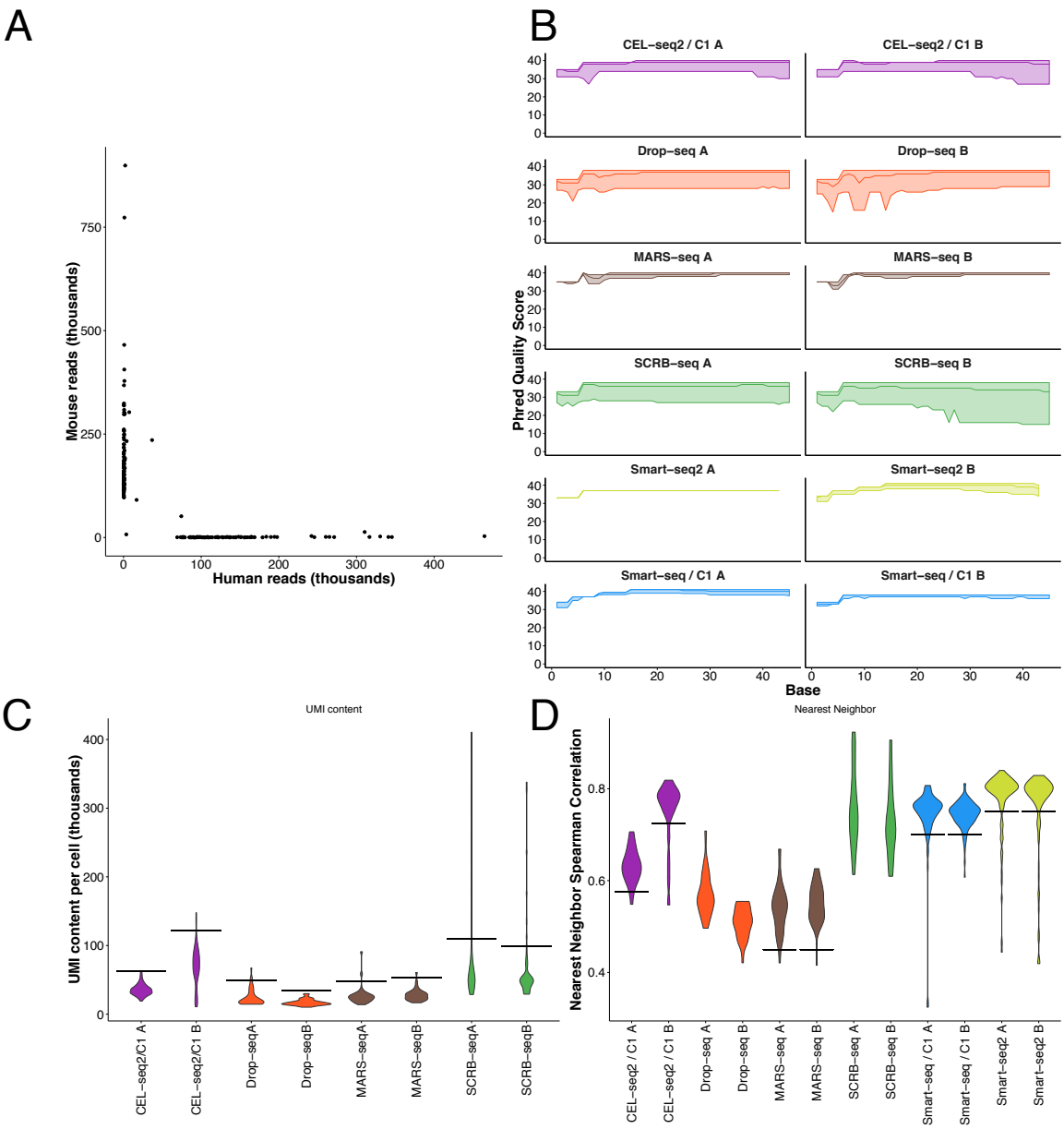


Figure S2

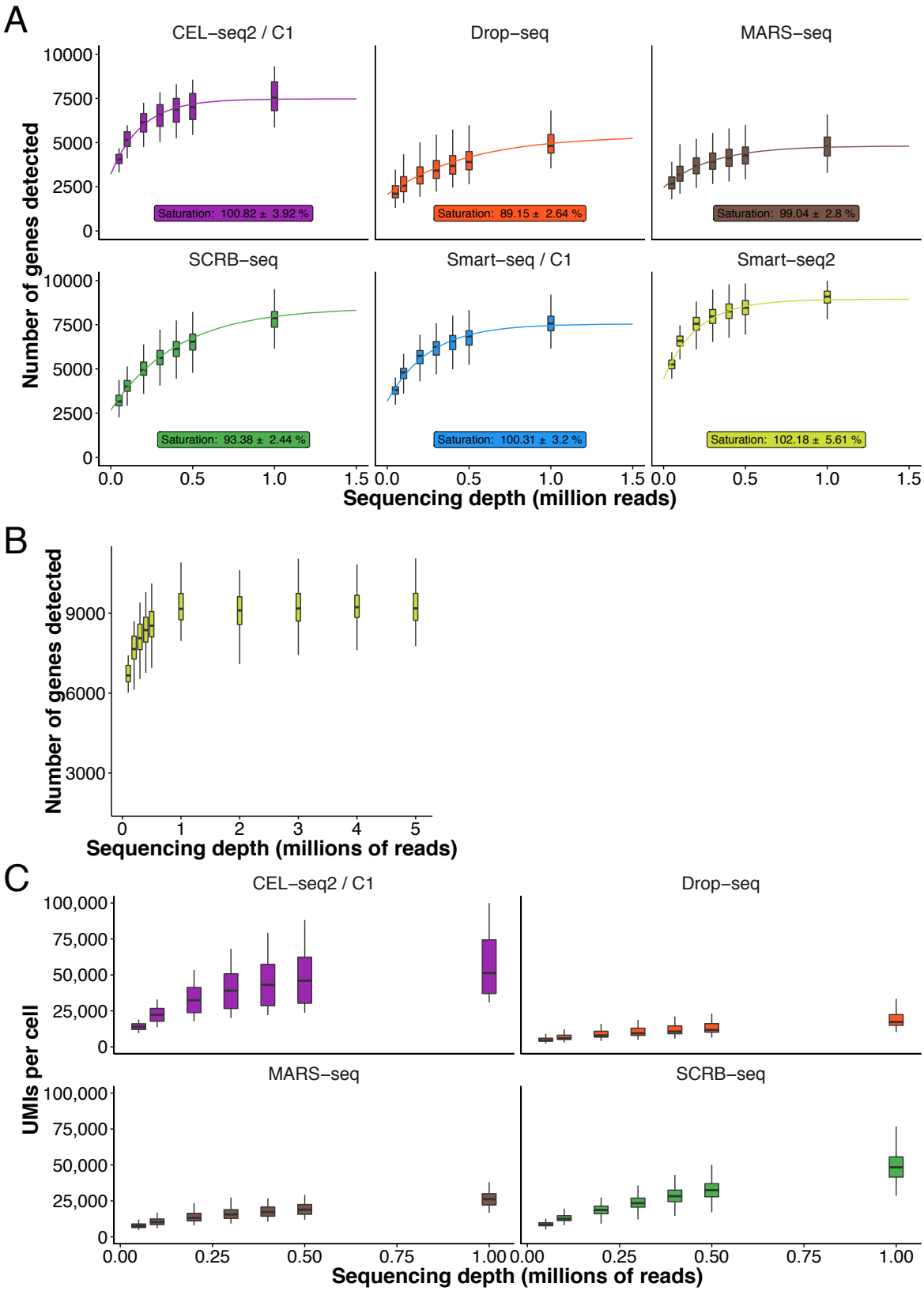

A
CEL−seq2 / C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq / C1
Smart−seq2
Saturation: 100.82 ± 3.92 %
Saturation: 89.15 ± 2.64 %
Saturation: 99.04 ± 2.8 %
Saturation: 93.38 ± 2.44 %
Saturation: 100.31 ± 3.2 %
Saturation: 102.18 ± 5.61 %
Number of genes detected
Sequencing depth (million reads)
B
Number of genes detected
Sequencing depth (millions of reads)
C
CEL−seq2 / C1
Drop−seq
MARS−seq
SCRB−seq
UMIs per cell
Sequencing depth (millions of reads)

Figure S3

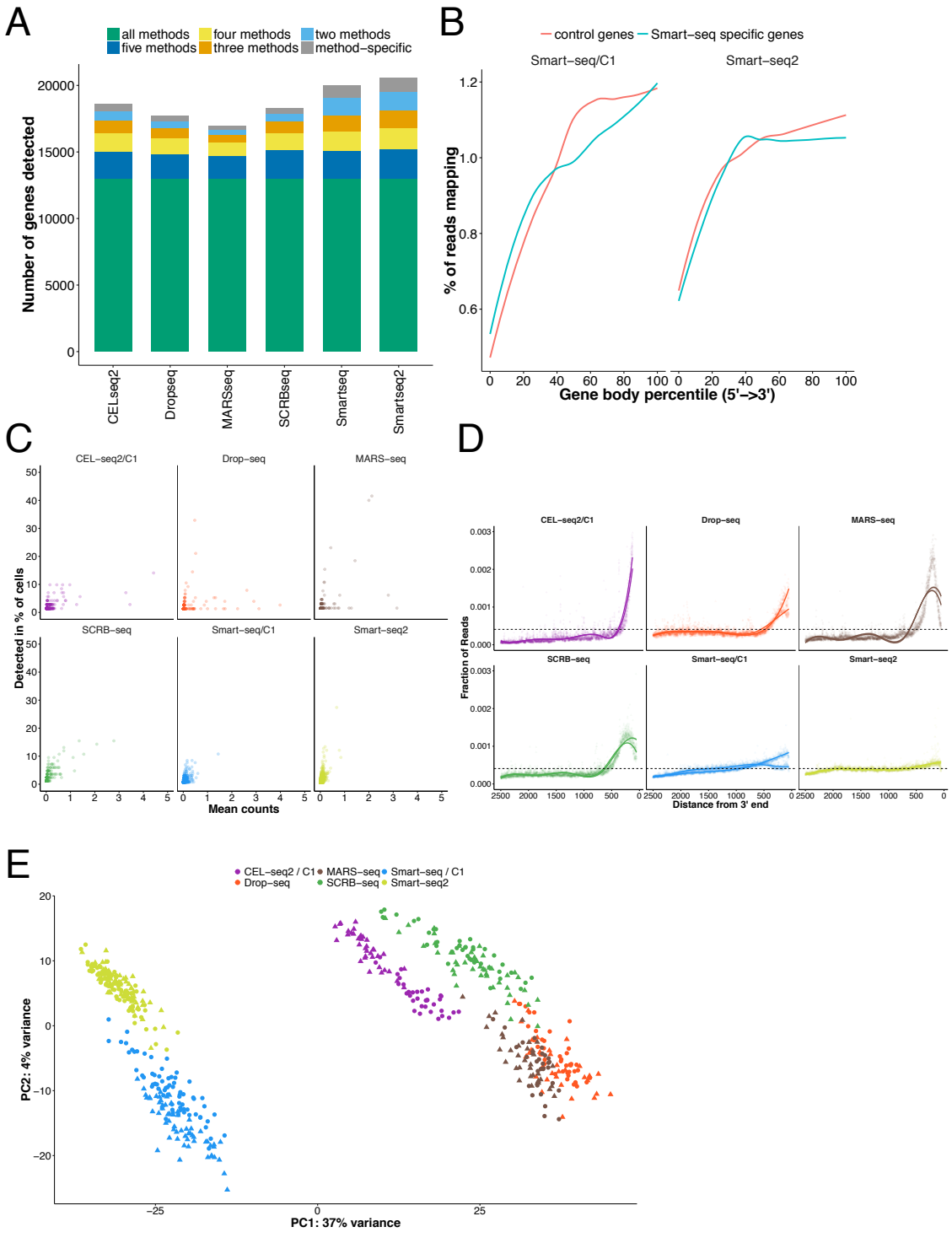

A
all methods
five methods
four methods
three methods
two methods
method−specific
Number of genes detected
20000
15000
10000
5000
0
CELseq2
Dropseq
MARSseq
SCRBseq
Smartseq
Smartseq2
B
control genes
Smart−seq specific genes
Smart−seq/C1
Smart−seq2
% of reads mapping
1.2
1.0
0.8
0.6
0
20
40
60
80
100
Gene body percentile (5'−>3')
C
CEL−seq2/C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq/C1
Smart−seq2
Detected in % of cells
Mean counts
D
CEL−seq2/C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq/C1
Smart−seq2
Fraction of Reads
0.003
0.002
0.001
0.000
2500
2000
1500
1000
500
0
Distance from 3' end
E
CEL−seq2 / C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq / C1
Smart−seq2
PC2: 4% variance
PC1: 37% variance
20
10
0
−10
−20
−25
0
25

Figure S4

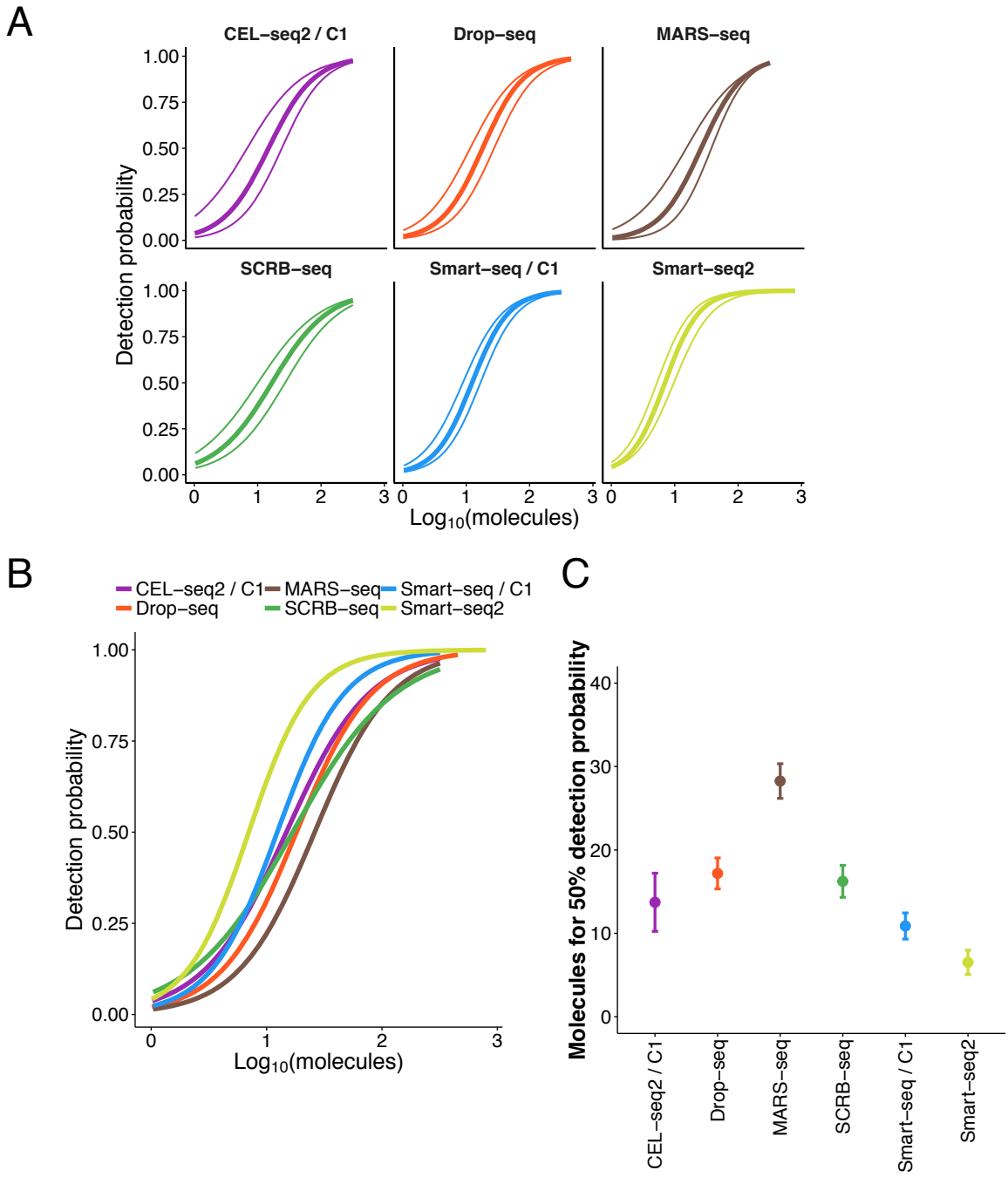

A
CEL−seq2 / C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq / C1
Smart−seq2
Detection probability
1.00
0.75
0.50
0.25
0.00
0
1
2
3
Log10(molecules)
B
CEL−seq2 / C1
MARS−seq
Smart−seq / C1
Drop−seq
SCRB−seq
Smart−seq2
Detection probability
Log10(molecules)
C
Molecules for 50% detection probability
40
30
20
10
0
CEL−seq2 / C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq / C1
Smart−seq2

Figure S5

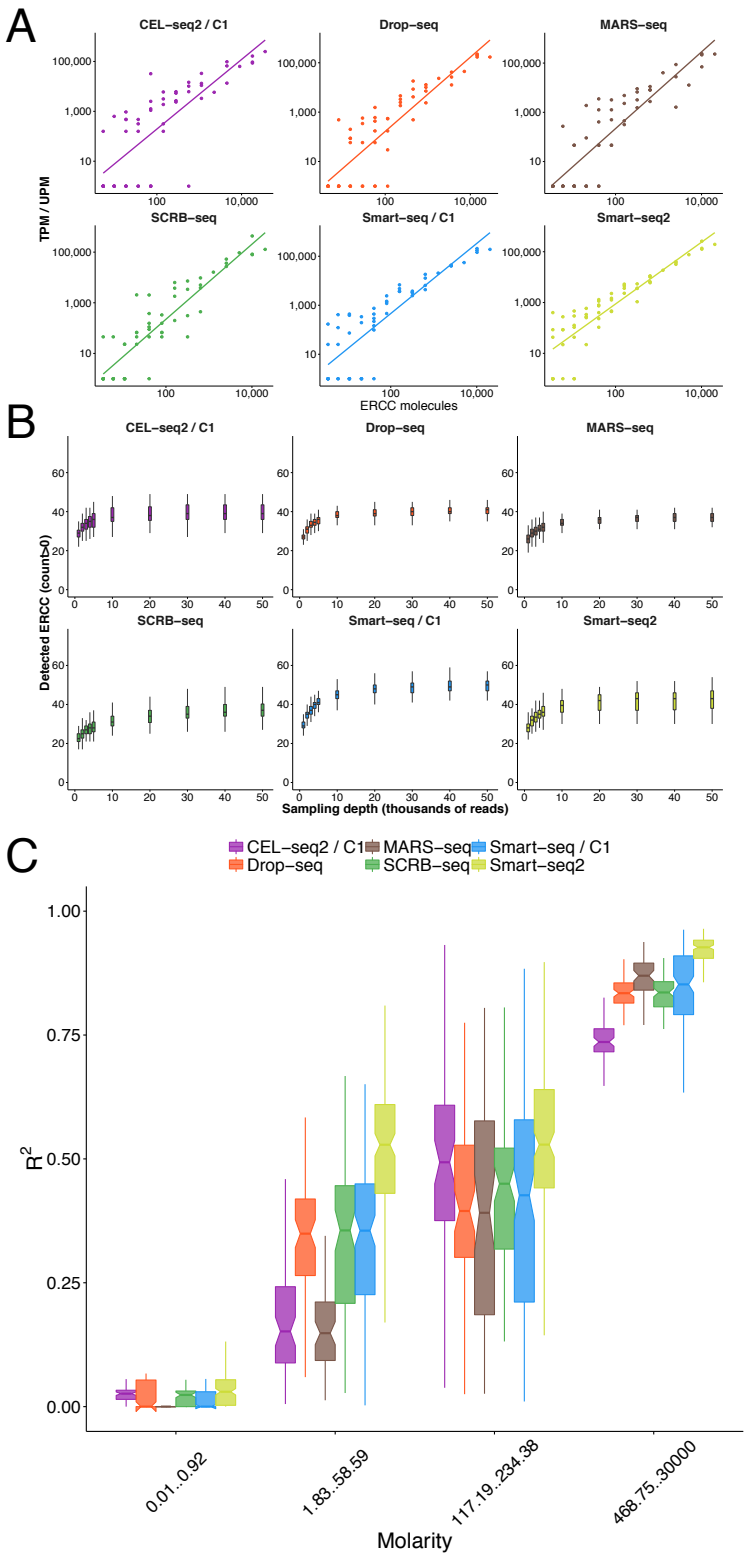

A
CEL−seq2 / C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq / C1
Smart−seq2
TPM / UPM
ERCC molecules
B
Detected ERCC (count>0)
Sampling depth (thousands of reads)
C
CEL−seq2 / C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq / C1
Smart−seq2
R2
Molarity
0.01..0.92
1.83..58.59
117.19..234.38
468.75..30000

Figure S6

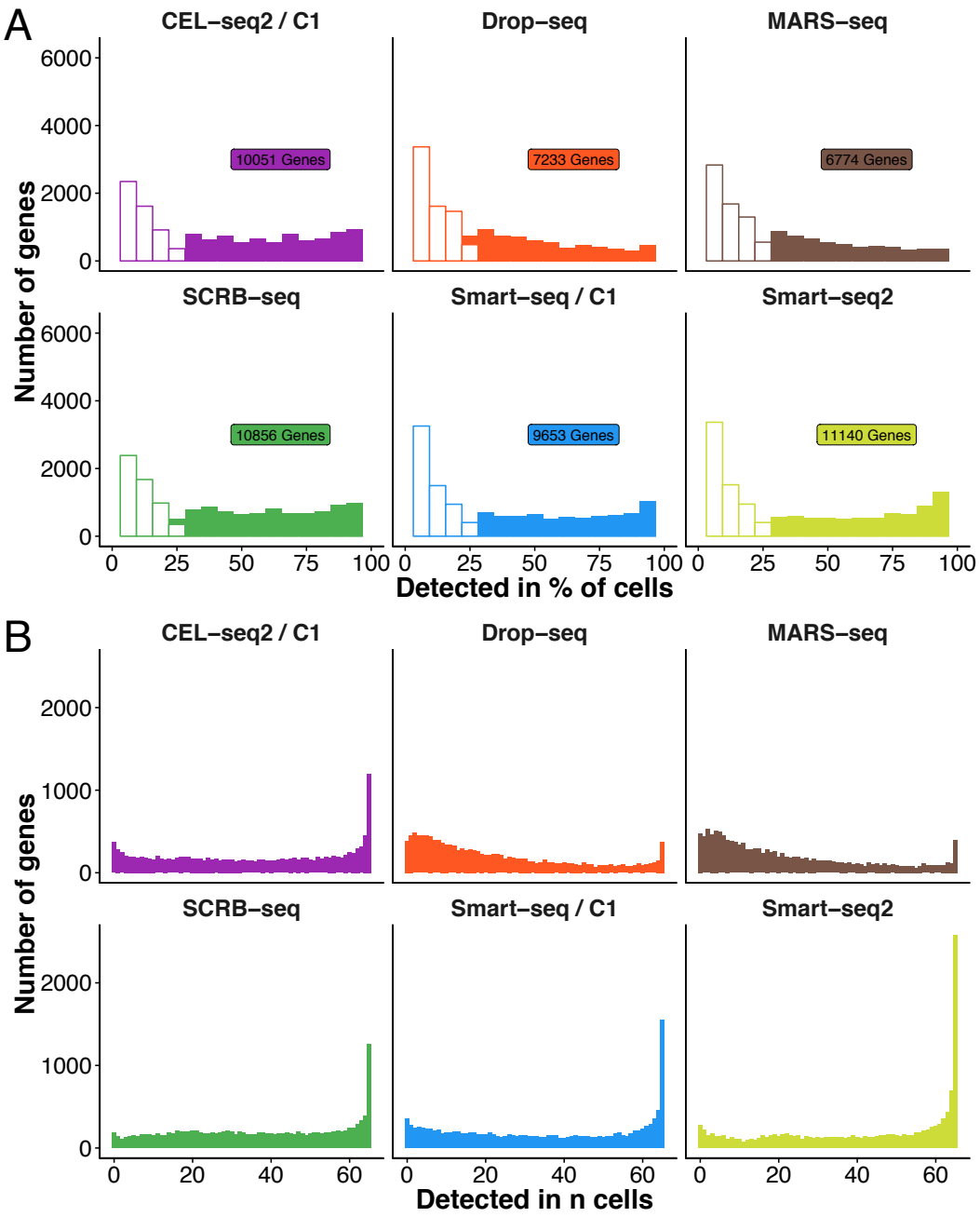

A
CEL−seq2 / C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq / C1
Smart−seq2
10051 Genes
7233 Genes
6774 Genes
10856 Genes
9653 Genes
11140 Genes
Number of genes
Detected in % of cells
B
Detected in n cells

Figure S7

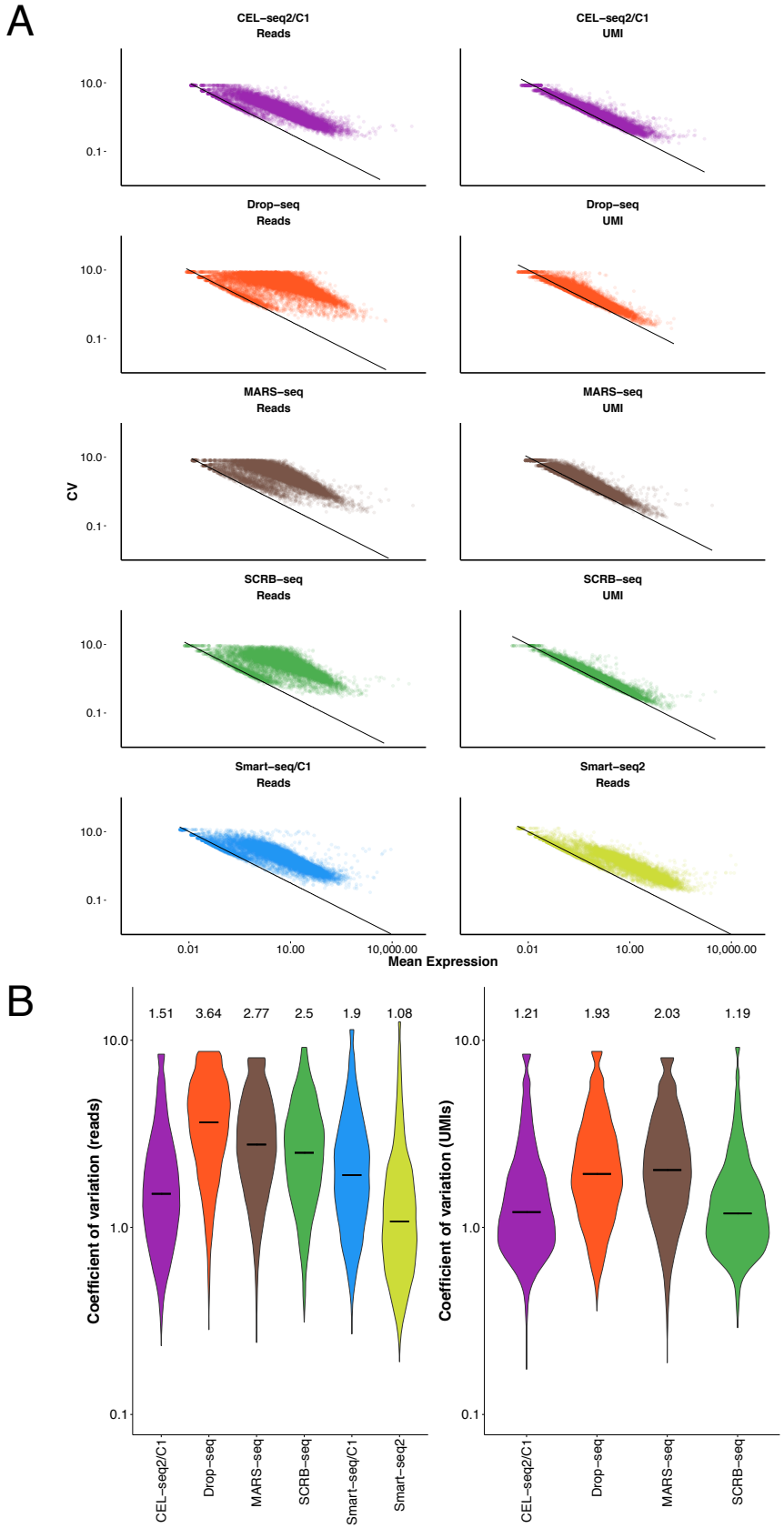
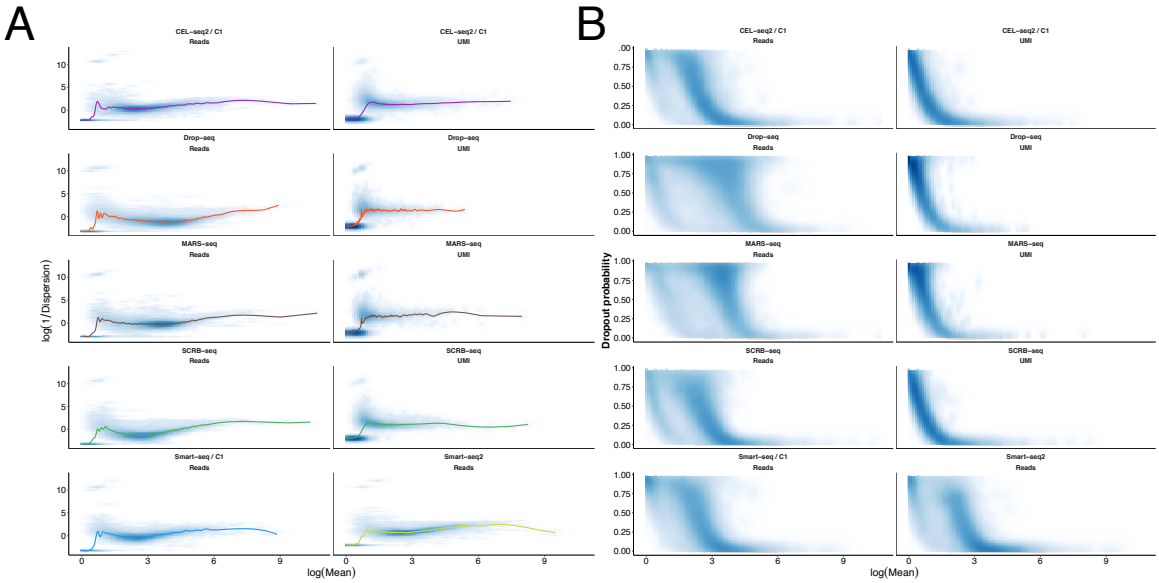
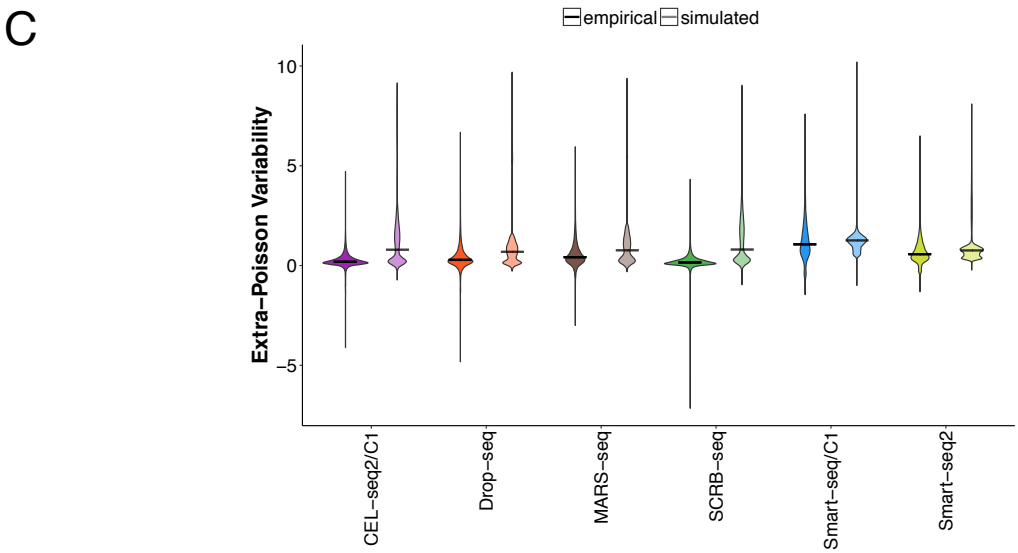

A
CEL−seq2/C1
Reads
CEL−seq2/C1
UMI
Drop−seq
Reads
Drop−seq
UMI
MARS−seq
Reads
MARS−seq
UMI
SCRB−seq
Reads
SCRB−seq
UMI
Smart−seq/C1
Reads
Smart−seq2
Reads
CV
10.0
0.1
0.01
10.00
10,000.00
Mean Expression
B
Coefficient of variation (reads)
Coefficient of variation (UMIs)
1.51
3.64
2.77
2.5
1.9
1.08
1.21
1.93
2.03
1.19
10.0
1.0
0.1
CEL−seq2/C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq/C1
Smart−seq2

Figure S8


A
B
CEL−seq2 / C1
Reads
UMI
Drop−seq
MARS−seq
SCRB−seq
Smart−seq / C1
Smart−seq2
log(1/Dispersion)
Dropout probability
log(Mean)


C
empirical
simulated
Extra−Poisson Variability
CEL−seq2/C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq/C1
Smart−seq2

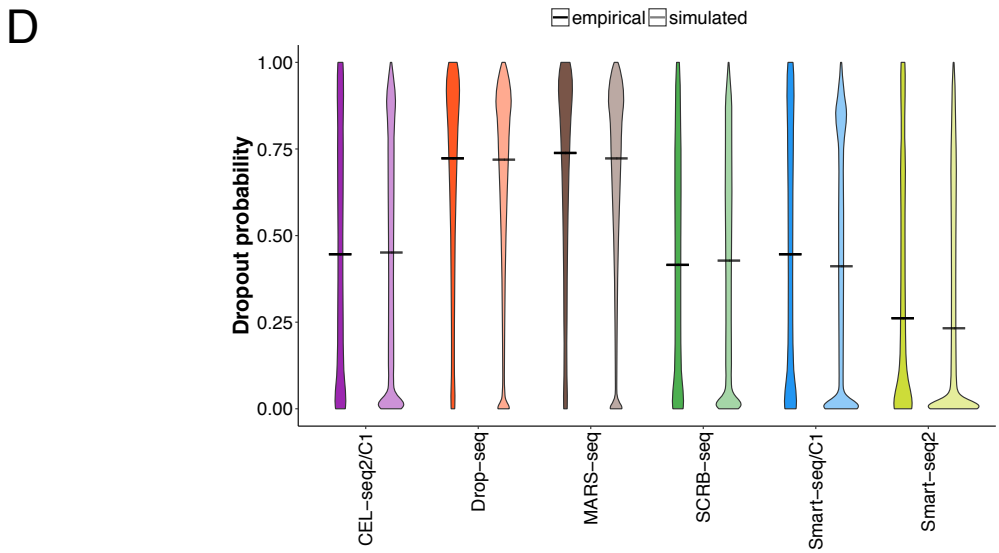

D
empirical
simulated
Dropout probability
CEL−seq2/C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq/C1
Smart−seq2

Figure S9

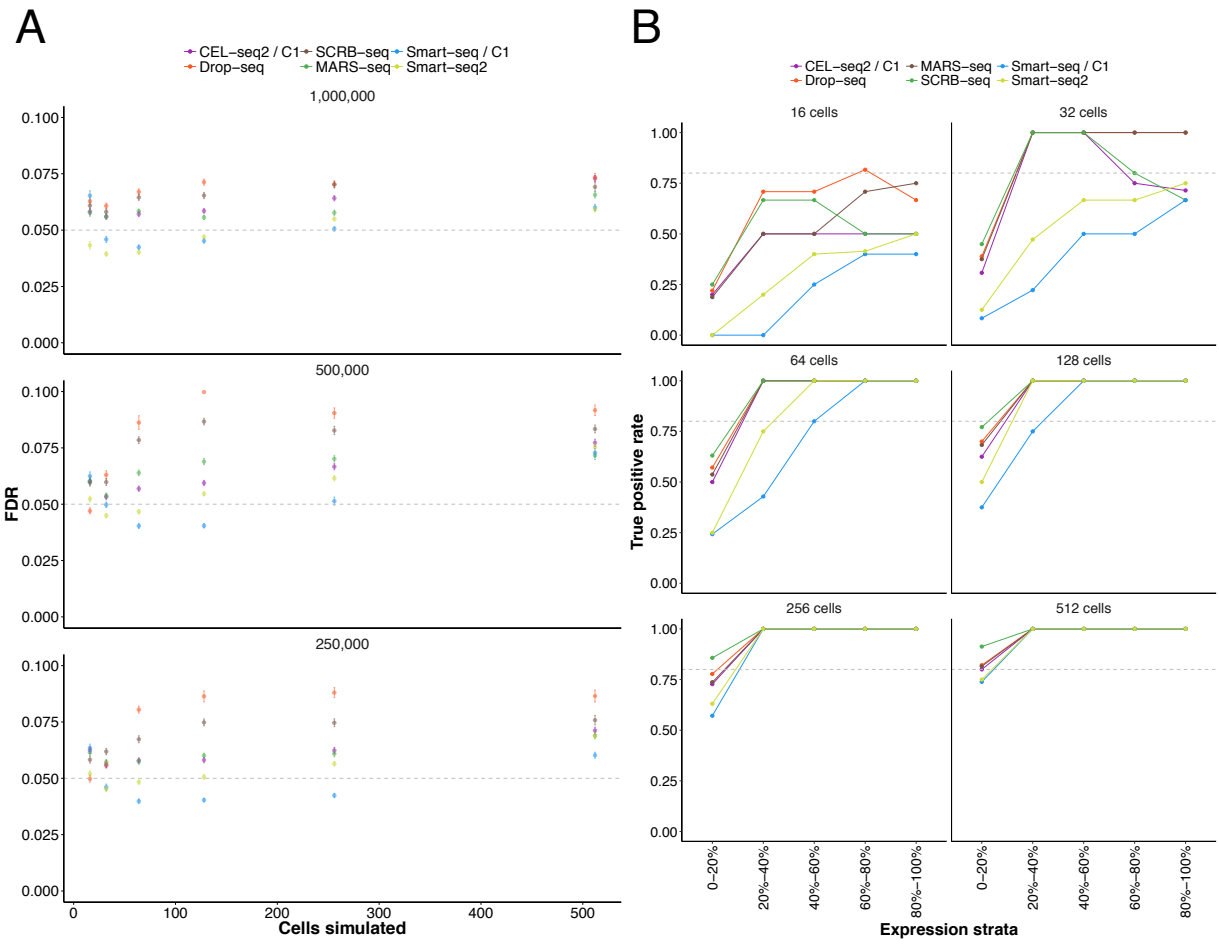
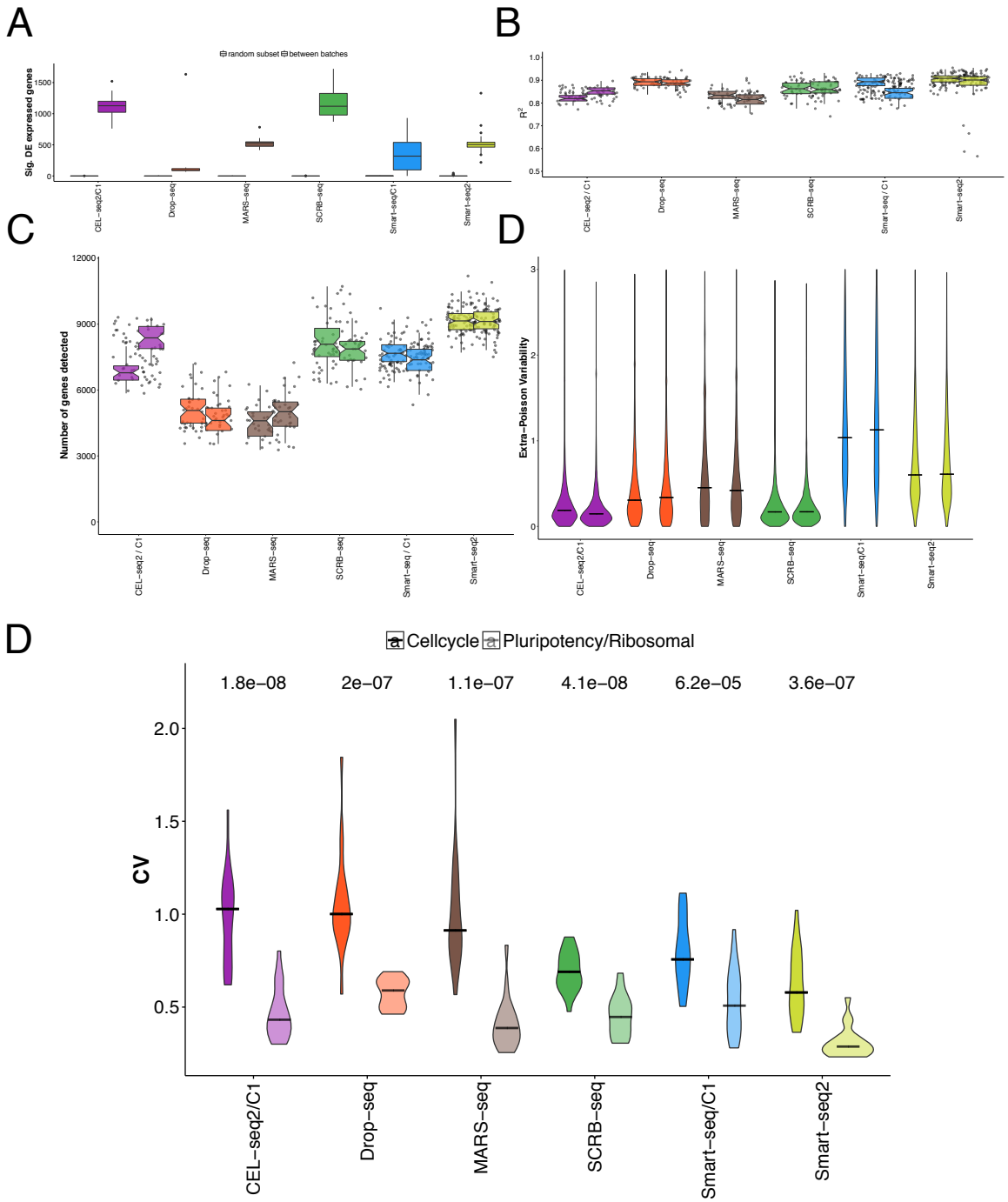

A
CEL−seq2 / C1
SCRB−seq
Smart−seq / C1
Drop−seq
MARS−seq
Smart−seq2
1,000,000
500,000
250,000
FDR
0.100
0.075
0.050
0.025
0.000
0
100
200
300
400
500
Cells simulated
B
16 cells
32 cells
64 cells
128 cells
256 cells
512 cells
True positive rate
1.00
0.75
0.50
0.25
0.00
0−20%
20%−40%
40%−60%
60%−80%
80%−100%
Expression strata

Figure S10


A
random subset between batches
Sig. DE expressed genes
CEL−seq2/C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq/C1
Smart−seq2
B
R^2
C
Number of genes detected
D
Extra−Poisson Variability

D
Cellcycle
Pluripotency/Ribosomal
1.8e−08
2e−07
1.1e−07
4.1e−08
6.2e−05
3.6e−07
CV
CEL−seq2/C1
Drop−seq
MARS−seq
SCRB−seq
Smart−seq/C1
Smart−seq2

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

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

Bioinformatics, 2017, 1–3
doi: 10.1093/bioinformatics/btx435
Advance Access Publication Date: 11 July 2017
Applications Note

OXFORD

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.

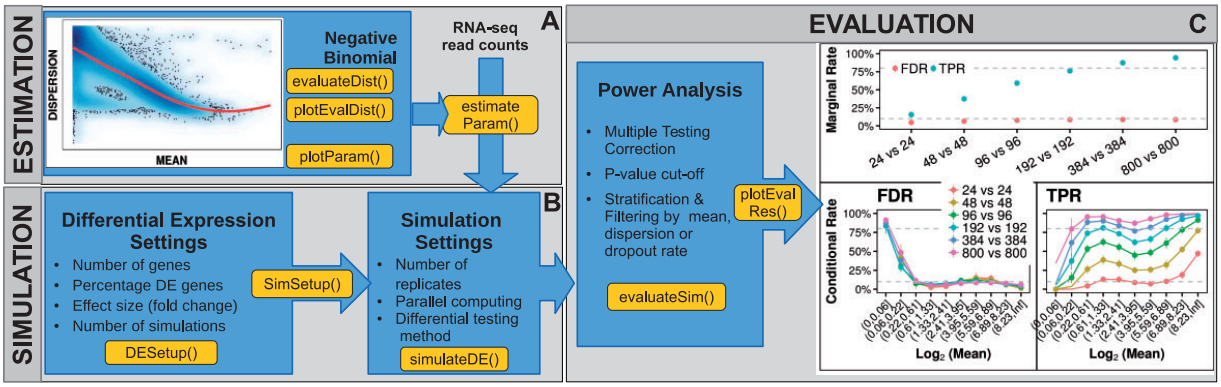


Fig. 1. powsimR schematic overview. (**A**) The mean-dispersion relationship is estimated from RNA-seq data, which can be either single cell or bulk data. The user can provide their own count tables or one of our five example datasets and choose whether to fit a negative binomial or a zero-inflated negative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

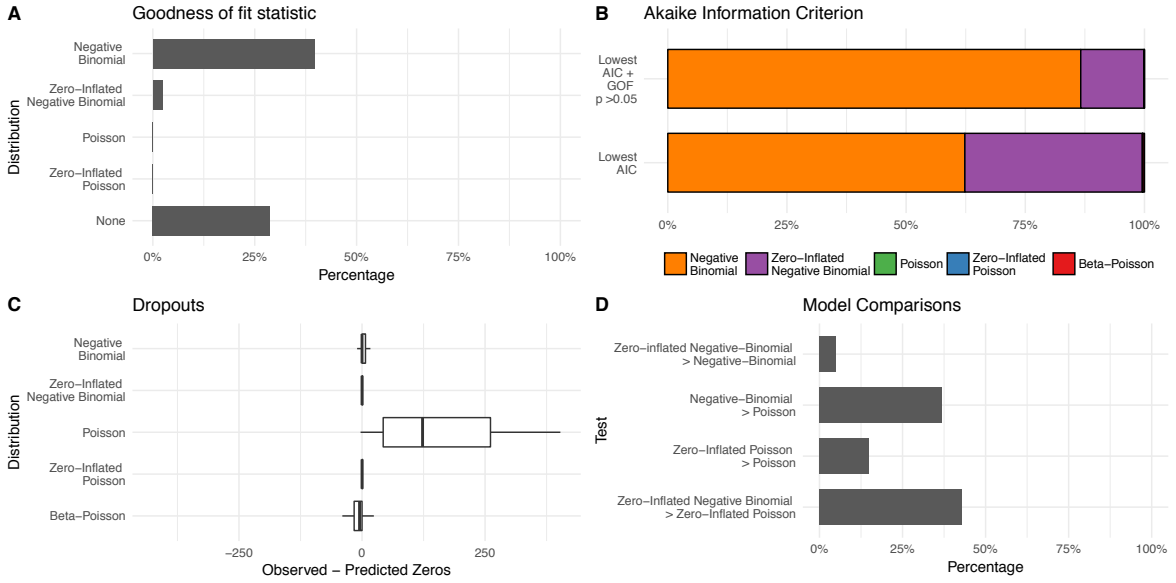


Figure S1: A) Goodness of fit of the model per gene assessed with a Chi-square test based on residual deviance and degrees of freedom. B) The fraction of genes for which the respective distribution has the lowest AIC and additionally the distribution with the lowest AIC as well as not rejected by the 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.

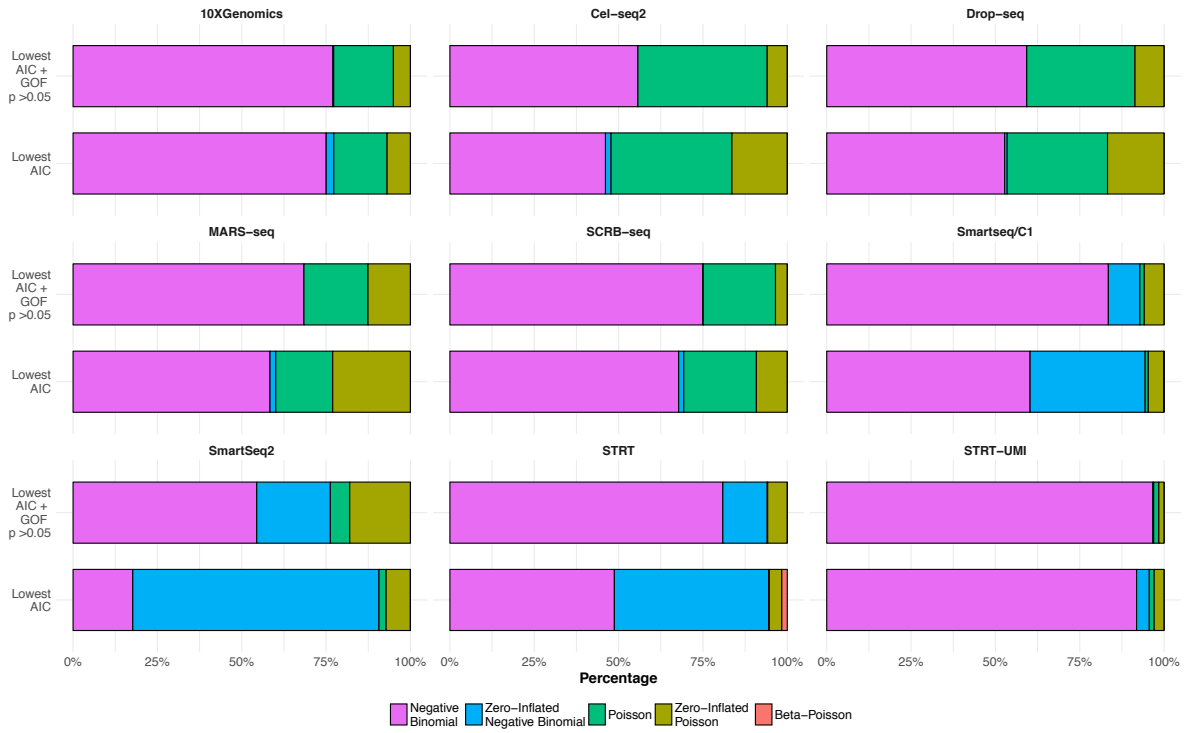


Figure S2: The negative binomial gives the best fit for the majority of genes (i.e. lowest AIC) for all UMI datasets. For protocols that do not account for PCR duplicates, the zero-inflated negative binomial often has a lower AIC, however this is mainly due to genes that cannot be fitted very well in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

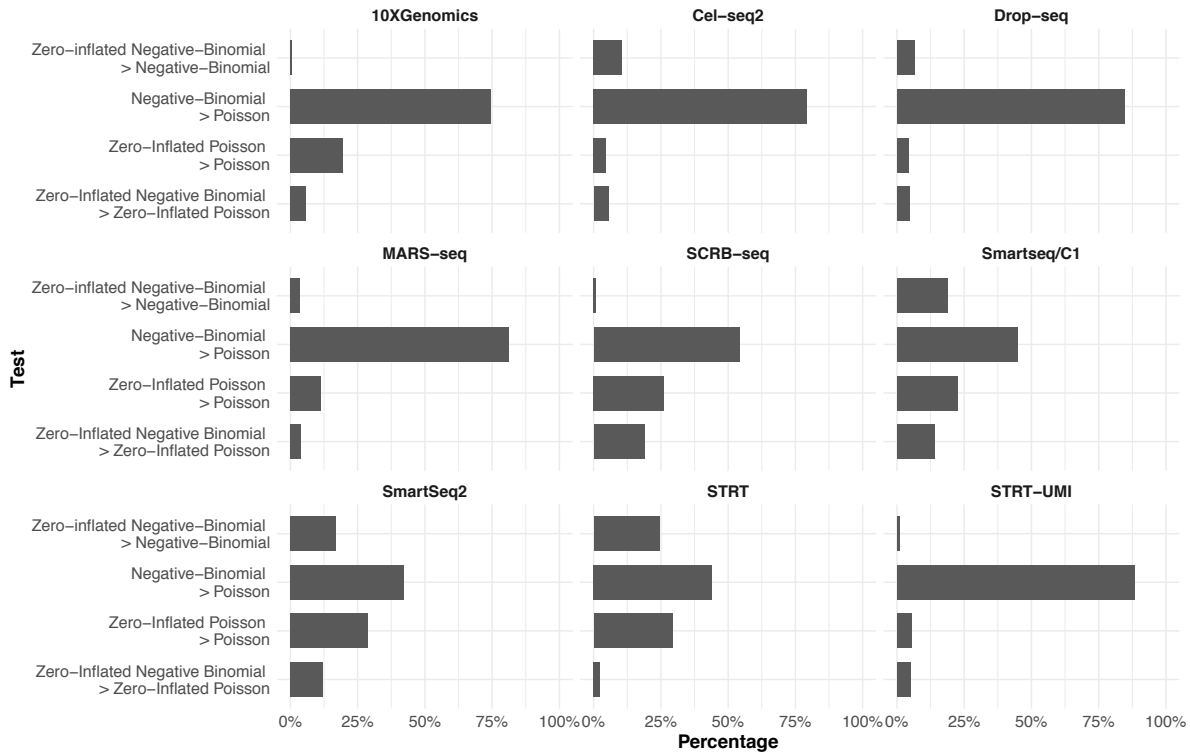


Figure S3: Model assessment per gene based on likelihood ratio test for nested models and Vuong test for non-nested models shows that zero-inflated negative binomial significantly improves the fit for maximally 25% of the genes (STRT protocol).

$\mu_i < \mu_{DP5}$ 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).

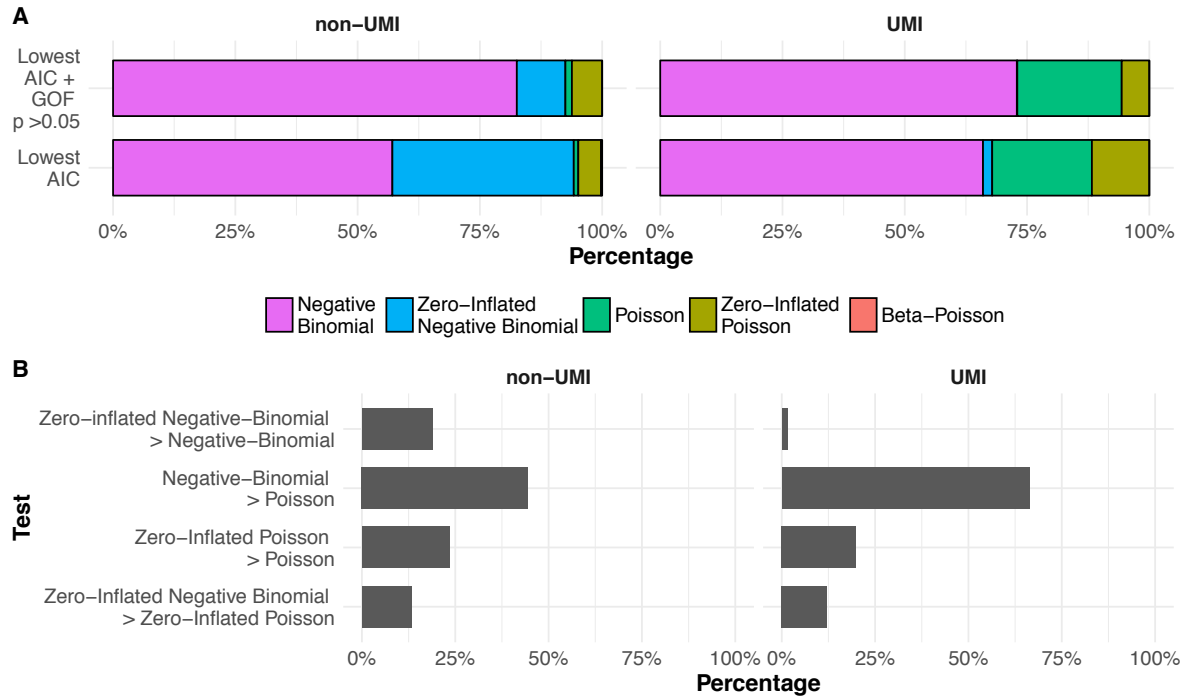


Figure S4: 6 UMI-protocols (STRT-UMI, Cel-SEq2, Drop-seq, MARS-seq, SCRB-seq,10XGenomics) are compared to 3 protocols not using UMIs (Smartseq/C1, SmartSeq2, STRT), showing that zero-inflation is only relevant for non-UMI-methods. A) The fraction of genes for which the respective distribution has the 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.

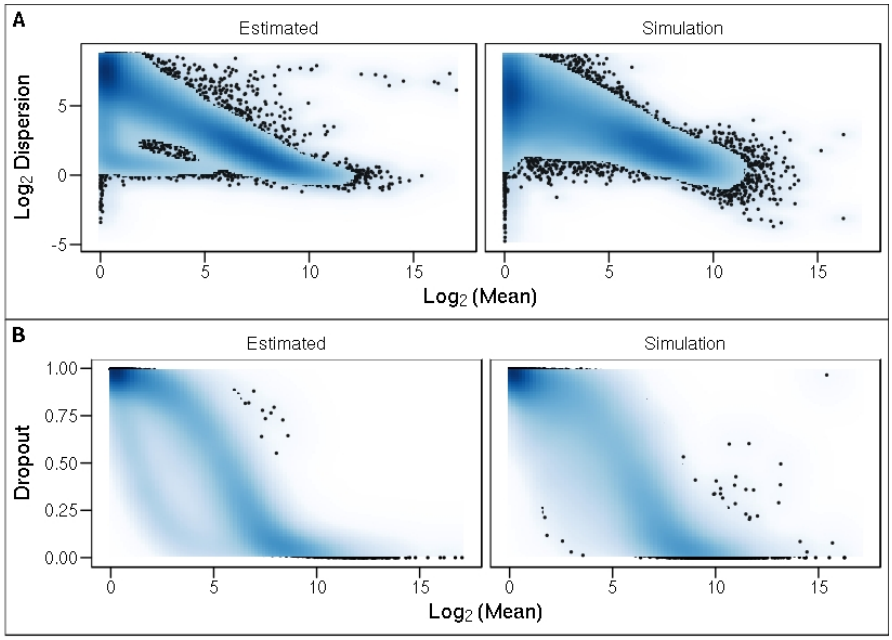


Figure S5: A) Dispersion versus mean. B) Dropout versus mean.

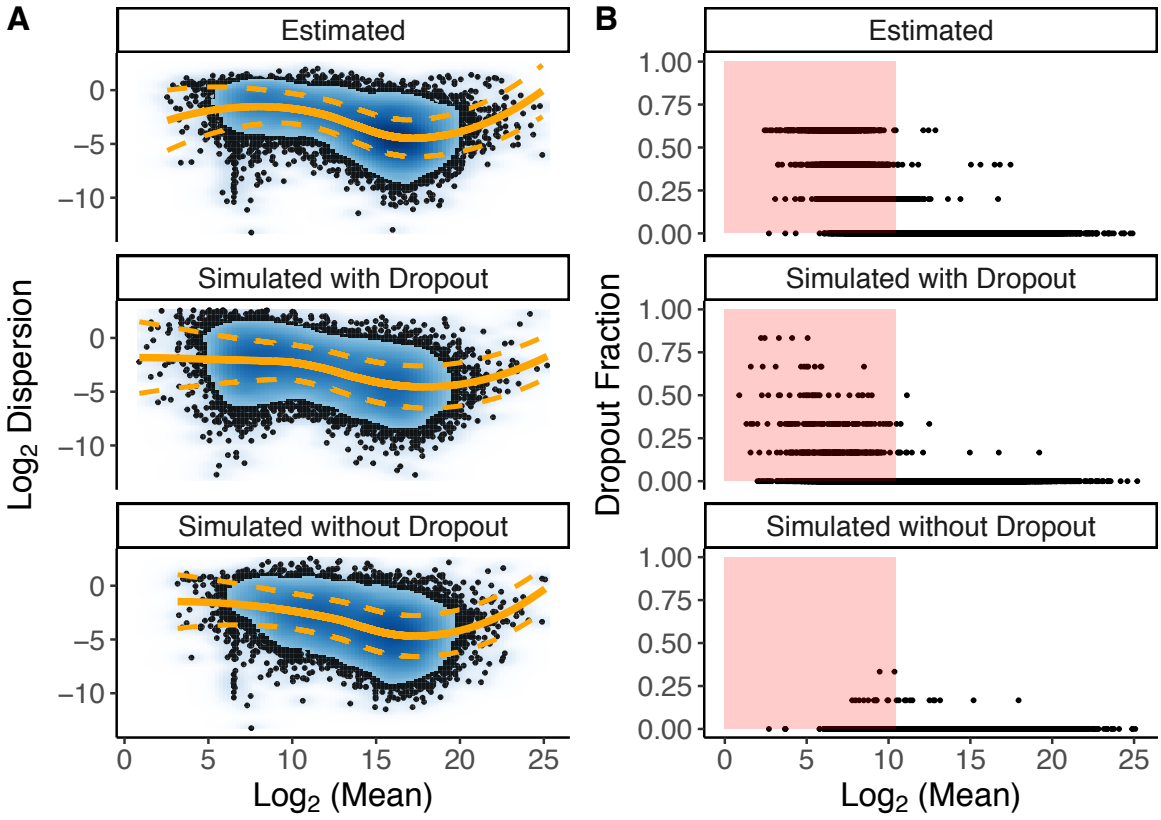


Figure S6: For bulk RNA-seq, the simulations include dropout sampling to better mimic the observed mean-dropout relation. A) Dispersion versus mean with locally weighted polynomial regression fit (orange line) and variability prediction band (dashed orange line). B) Dropout versus mean with red box indicating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

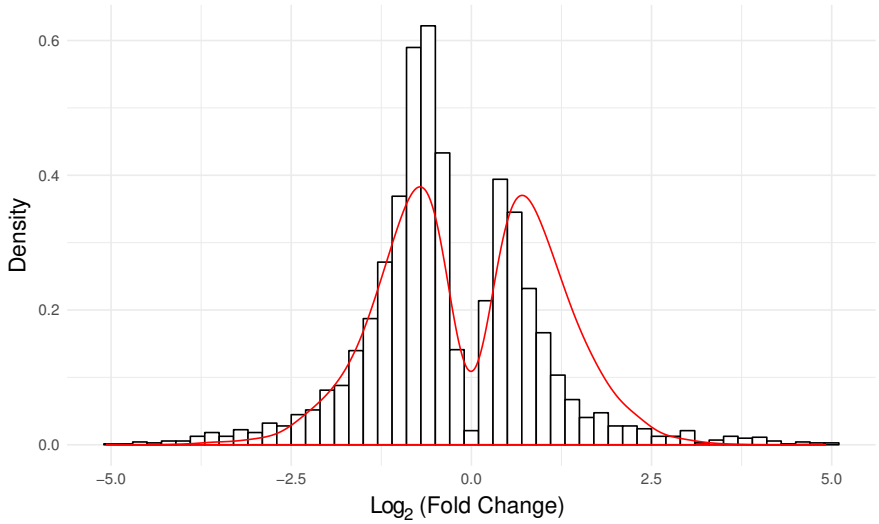


Figure S7: Log2 fold changes between serum+LiF and 2i+LiF cultured cells (Kolodziejczk et al. 2015). Red line indicates the density of a theoretical narrow gamma distribution (shape and rate equal to 3).

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.

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

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

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

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

*** contributed equally**

Abstract

RNA sequencing is increasingly performed with less starting material and at a higher sample throughput, e.g. to analyse single-cell transcriptomes. In this context, unique molecular identifiers (UMIs) are used to reduce amplification noise and sample-specific barcodes are used to track libraries. Here, we present a fast and flexible pipeline to process data from such RNA-seq protocols.
Availability: https://github.com/sdparekh/zUMIs

1 Introduction

The recent development of sensitive protocols allows to generate RNA-seq libraries of single cells [1]. The throughput of such scRNA-seq protocols is rapidly increasing, enabling the profiling of tens of thousands of cells [2, 3] and opening exciting possibilities to analyse cellular identities [4, 5]. As the required amplification from such low starting amounts introduces substantial amounts of noise [6], many scRNA-seq protocols incorporate unique molecular identifiers (UMIs) to label individual cDNA molecules with a random nucleotide sequence before amplification [7]. This allows to computationally remove amplification noise and thus increases the power to detect expression differences [8, 9]. To increase the throughput, many protocols also incorporate sample-specific barcodes (BCs) to label all cDNA molecules of a single cell with a nucleotide sequence before library generation [10, 2]. Additionally, for cell types such as neurons it has

Name	Reference	Open Source	Quality UMI/BC	Mapper	intron counting	Down-sampling
CellRanger	[2]	no	no	STAR	no	yes
Drop-seq	[10]	no	yes	STAR	no	no
CEL-seq	[13]	yes	yes	bowtie2	no	no
umis	[14]	yes	no	Kallisto	no	no
zUMIs	This work	yes	yes	STAR	yes	yes

Table 1. Pipelines handling UMI expression data

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 as part of nascent RNAs. However, the few bioinformatic tools that process RNA-seq data with UMIs and BCs have limitations with respect to availability, mapping, quality assessment and/or can not consider intronic reads (Table 1). Here, we present *zUMIs*, a fast and flexible pipeline to overcome such limitations.

2 zUMIs

zUMIs is a pipeline that processes paired fastq files containing the UMI and BC in one read and the cDNA sequence in the other read, filters out reads with bad BCs or UMIs based on sequence quality, maps reads to the genome and outputs count tables of unique UMIs or reads per gene (Figure 1). To allow the quantification of intronic reads that are generated from unspliced RNAs especially when using nuclei as input material, three separate count tables for exons, introns and exon+introns are provided. Another unique feature of *zUMI* is that it allows for downsampling of reads before summarizing UMIs per feature, which is recommended for cases of highly different read numbers per sample [15]. *zUMIs* is flexible with respect to the length and sequences of the BC and UMIs, making it compatible with a large number of protocols [16, 17, 10, 13, 3, 2].

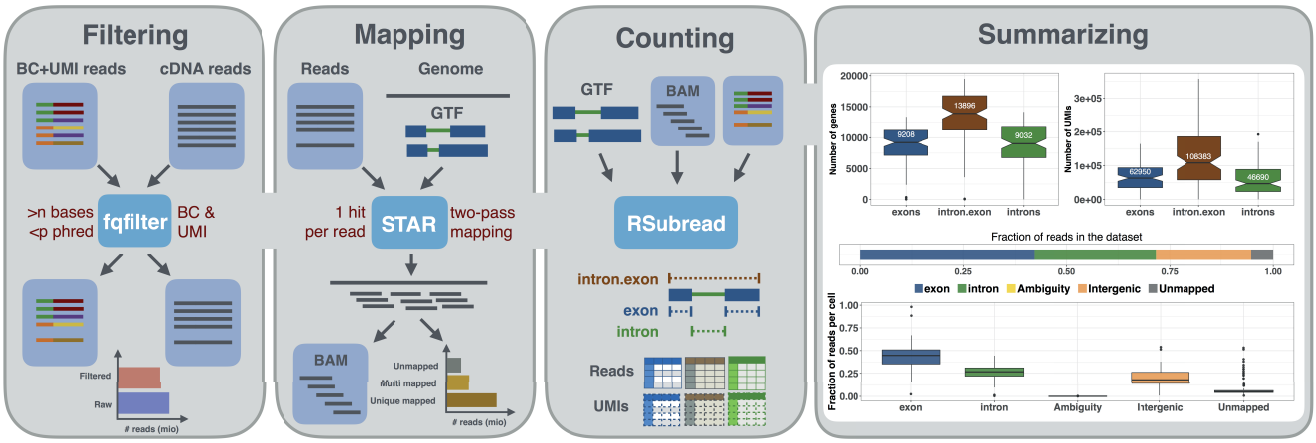


Figure 1. zUMIs schematic overview.

A Each of the grey panels from left to right depicts a step of the *zUMIs* pipeline. First, paired fastq files are filtered according to user-defined BC and UMI quality thresholds. Next, the remaining cDNA reads are mapped to the reference genome using STAR. Then 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. Optionally, *zUMIs* also generates data and plots for several quality measures, such as the number of detected Genes/UMIs per barcode and distribution of reads into mapping feature categories (Supplementary Figure 3).

2.1 Processing pipeline

The input for *zUMIs* is a pair of fastq files, whereas one file contains the cDNA sequences and the other one the read containing the BC and UMI. The exact location and length of UMI and BC are specified by the user. Note that both fastq files need to be ordered by read name, which is usually the case if unprocessed files are used. The first step in our pipeline is to filter reads where the BC or the UMI fails a user-defined quality threshold. This helps to eliminate spurious BCs and is expected to reduce noise. The cleaned-up reads are then mapped to the genome using the splice-aware aligner STAR [18]. The user is free to adapt the STAR options to their data, however *zUMIs* requires that only one mapping position per read is reported. Next, reads are assigned to genes and to exons or introns based on the provided gtf file, whereas introns are defined as not overlapping with any exon. Rsubread featureCounts [19] is used to first assign reads to exons and afterwards to check whether the remaining reads fall into introns. The resulting output is then read into R using data.table [20] and count tables for UMIs and reads are generated. *zUMIs* tabulates the UMIs/gene either for user-specified BCs or for the n BCs with the highest read counts.

2.2 Output and statistics

zUMis outputs three UMI and three read count tables: one for traditional exon mapping gene-wise counts, one for intron and one for intron+exon counts. If a user chooses the downsampling option, 6 additional count-tables are provided in which samples with an excess of reads are downsampled and samples with too few reads are dismissed (Supplementary Figures 4). We highly recommend to use this option, because normalizing across samples with vastly different library sizes does not work well [15, 21]. *zUMIs* also reports descriptive statistics. To evaluate library quality *zUMIs* summarizes the fractions of unmapped, ambiguously mapped, exon and intron mapped reads and to evaluate library complexity, the numbers of detected genes and UMIs per sample are provided (Supplementary Figures 2,3).

We processed 227 million reads with *zUMIs* and quantified expression levels for exonic and intronic counts on a unix machine using up to 16 threads, which took barely 3 hours. Increasing the number of reads increases the processing time approximately linearly, whereas filtering, mapping and counting each take up roughly one third of the total time (Supplementary Figure 1).

3 Conclusions

zUMIs is a fast and flexible pipeline to process raw reads to count tables for RNA-seq data using UMIs. To our knowledge it is the only open source pipeline that has a barcode and UMI quality filter, allows intron counting and has an integrated downsampling function (Table 1). These features ensure that *zUMIs* is applicable for most experimental designs of RNA-seq data, such as single-nuclei sequencing techniques [11, 12, 22], droplet based methods where the BC is unknown and the library sizes can vary a lot as well as plate-based UMI-methods with known BCs.

Funding

This work has been supported by the DFG through SFB1243 subprojects A14/A15.

Availability

The pipeline is freely available at https://github.com/sdparekh/zUMIs.

References

1. Rickard Sandberg. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods*, 11(1):22–24, January 2014.

2. 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.

3. 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.

4. Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, 34(11):1145–1160, 8 November 2016.

5. Aviv Regev, Sarah 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 Gottgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung K Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Joakim Lundeberg, Partha Majumder, John Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Philipakis, Chris P Ponting, Stephen R Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua R Sanes, Rahul Satija, Ton Shumacher, Alex K Shalek, Ehud Shapiro, Padmanee Sharma, Jay Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Alexander van Oudenaarden, Allon Wagner, Fiona M Watt, Jonathan S Weissman, Barbara Wold, Ramnik J Xavier, Nir Yosef, and Human Cell Atlas. The human cell atlas. 8 May 2017.

6. 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, 9 May 2016.

7. 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, January 2012.

8. 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.

9. Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: Power analysis for bulk and single cell RNA-seq experiments. 15 March 2017.

10. 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.

11. 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, 24 June 2016.

12. 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.

13. 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.

14. 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.

15. Dominic Grün and Alexander van Oudenaarden. Design and analysis of Single-Cell sequencing experiments. *Cell*, 163(4):799–810, 5 November 2015.

16. 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.

17. 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.

18. 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, 1 January 2013.

19. Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 1 April 2014.

20. Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*, 2017. URL `https://CRAN.R-project.org/package=data.table`. R package version 1.10.4.

21. Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.*, 27 February 2017.

22. Suguna Rani Krishnaswami, Rashel V Grindberg, Mark Novotny, Pratap Venepally, Benjamin Lacar, Kunal Bhutani, Sara B Linker, Son Pham, Jennifer A Erwin, Jeremy A Miller, Rebecca Hodge, James K McCarthy, Martin Kelder, Jamison McCorrison, Brian D Aevermann, Francisco Diez Fuertes, Richard H Scheuermann, Jun Lee, Ed S Lein, Nicholas Schork, Michael J McConnell, Fred H Gage, and Roger S Lasken. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.*, 11(3):499–524, March 2016.

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 Characterization of zUMIs

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

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

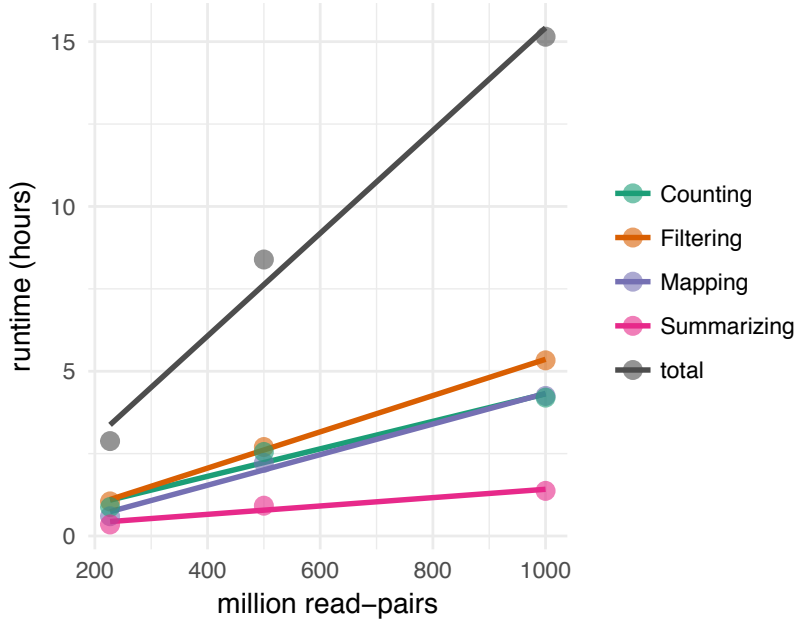


Figure S1: *zUMIs* runtime for three datasets with 227, 500 and 1000 million read-pairs. The runtime is divided in the main steps of the *zUMIs* pipeline: Filtering, Mapping, Counting and Summarizing. Each dataset was processed using up to 16 threads ("-p 16").

2 zUMIs example dataset

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

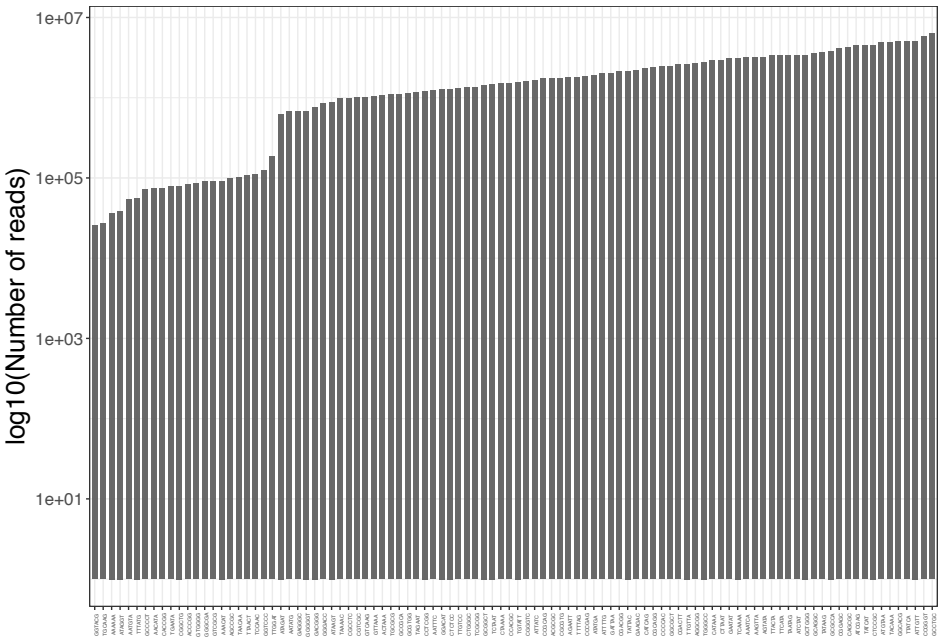


Figure S2: Reads per barcode. Bars show the number of reads assigned to each sample barcode.

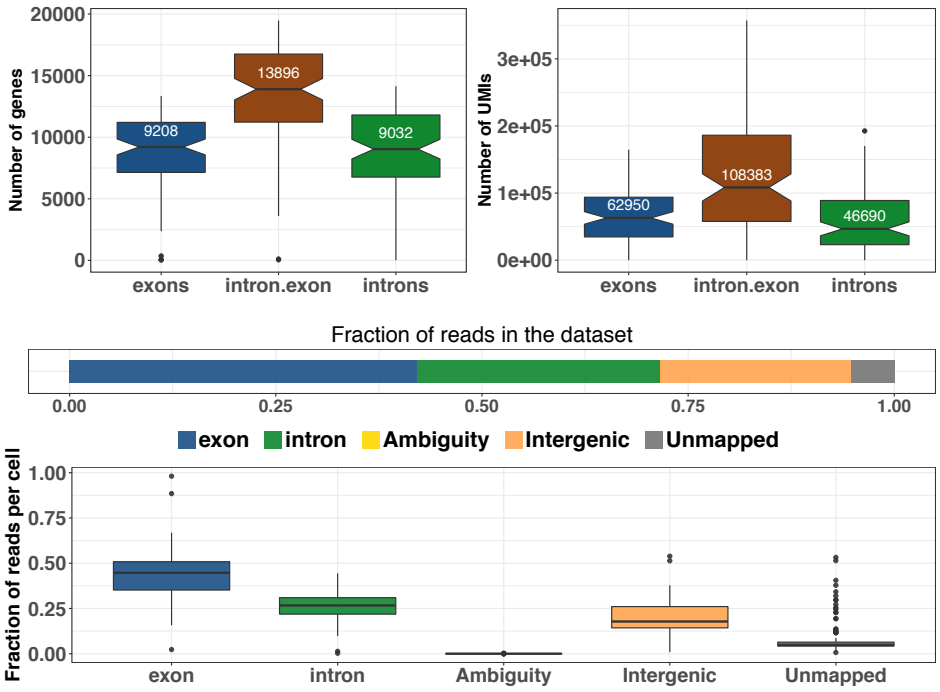


Figure S3: Summary statistics. The boxplot in the left panel shows number of genes (left) and number of UMIs(right) detected per barcode while considering only intronic/exonic counts and intornic+exonic counts. The horizontal relative barplot in the middle indicates total fraction of reads assignment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

3 Downsampling

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

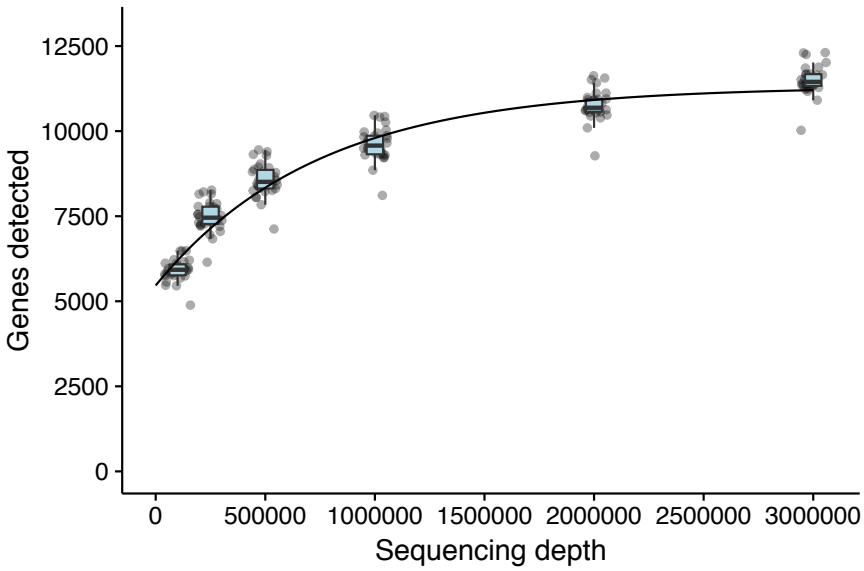


Figure S4: Downsampling. Cells were downsampled to six depths from 100,000 to 3,000,000 reads. For each sequencing depth the reads detected per cell is shown. Here the increase in the number of genes detected using 1 million as compared to 3 million reads is small, suggesting that 1 million reads per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

References

[1] Dominic Grün and Alexander van Oudenaarden. Design and analysis of Single-Cell sequencing experiments. *Cell*, 163(4):799–810, 5 November 2015.

[2] Magali Soumillon, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S Mikkelsen. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*, 5 March 2014.

[3] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, 65(4):631–643.e4, 16 February 2017.

mcSCRB-seq: sensitive and powerful single-cell RNA sequencing

mcSCRB-seq: sensitive and powerful single-cell RNA sequencing

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

* contributed equally

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

[+] **Corresponding author, Lead contact:**
Wolfgang Enard
Anthropology and Human Genomics
Department of Biology II
Ludwig-Maximilians University
Großhaderner Str. 2
82152 Martinsried, Germany
Phone: +49 (0)89 / 2180 - 74 339
Fax: +49 (0)89 / 2180 - 74 331
E-Mail: enard@bio.lmu.de

Summary

Single-cell RNA sequencing (scRNA-seq) has emerged as the central genome-wide method to characterize cellular identities and processes. While performance of methods is improving, an optimum in terms of sensitivity, efficiency and flexibility has not been reached yet. Among the flexible plate-based counting methods, “Single-Cell RNA-Barcoding and Sequencing” (SCRB-seq) is one of the most sensitive and efficient ones. Based on this protocol, we systematically evaluated reverse transcriptases, buffer modifications and PCR polymerases and found that the addition of polyethylene glycol increased the sensitivity considerably. Based on this and other improvements, we developed molecular crowding SCRB-seq (mcSCRB-seq), a fast, cost-efficient and sensitive protocol. By analyzing mouse embryonic stem cells and ERCC spike-ins we show that mcSCRB-seq is the most sensitive scRNA-seq method to date.

Keywords

single-cell RNA sequencing, sensitivity, SCRB-seq, power analysis, molecular crowding

Introduction

Whole transcriptome single-cell RNA sequencing (scRNA-seq) is a transformative tool with wide applicability to biological and biomedical questions (Wagner, Regev, and Yosef 2016). In the last few years, many new scRNA-seq protocols have been developed to overcome the challenge of isolating, reverse transcribing and amplifying the small amounts of mRNA in single cells to generate high-throughput sequencing libraries (Macaulay and Voet 2014; Kolodziejczyk, Kim, Svensson, et al. 2015). An idealized protocol would be able to generate one cDNA library molecule for each mRNA molecule in the cell. Such a protocol would be 100 % sensitive as all mRNAs would be turned into sequenceable cDNA fragments, 100 % accurate as the concentration of mRNAs would fully correlate with the number of sequenced cDNA fragments and 100% precise as the measurement error would only depend on the sampling error of sequencing reads. The lower the cost per cell for generating and sequencing a library the more efficient the protocol would be. Furthermore, cells would need to come from independent replicates and hence how flexible different numbers of cells from different biological samples could be combined would also be a relevant property of the protocol. Obviously such an optimal, one-size-fits all protocol does not exist and probably will never exist as real protocols are likely to have inherent trade-offs and hence, optimal

protocols will differ for different research questions. While many improvements have been made to scRNA-seq protocols, it is likely that further improvements are still possible. Given the importance of scRNA-seq method (Regev et al. 2017), it is also likely that further improvements of sensitivity, efficiency and/or flexibility are worth the effort.

The sensitivity of scRNA-seq methods is limited by the effectiveness of the reverse transcription and the subsequent second strand synthesis. Protocols have improved this step by optimizing enzymes, buffers and reaction volumes, resulting in conversion rates of mRNA into cDNA of 10-20% for sensitive protocols (Grün, Kester, and van Oudenaarden 2014; Svensson et al. 2017; Hashimshony et al. 2016). Amplification of the resulting minute amounts of cDNA leads to bias and noise when quantifying gene expression levels and hence reduce the accuracy and the precision of a scRNA-seq protocol. By incorporating random nucleotides - so called unique-molecular identifiers (UMIs) (Kivioja et al. 2012) - into the primers used for generating cDNA, amplification bias and noise can be removed by only counting cDNA fragments of a gene that have different UMIs. This increase in precision leads to a substantial increase in the power to detect differentially expressed genes in scRNA-seq protocols (Parekh et al. 2016; Ziegenhain et al. 2017). In most protocols the UMI is incorporated either in the oligo-dT primer or in the primer used for the second-strand synthesis (see (Sheng et al. 2017) for an - albeit expensive - exception of internal priming with UMIs). Hence, the use of UMIs results in a 5' or 3' tag counting method and sacrifices full transcript coverage. While this can be a severe drawback when splicing and/or sequence information across the entire transcript is required, it is usually sufficient when quantifying gene expression levels to identify cell types or regulatory processes. An additional and decisive advantage when reading information from the incorporated primers is that cell-specific barcodes can be incorporated during cDNA generation. This "early-barcoding" reduces costs tremendously and has allowed to develop scRNA-seq approaches that efficiently can generate libraries of tens or even hundreds of thousands of cells, especially when combined with microdroplet isolations (Macosko et al. 2015; Klein et al. 2015; Zheng et al. 2017). Hence, by incorporating early barcoding and UMIs, counting methods have made scRNA-seq protocols more precise and more efficient. Notably, higher amplification noise and bias still decreases the efficiency of the protocol, as more sequencing is necessary to obtain the same information.

Comparing protocols would ideally involve to compare the "bang for the buck", i.e. to compare the costs of protocols at a given power to detect differentially expressed genes and/or cell types for a given amount of money spent (Ziegenhain et al. 2017; Vieth et al.

2017). This is challenging, especially because there are no standardized cells with known concentrations of mRNAs available. The next best proxy are 92 standardized mRNAs known as ERCC spike-ins (Baker et al. 2005) that are used in many protocols. Recently, this ERCC data has been used to compare 19 different protocols, showing that median sensitivity, measured as the 50% detection probability, differs from 2.2 to >302 molecules (Svensson et al. 2017). Accuracy, assessed as the correlation coefficient of the known ERCC concentrations and measured expression signal in a cell, differs less, but is also difficult to interpret as it reflects a combination of sensitivity, accuracy and precision and how this translates into the power to detect differentially expressed genes is not clear. A further limitation of using ERRCs to compare scRNA-seq protocols is that it has been questioned whether the 92 ERCC transcripts really are representative mRNAs as they are shorter, have smaller poly-A tails, do not represent the relevant concentration range with enough transcripts and are purified (Tung et al. 2017; Risso et al. 2014). Indeed, it seems that some protocols are more sensitive for ERCCs than for real mRNAs and vice versa (Ziegenhain et al. 2017). An alternative approach to compare methods is to use the same cells in the same lab with different methods and compare their power to detect differentially expressed genes using simulations (Ziegenhain et al. 2017; Vieth et al. 2017). However, this is approach is expensive and difficult to realize for more than a handful of methods, especially if fresh cells are used. From these two comparisons (Ziegenhain et al. 2017; Svensson et al. 2017) , as well as from an earlier study (Wu et al. 2014) it has emerged that protocols differ considerably in their sensitivity and that low reaction volumes, as available in Fluidigm's microfluidic chips, increase sensitivity. However, the efficiency of Fluidigm chips is low due to their high costs and methods that are even more sensitive like Smart-seq2 and SCRB-seq are possible also in microliter volumes of plate-based methods. Microdroplet methods like Drop-seq, inDrop or 10x do currently not reach the sensitivity of these plate-based methods, but are very efficient due to their low costs to generate libraries, especially when large numbers of cells per sample are analyzed. Indeed, when comparing the costs at 80% power to detect differentially expressed genes, the most efficient method was found to be SCRB-seq (Ziegenhain et al. 2017), a plate-based method using barcoded, UMI-containing -oligodT primers, template switching and PCR amplification to generate scRNA-seq libraries (Soumillon et al. 2014). Here, we set out to systematically further improve the sensitivity and efficiency of SCRB-seq. Based on these evaluations, we developed molecular crowding SCRB-seq (mcSCRB-seq), a highly flexible, efficient, plate-based protocol with low set-up costs that is the most sensitive scRNA-seq protocol to date.

Design

As described above, there is the possibility and the need to improve scRNA-seq methods in terms of sensitivity and efficiency. Among plate-based methods that are efficient when processing many samples and isolating cells via FACS, SCRB-seq has been shown to be very efficient (Ziegenhain et al. 2017). However, sensitivity and amplification bias are still worse for SCRB-seq than for Smart-seq2, a methodologically similar protocol that allows to generate full-length scRNA-seq libraries, but is less precise and more costly due to the lack of UMIs and early barcoding. As the Smart-seq2 protocol has been developed by optimizing conditions for cDNA generation (Picelli et al. 2013), this suggested that sensitivity and efficiency could also be increased for SCRB-seq. Hence, we systematically and robustly assessed how different reverse transcriptases, buffer and primer modifications impact cDNA yield from low amounts of the standardized universal human reference RNA (UHRR) (SEQC/MAQC-III Consortium 2014). We then combined the most promising improvements, in particular the addition of polyethylene glycol, and could show by sequencing the generated UHRR libraries that the new molecular crowding SCRB-seq protocol represents a x-y fold increase in the number of transcripts detected compared to prior versions of SCRB-seq (Soumillon et al. 2014; Ziegenhain et al. 2017). To further improve the efficiency of the new protocol by reducing the PCR amplification bias, we tested two PCR enzymes that had generated sufficient cDNA yield (KAPA HiFi and Terra) and found Terra to approximately double the library complexity at read depths below complete saturation. We then compared this optimized protocol mcSCRB-seq directly to a previous SCRB-seq version (Ziegenhain et al. 2017) using mouse ES cells and ERCC spike-ins and find that it is twice as powerful to detect differentially expressed genes than the previous SCRB-seq protocol and with a 50% detection probability for 2.2 transcripts the most sensitive protocol among all ERCC benchmarked protocols today. Together with a low cost per sample and minimal hands-on time, our optimizations led to the highly flexible, sensitive and efficient mcSCRB-seq protocol.

Results

A streamlined assay for cDNA yield

In order to easily quantify the effects of changes to our protocol on reverse transcription (RT), second strand synthesis and PCR amplification, we first developed a streamlined

assay to use cDNA yield as a proxy for sensitivity (Figure 1). To quantify changes to the protocol independent of biological noise, we used 1 ng or less of universal human reference RNA (UHRR) as template (SEQC/MAQC-III Consortium 2014), as single-cells also from homogenous cell populations show biological variation (Kolodziejczyk, Kim, Tsang, et al. 2015; Grün, Kester, and van Oudenaarden 2014). To accommodate additions to the cDNA generation reaction more easily, we increased its volume from 2 µl (Soumillon et al. 2014) to 10 µl and confirmed that this change did not influence cDNA yield (data not shown). To quantify the cDNA yield of a single reaction, we omitted the pooling, clean-up and Exonuclease I digestion step. Instead, we heat-inactivated the reverse transcriptase enzyme and directly proceeded with PCR amplification. We measured the resulting cDNA yield by fluorometry and the cDNA length-distribution for a subset of samples by a Bioanalyzer system.

cDNA yield is highest with *Maxima H-*

First, we optimized the reverse transcription reaction. In the SCRB-seq protocol, RNA is desiccated prior to reverse transcription (Soumillon et al. 2014). Our change to 10 µl reverse transcription volume allowed us to omit this step. Furthermore, we include barcoded oligo-dT primers in the lysis buffer, saving a time-consuming pipetting step in the critical phase of any scRNA-seq protocol before reverse transcription of RNA into more stable cDNA. Together, these changes resulted in a small (~10%) increase in yield (Figure 2A).

Similar to many scRNA-seq protocols (Ramsköld et al. 2012; Picelli et al. 2013; Islam et al. 2014; Macosko et al. 2015), our method relies on oligo-dT priming to initiate reverse transcription and a template switching reaction at the 5' end to incorporate a priming site for preamplification. As enzyme sensitivity and processivity may be highly variable, we compared the performance of nine moloney murine leukemia virus (MMLV) reverse transcriptase enzymes with described template-switching properties at . When analyzing the reaction yield in response to input amounts of RNA, Maxima H- (Thermo Fisher) and SmartScribe (Clontech) performed best (Figure 2B). Furthermore, non-MMLV reverse transcriptase enzymes (SunScript, SuperScript IV and PrimeScript II) did not yield satisfactory cDNA quality (data not shown). Notably, SuperScript II (Thermo Fisher) performed significantly worse in our experiments, contrary to other protocols (Picelli et al. 2013; Hashimshony et al. 2016).

Since pooling of cells can only occur after incorporation of cell-specific barcodes by reverse transcription, the costs for this step are a major factor in overall costs. In order to reduce enzyme costs, we showed that lowering RT enzyme to 20 units per reaction (20% reduction)

does not measurably affect cDNA yield (Supplementary Figure 1A). Similarly, oligo-dT primer amounts can be reduced by 80% without repercussions (Supplementary Figure 1B). Lastly, we showed that an unblocked template-switching oligo is cheaper while retaining the same performance without primer artefacts (Supplementary Figure 1C,D).

Molecular crowding significantly increases cDNA yield

To explore additional optimizations of the RT reaction, we evaluated additives that had led to the increased sensitivity of the Smart-seq2 protocol in a previous study (Picelli et al. 2013). Both SCRB-seq and Smart-seq2 use oligodT priming and template switching to generate cDNA, but surprisingly the additives that have improved cDNA yield for Smart-seq2 do not improve SCRB-seq: In our experiments, the addition of $MgCl_2$ prevented the generation of full-length transcripts, while additives Betaine and Trehalose did not increase yield (Supplementary Figure 2A). What had not been explored so far for scRNA-seq protocols is adding agents such as polyethylene glycol that mimic macromolecular crowding and can drastically increase reaction rates (see (Rivas and Minton 2016) for a recent review). This effect is largely attributed to excluding solvent volume and thereby increasing the effective concentrations of reacting molecules. This can lead e.g. to more efficient ligation reactions (Zimmerman and Pheiffer 1983) and as a small reaction volume has been shown to increase the sensitivity of scRNA-seq protocols (Wu et al. 2014; Hashimshony et al. 2016; Svensson et al. 2017), we hypothesized that molecular crowding could increase the sensitivity of reverse transcription. Indeed, we observed that adding polyethylene glycol (PEG 8000) increased cDNA yield in a concentration-dependent manner (Supplementary Figure 2B). Because negative controls showed unspecific products at higher PEG-concentrations, we chose 7.5% PEG 8000 as an optimal concentration balancing yield and high specificity (Supplementary Figure 2C). With the addition of PEG 8000, yield increased dramatically, making it possible to detect RNA inputs under 1 pg (Figure 2C).

Increases in cDNA yield translate to increased sensitivity

In order to demonstrate that our increases in cDNA yield correspond to increases in sensitivity, we constructed libraries from eight replicates of 10 pg total RNA input with four protocol variants (Supplementary Table 1). Variant 1 (“Soumillon”) corresponds to the original SCRB-seq protocol (Soumillon et al. 2014), variant 2 (“Ziegenhain”) corresponds to the SCRB-seq protocol substituted with KAPA HiFi (Ziegenhain et al. 2017), variant 3 (“SmartScribe”) uses SmartScribe and KAPA HiFi, while variant 4 (“molecular crowding”) combined Maxima H-, 7.5% PEG 8000 and KAPA HiFi.

Here, the molecular crowding protocol yielded the most cDNA, while variant 1 yielded the least, confirming our systematic optimization (Figure 3A). Interestingly, variant 2 clearly outperformed variant 3, substantiating that Maxima H- is the most sensitive reverse transcriptase enzyme evaluated here. Next, we pooled all 32 libraries and sequenced 81 million reads. We used *zUMIs* (Parekh et al. 2017) to process and downsample sequencing data to one million reads per sample (Supplementary Figure 3), which has been suggested to correspond to reasonable saturation for single-cell RNA-seq experiments (Svensson et al. 2017; Ziegenhain et al. 2017). Libraries that did not obtain 1 million reads were excluded from the analysis. Taking the number of detected (>= 1 UMI) genes per sample (Figure 3B) as a first proxy for sensitivity confirmed that the molecular crowding method is the most sensitive protocol ($p = 7 \times 10^{-7}$, Welch Two Sample t-test, compared to variant 2) with 7,898 genes on average, while variants 1-3 detected only 3,938, 5,542, 3,805 genes, respectively.

As our data contained UMIs, we could then use the number of total detected molecules per sample as a second measure of sensitivity (Figure 3C). Although more variable, this corroborated our findings on detected genes. Next, we asked whether the increase in sensitivity translates not only in more detected genes but also in more reproducible detection of genes. For this, we calculated the dropout probabilities of genes, excluding stochastically detected genes (<0.2 UMIs mean expression) (Lun, Bach, and Marioni 2016a). Confirming our previous findings, molecular crowding markedly improved detection rates. Clearly visible, genes had lower overall dropout probabilities and a significantly larger number of genes was detected in all samples (Figure 3D).

Terra polymerase retains library complexity during PCR

Single-cell RNA sequencing methods rely on amplification of very low amounts of input material. It is well established that noise and bias may be introduced during library PCR, depending on the number of cycles, reaction conditions and polymerases (Parekh et al. 2016; Quail et al. 2012). While UMIs can largely correct the effects of noise and bias, it still requires more reads to reach the same information, resulting in a higher efficiency of scRNA-seq methods that have less amplification noise and bias (Ziegenhain et al. 2017; Sasagawa et al. 2017). To optimize PCR conditions we first evaluated the effect of various high fidelity polymerases in the amplification step on the cDNA yield. In total, twelve enzymes from eight vendors were examined. Three polymerases (KAPA HiFi, SeqAmp and Terra) yielded significantly more amplified cDNA after 18 PCR cycles (Supplementary Figure 4A) than the enzyme used by our baseline protocol SCRB-seq (Advantage2)

(Soumillon et al. 2014). We discarded SeqAmp because of a decreased median length of the amplified cDNA molecules (Supplementary Figure 4B) and compared amplification and noise of the KAPA and Terra polymerases by generating libraries from single mouse embryonic stem cells (mESCs) using our optimized molecular crowding protocol to generate cDNA. We pooled cDNA from 32 cells and amplified cDNA using either KAPA or Terra polymerase. After sequencing both library pools, we processed the data and downsampled each transcriptome to the same number of raw reads to exclude bias from varying coverage (Parekh et al. 2017). Taking the number of detected UMIs per cell as a measure, we found that PCR amplification using the Terra polymerase yielded twice as much library complexity than with KAPA HiFi (Supplementary Figure 4C). Thus, we chose Terra polymerase for the mcSCRB-seq protocol in order to retain as much as possible of the initial transcriptome complexity through preamplification. Importantly, the higher yield of the molecular crowding reverse transcription allowed us to reduce the number of PCR cycles further reducing amplification bias (Parekh et al. 2016).

Sensitivity in mouse embryonic stem cells is increased 2.5-fold

In order to assess the improvements of the molecular crowding SCRB-seq protocol (Supplementary Table 2) for single-cell transcriptomics, we sequenced further single mESCs. To provide quantitative information relative to previous benchmarking of scRNA-seq protocols (Ziegenhain et al. 2017), we prepared libraries from mESCs using the *Ziegenhain et al.* and mcSCRB-seq protocols. We used a single sample of mESCs and sorted two plates containing 96 and 48 cells for each of both methods. Libraries were prepared on the same day and multiplexed for sequencing in order to avoid batch effects.

Following sequencing, we filtered cells by excluding doublets identified from the distribution of per-cell total UMI counts (Ziegenhain et al. 2017). Furthermore we discarded broken cells and failed libraries by inspecting nearest-neighbor correlation of gene expression values (Petropoulos et al. 2016), yielding 249 high-quality libraries (Supplementary Figure 5). Importantly, the mcSCRB-seq protocol showed a high rate of reads mapping to the genome, allowing to quantify gene expression from most of the sequenced reads (Supplementary Figure 6).

Next, we assessed the sensitivity and library complexity relative to sequencing coverage. We used the *zUMIs* pipeline (Parekh et al. 2017) to downsample reads of each cell to fixed depths. Interestingly, libraries were not yet sequenced to saturation at a million reads (Supplementary Figure 7A). Still, already at low sequencing coverages, the mcSCRB-seq

protocol clearly outperformed SCRB-seq, detecting on average over 2.5 times as many unique molecules per cell at sequencing depths above 200,000 reads (Figure 4A). Furthermore, at the gene level mcSCRB-seq detected ~1,500 genes more per cell (Supplementary Figure 7B). In order to judge the absolute sensitivity of mcSCRB-seq, we used ERCC spike-ins to estimate the RNA content per cell by dividing the number of detected transcriptomic UMIs by the fraction of ERCC UMIs detected from the annotated molecule number (Supplementary Figure 8). Fitting with previous reports (Islam et al. 2014), the median mRNA content of our mouse ES cells was 227,467 molecules. Using this estimate, we could then convert the number of transcriptomic UMIs detected to the fraction of the cellular mRNAs that was observed (Figure 4B). At high sequencing depths, mcSCRB-seq could detect above 50% of the cellular mRNA content, far exceeding estimated efficiency of previous protocols (Grün, Kester, and van Oudenaarden 2014).
Furthermore, the higher sensitivity of mcSCRB-seq lead to the detection of a larger geneset overall when pooling cells (Supplementary Figure 9A). Similarly, dropout rates for detected genes were higher in the original SCRB-seq protocol (Supplementary Figure 9B). This shows that mcSCRB-seq outpowers its source method not only for sensitivity but also for consistency, as transcripts are detected more reliable in the different cells. Lastly, we could also confirm that the optimization of the preamplification enzyme yielded more uniform amplification, because amplification bias measured as extra-poisson variability was lower in the mcSCRB-seq protocol (Supplementary Figure 9C,D). Although both methods use UMIs to remove PCR bias (Supplementary Figure 9D), the reduction of the preamplification variance leads to higher information content at the same sequencing depth.

mcSCRB-seq is the most sensitive protocol as determined by ERCCs

After the characterisation of mcSCRB-seq relative to the original SCRB-seq protocol, we proceed to analyze the absolute sensitivity using ERCCs (Baker et al. 2005). For this, we spiked ERCC Mix 1 at 1:80,000 dilution to mouse ES cells, equaling to a total number of 77,923 spiked mRNA molecules per cell. Next, we use a binomial logistic regression to compute sensitivity as the probability for detection of ERCC genes relative to their spiked-in molecule number, as proposed by others (Figure 5A) (Svensson et al. 2017). In our mcSCRB-seq dataset, 50% detection probability was reached on average from 2.2 molecules input (Figure 5A). Because of the large spread in copy numbers of ERCC genes, the capture of low abundance mRNA species was only possible with high sequencing depth. Thus, the absolute sensitivity estimation depends on the number of sequenced reads but

stabilizes after 1-2 million reads (Figure 5B). Because of the wealth of single-cell RNA sequencing protocols that have become available in the recent years, method comparisons are important for users to make an informed choice. Here, we place our new mcSCRB-seq protocol into the context of other protocols by integrating ERCC spike-in data from the two major independent protocol comparisons (Svensson et al. 2017; Ziegenhain et al. 2017) and additional important protocols such as the Quartz-seq2 protocol (Sasagawa et al. 2017) and the 10x Genomics Chromium chemistry (Zheng et al. 2017). For each method, we either used the published detection limits for ERCC molecules (Svensson et al. 2017) or computed them using the binomial logistic regression described above. Here, mcSCRB-seq needed the lowest number of ERCC molecules for a 50% detection probability, making it the most sensitive protocol to date, followed by CEL-seq/C1 and Smart-seq/C1.

mcSCRB-seq combines high power, fast processing and low costs

After characterizing the increased sensitivity of mcSCRB-seq, we quantify these improvements in regard to the detection of differentially expressed genes. For this, we utilize scRNA-seq simulations (Vieth et al. 2017) of two-group comparisons with varying sample sizes. As expected, our newly developed mcSCRB-seq protocol increased the true positive rate significantly (Figure 6A). Importantly, the increase in statistical power was very large (up to ~2x higher TPR) at very small sample sizes, proving that mcSCRB-seq extracts most of the information of each sequenced cell. In order to reach a power level of 80%, mcSCRB-seq needed only 192 cells per group, while SCRB-seq needed roughly 384 cells. The false discovery rate (“FDR”) was well controlled in all conditions below the nominal level of 10% (Figure 6A). Furthermore, mcSCRB-seq showed higher consistency, as batch effects between the two processed plates were greatly reduced (Supplementary Figure 10A). As with SCRB-seq, data quality is excellent and features minimal GC or length bias (Supplementary Figure 10B,C) (Phipson, Zappia, and Oshlack 2017). In our recent protocol comparison (Ziegenhain et al. 2017), SCRB-seq was already the most powerful and cost-efficient protocol for single-cell RNA sequencing studies. mcSCRB-seq not only further improves upon this high efficiency by larger statistical power, but also by significantly reduced costs. Considering all relevant cost factors including enzymes, kits and plasticware, library preparation costs are below 60 cents per cell, down from 2 Euro, when performed in 96-well plates (Figure 6B, Supplementary Table 3). Due to the pooling of barcoded transcriptomes after reverse transcription, experiments conducted in 384-well plates increase cost efficiency even further. Moreover, owing to an optimized workflow, we could

drastically reduce the working time required to complete the protocol making it possible to create sequencing-ready libraries in one working day with minimal hands-on time (Figure 6C, Supplementary Table 4). Taken together, we show that the mcSCRB-seq protocol presented here greatly increases sensitivity and power while reducing amplification bias, costs and processing time.

Discussion

Here, we have presented an optimized and improved protocol for highly efficient single-cell RNA sequencing (key characteristics listed in Supplementary Table 3). We have shown that molecular crowding using polyethylene glycol can strongly increase the efficiency and yield of reverse transcription reactions (Figure 2,3). Presumably, the molecular crowding reduces the effective accessible volume of the reaction, similar to the biophysical properties of cells where 20-30% of the volume is occupied by macromolecules (Han and Herzfeld 1993). This volume exclusion can increase the rate of reactions significantly (Ellis 2001), as demonstrated previously for DNA ligation (Zimmerman and Pheiffer 1983). Furthermore, the optimized reaction conditions in reverse transcription allowed us to decrease the amount of required enzyme leading to a decrease in costs for the most expensive step (Figure 6B). Although others (Picelli et al. 2013; Hashimshony et al. 2016) have presented systematically optimized scRNA-seq protocols as well, improvements found in these methods did not necessarily translate to the SCRB-seq method. This is especially surprising for the case of Smart-seq2, which also relies on template-switching and PCR amplification. Consequently, there seem to be complex interactions of primers, enzymes and reaction conditions that necessitate individual optimization of each protocol. It will be interesting to see whether the molecular crowding described here is a general enough mechanism to be a useful improvement for a large variety of other scRNA-seq protocols.

Judging the performance of mcSCRB-seq relative to the large number of protocols already described in the literature (Svensson et al. 2017; Ziegenhain et al. 2017), analysis of ERCC spike-in data showed that it is the most sensitive molecule-counting protocol to date (Figure 5C). Although ERCC spike-ins have been criticized for poorly modelling endogenous mRNA molecules for instance due to short poly-A tails and lack of cap-structure (Stegle,

Teichmann, and Marioni 2015; Grün and van Oudenaarden 2015), they are still a valuable tool to easily compare sensitivity amongst protocols (Svensson et al. 2017).

Furthermore, we found that a higher yield and larger transcriptome complexity after reverse transcription could reduce PCR bias as less amplification was now necessary. In addition, we show that using Terra polymerase for amplification reduces PCR noise even further. This is crucial in order to retain maximum cDNA complexity throughout amplification. Importantly, less amplification bias leads to a larger fraction of unique molecular counts detected per raw sequencing coverage, especially for low sequencing depths (Sasagawa et al. 2017). This higher slope of UMI detection makes our mcSCRB-seq protocol even more cost-efficient in practice and large amounts of information per cell can be extracted already from low sequencing depths. When sequencing to higher coverages, we show that mcSCRB-seq is able to measure a large fraction of the cellular mRNA molecules (Figure 4B) and is thus ideally suited for studying rare cells in depth. Taken together, mcSCRB-seq is a fast, inexpensive, powerful and highly flexible protocol that is equally as useful for the survey of large numbers of cells isolated by FACS sorting as for the high resolution analysis of few cells.

Limitations

In order to incorporate unique molecular identifiers and cellular barcodes, mcSCRB-seq is limited to sequencing 3’ ends of transcripts. Thus, most splicing information will be missed by our method. While other methods, especially those based on droplet microfluidics or combinatorial indexing (Macosko et al. 2015; Klein et al. 2015; Zheng et al. 2017; Cao et al. 2017; Rosenberg et al. 2017), feature higher massively parallel throughput, our plate-based mcSCRB-seq protocol can still produce medium to high amounts of data in short amount of time at very competitive low costs.

STAR Methods

Key Resources Table

Reagent/Resource	Source	Identifier
Chemicals, Peptides, recombinant proteins		
2-Mercaptoethanol (50 mM)	Thermo Fisher	21985-023
AccuPrime Pfx	Invitrogen	12344-024
AccuStart Taq HiFi	Quantabio	1706-25-BL
Advantage 2	Clontech	639207
Betaine 5M	Sigma-Aldrich	B0300-5VL
CHIR99021	Sigma-Aldrich	SML1046-25MG
Clontech Lysis Buffer	Clontech	ST0361
D(+)-Trehalose Dihydrate	Sigma-Aldrich	90210-50G
dNTPs (100 mM each)	Thermo Fisher	R0182
Dulbecco's modified Eagle medium	Thermo Fisher	41965062
EDTA	Sigma Aldrich	E7889
EnzScript	Biozym	280560
Esgro recombinant mouse LIF	Millipore	ESG1107
Ethanol, absolute	Carl Roth	9065.4
Exonuclease I (20 U/uL)	Thermo Fisher	EN0582
Exonuclease I Reaction Buffer (10x)	Thermo Fisher	EN0582
Fetal bovine serum	Thermo Fisher	10500-064
FideliTaq	Affymetrix	71156
Gelatin (from porcine skin)	Sigma Aldrich	G1890-1KG
GoScript	Promega	A5003
Guanidine Hydrochloride	Sigma-Aldrich	G3272
Igepal CA-630	Sigma-Aldrich	I8896
KAPA HiFi 2x ReadyMix	KAPA Biosystems	KR0370
L-Glutamine	Thermo Fisher	25030-024
Magnesium Chloride Solution	Sigma-Aldrich	M1028
Maxima H- Reverse Transcriptase	Thermo Fisher	EP0753
Maxima RT Buffer (5x)	Thermo Fisher	EP0753

M-MLV Reverse Transcriptase, RNase H Minus, Point Mutant	Promega	M3682
Nonessential Amino Acids (NEAA)	Thermo Fisher	11140-035
NucBlue Live	Molecular Probes	R37605
PBS	Gibco	10010-023
PD0325901	Sigma-Aldrich	PZ0162-25MG
Penicillin-Streptomycin (Pen-Strep)	Thermo Fisher	157070-063
Phusion Flash	Thermo Fisher	F-548S
Phusion HF Buffer	New England Biolabs	B0518
PicoMaxx	Agilent	600650
Platinum SuperFi	Thermo Fisher	12358-010
Polyethylene glycol	Sigma-Aldrich	89510
Precisor	BioCat	1706-25-BL
Proteinase K	Ambion	AM2546
ProtoScript II	New England Biolabs	M0368
Q5	New England Biolabs	M0493L
RLT Plus Buffer	Qiagen	1053393
RevertAid Reverse Transcriptase	Thermo Fisher	EP0441
RevertUP II Reverse Transcriptase	Biozym	350400501
RNAprotect Cell Reagent	Qiagen	76526
SeqAmp	Clontech	638504
SmartScribe	Clontech	639537
Sodium Azide Reagent Plus 99.5%	Sigma-Aldrich	S2002-100G
Sodium Chloride (NaCl)	Sigma-Aldrich	S5150-1L T2694
SuperScript II	Thermo Fisher	18064-014
Terra PCR Direct Polymerase Mix	Clontech	639271
Triton X-100	Sigma Aldrich	T8787
Trizma hydrochloride solution	Sigma-Aldrich	T2694
Trypsin / EDTA	Thermo Fisher	25200-056
UltraPure Distilled Water	Invitrogen	10977-049
Critical Commercial Assays		
Clean & Concentrator-5 Kit	Zymo Research	D4013
High Sensitivity DNA Analysis Kits	Agilent	5067-4626

MinElute Gel Extraction Kit	Qiagen	28606
Nextera XT DNA Sample Preparation Kit	Illumina	FC-131-1096
Quant-iT PicoGreen dsDNA Assay Kit	Thermo Fisher	P7589
Deposited data		
Chromium ERCC data	Zheng et al., 2017	http://support.10xgenomics.com/single-cell/datasets
Comparative scRNA-seq ERCC data	Ziegenhain et al., 2017	GEO: GSE75790
ERCC binomial regression data (15 protocols	Svensson et al., 2017	nmeth.4220-S3
Quartz-seq2 ERCC data	Sasagawa et al., 2017	GEO: GSE99866
Single-cell RNA-seq data	This paper	GEO: GSE103568
UHRR RNA-seq data	This paper	GEO: GSE103568
Experimental Models: Cell Lines		
J1 mESCs	Li et al., 1992	129S4/SvJae
JM8 mESCs	Pettitt et al., 2009	C57BL/6N
Sequence-Based Reagents		
Universal Human Reference RNA (UHRR)	Agilent	740000
ERCC RNA Spike-In Mix	Ambion	4456740
Nextera XT i7 Index primer	IDT	"TruGrade Ultramer"
SCRB-seq P5 primer, AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG*A*T*C*T, * PTO bond	IDT	NA
SCRB-seq oligo-dT primer, Biotin-ACACTCTTTCCCTACACGACGCTCTTCCGATCT[BC6][N10][T30]VN	IDT	"TruGrade Ultramer"
SCRB-seq template-switch oligo, iCiGiCACACTCTTTCCCTACACGACGCrGrGrG	Eurogentec	NA
mcSCRB-seq template-switch oligo	IDT	HPLC

unblocked, ACACTCTTTCCCTACACGACGCrGrGrG		
SCRB-seq SINGV6 PCR primer, Biotin-ACACTCTTTCCCTACACGACGC	IDT	NA
Software and Algorithms		
R (v 3.4.0)	R Development Core Team, 2008	https://cran.r-project.org
RStudio (v 1.1.364)	RStudio Team, 2015	https://www.rstudio.com
STAR (v 2.5.3a)	Dobin et al., 2013	https://github.com/alexdobin/STAR
zUMIs	Parekh et al., 2017	https://github.com/sdparekh/zUMIs/
powsimR (v 0.0.905)	Vieth et al., 2017	https://github.com/bvieth/powsimR
MASS (v 7.3-47)	Venables & Ripley 2002	https://cran.r-project.org/web/packages/MASS/
Other		
2% E-Gel Agarose EX Gels	Life Technologies	G402002
Sera-Mag Speed Beads	Thermo Fisher	65152105050250

Optimization experiments

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

Comparison of reverse transcriptases

Nine reverse transcriptases, Maxima H- (Thermo Fisher), SMARTScribe (Clontech), Revert Aid (Thermo Fisher), EnzScript (Biozym), ProtoScript II (New England Biolabs), Superscript II (Thermo Fisher), GoScript (Promega), Revert UP II (Biozym), M-MLV Point Mutant (Promega), were compared to determine which enzyme resulted in the largest cDNA yield. Several dilutions ranging from 10 to 1000 pg of universal human reference RNA (UHRR; Agilent) were used as input into the RT reactions.

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

Effect of reaction enhancers

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

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

Comparison of PCR DNA polymerases

The following twelve DNA polymerases were evaluated in preamplification: KAPA HiFi HotStart (KAPA Biosystems), SeqAmp (Clontech), Terra direct (Clontech), Platinum SuperFi (Thermo Fisher), Precisor (Biocat), Advantage2 (Clontech), AccuPrime Taq (Invitrogen), Phusion Flash (Thermo Fisher), AccuStart (QuantaBio), PicoMaxx (Agilent), FideliTaq (Affymetrix), Q5 (New England Biolabs). For each enzyme, at least three replicates of 1 ng UHRR were reverse transcribed using the optimized molecular crowding reverse transcription in 10 µl reactions. Optimal concentrations for dNTPs, reaction buffer, stabilizers, and enzyme were determined using manufacturer's recommendations. For all amplification reactions, we used the original SCRB-seq PCR cycling conditions (Soumillon et al. 2014).

Cell culture of mouse embryonic stem cells

J1 (Li, Bestor, and Jaenisch 1992) and JM8 (Pettitt et al. 2009) mouse embryonic stem cells were cultured under feeder-free conditions on gelatine-coated dishes in high-glucose Dulbecco's modified Eagle's medium (Thermo Fisher) supplemented with 15% fetal bovine serum (FBS, Thermo Fisher), 100 U/ml penicillin, 100 µ g/ml streptomycin (Thermo Fisher), 2 mM L-glutamine (Thermo Fisher), 1x MEM non-essential amino acids (NEAA, Thermo Fisher), 0.1 mM β-mercaptoethanol (Thermo Fisher), 1000 U/ml recombinant mouse LIF (Merck Millipore) and 2i (1 µM PD032591 and 3 µM CHIR99021 (Sigma-Aldrich)). mESCs were routinely passaged using 0.25% trypsin (Thermo Fisher).
mESC cultures were confirmed to be free of mycoplasma contamination by a PCR-based test (Young et al. 2010).

SCRB-seq cDNA synthesis

Cells were dissociated using trypsin and resuspended in 100 µL of RNAprotect Cell Reagent (Qiagen) per 100 000 cells. Directly prior to FACS sorting, the cell suspension was diluted with PBS (Gibco). Single cells were sorted into 96-well DNA LoBind plates (Eppendorf) containing lysis buffer using a Sony SH800 sorter (Sony Biotechnology; 100 µm chip) in "Single Cell (3 Drops)" purity. Lysis buffer consisted of a 1:500 dilution of Phusion HF buffer (New England Biolabs). After sorting, plates were spun down and frozen at -80 °C.
Libraries were prepared as described previously (Ziegenhain et al. 2017; Soumillon et al. 2014). Briefly, proteins were digested with Proteinase K (Ambion) followed by desiccation to inactivate Proteinase K and reduce the reaction volume. RNA was then reverse transcribed in a 2 µL reaction at 42°C for 90 min. Unincorporated barcode primers were digested using Exonuclease I (Thermo Fisher). cDNA was pooled using the Clean & Concentrator-5 kit (Zymo Research) and PCR amplified with the KAPA HiFi HotStart polymerase (KAPA Biosystems) in 50 µL reaction volumes.

mcSCRB-seq cDNA synthesis

Cells were dissociated using trypsin and resuspended in PBS. Single cells were sorted into 96-well DNA LoBind plates (Eppendorf) containing 5 µl lysis buffer using a Sony SH800 sorter (Sony Biotechnology; 100 µm chip). Lysis buffer consisted of a 1:500 dilution of Phusion HF buffer (New England Biolabs), 1.25 µg/µl Proteinase K (Clontech) and 0.4 µM barcoded oligo-dT primer (E3V6NEXT, IDT). After sorting, plates were immediately spun

down and frozen at -80 °C. For libraries containing ERCCs, 0.1 µl of 1:80,000 dilution of ERCC spike-in Mix 1 was used.

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

Library Preparation

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

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

Sequencing

Libraries were paired-end sequenced on high output flow cells of an Illumina HiSeq 1500 instrument. 16 bases were sequenced with the first read to obtain cellular and molecular barcodes and 50 bases were sequenced in the second read into the cDNA fragment. When several libraries were multiplexed on sequencing lanes, an additional 8 base i7 barcode read was done.

Primary Data Processing

All raw fastq data was processed using zUMIs together with STAR to efficiently generate expression profiles for barcoded UMI data (Parekh et al. 2017; Dobin et al. 2013). For UHRR experiments, we mapped to the human reference genome (hg38) while mouse cells were mapped to the mouse genome (mm10) concatenated with the ERCC reference. Gene annotations were obtained from Ensembl (GRCh38.84 or GRCm38.75). Downsampling to fixed numbers of raw sequencing reads per cell were performed using the "-d" option in zUMIs.

Filtering of scRNA-seq libraries

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

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

Estimation of cellular mRNA content

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

ERCC Analysis

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

For public data from *Svensson et al.*, we used their published molecular abundances calculated using the same logistic regression model obtained from "Supplementary Table 2"

(https://www.nature.com/nmeth/journal/v14/n4/extref/nmeth.4220-S3.csv) (Svensson et al. 2017). For Quartz-seq2 (Sasagawa et al. 2017), we obtained expression values for ERCCs from Gene Expression Omnibus (GEO; GSE99866), sample GSM2656466; for Chromium (Zheng et al. 2017) we obtained expression tables from the 10x Genomics webpage (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/ercc) and for SCRB-seq, Smart-seq2, CEL-seq2/C1, MARS-seq and Smart-seq/C1 (Ziegenhain et al. 2017), we obtained count tables from GEO (GSE75790). For these methods, we calculated molecular detection limits given their published ERCC dilution factors.

Power Simulations

For power simulation studies, we used the *powsimR* package (Vieth et al. 2017). Parameter estimation of the negative binomial distribution was done using scran normalized counts (Lun, Bach, and Marioni 2016a). Next, we simulated two-group comparisons with 10% differentially expressed genes. Log2 fold-changes were drawn from a normal distribution with mean of 0 and a standard deviation of 1.5. In each of the 25 simulation iterations, we draw equal sample sizes of 24, 48, 96, 192 and 384 cells per group and test for differential expression using ROTS (Seyednasrollah et al. 2015) and scran normalization (Lun, Bach, and Marioni 2016b).

Batch Effect Analysis

In order to detect genes differing between batches of one scRNA-seq protocol, data were normalized using scran (Lun, Bach, and Marioni 2016a). Next, we tested for differentially expressed genes using limma-voom (Ritchie et al. 2015; Law et al. 2014). Genes were labelled as significantly differentially expressed between batches with Benjamini-Hochberg adjusted p-values < 0.01.

References

Baker, Shawn C., Steven R. Bauer, Richard P. Beyer, James D. Brenton, Bud Bromley, John Burrill, Helen Causton, et al. 2005. “The External RNA Controls Consortium: A Progress Report.” *Nature Methods* 2 (10): 731–34.

Cao, Junyue, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, et al. 2017. “Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism.” *Science* 357 (6352): 661–67.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Ellis, R. J. 2001. “Macromolecular Crowding: Obvious but Underappreciated.” *Trends in Biochemical Sciences* 26 (10): 597–604.

Gierahn, Todd M., Marc H. Wadsworth 2nd, Travis K. Hughes, Bryan D. Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J. Christopher Love, and Alex K. Shalek. 2017. “Seq-Well: Portable, Low-Cost RNA Sequencing of Single Cells at High Throughput.” *Nature Methods*, February. doi:10.1038/nmeth.4179.

Grün, Dominic, Lennart Kester, and Alexander van Oudenaarden. 2014. “Validation of Noise Models for Single-Cell Transcriptomics.” *Nature Methods* 11 (6): 637–40.

Grün, Dominic, and Alexander van Oudenaarden. 2015. “Design and Analysis of Single-Cell Sequencing Experiments.” *Cell* 163 (4): 799–810.

Han, J., and J. Herzfeld. 1993. “Macromolecular Diffusion in Crowded Solutions.” *Biophysical Journal* 65 (3): 1155–61.

Hashimshony, Tamar, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, et al. 2016. “CEL-Seq2: Sensitive Highly-Multiplexed Single-Cell RNA-Seq.” *Genome Biology* 17 (1): 77.

Hashimshony, Tamar, Florian Wagner, Noa Sher, and Itai Yanai. 2012. “CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification.” *Cell Reports* 2 (3): 666–73.

Hochgerner, H., P. Lönnerberg, R. Hodge, and J. Mikes. 2017. “STRT-Seq-2i: Dual-Index 5′ Single Cell and Nucleus RNA-Seq on an Addressable Microwell Array.” *bioRxiv*. biorxiv.org. http://biorxiv.org/content/early/2017/04/20/126268.abstract.

Islam, Saiful, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. 2011. “Characterization of the Single-Cell Transcriptional Landscape by Highly Multiplex RNA-Seq.” *Genome Research* 21 (7): 1160–67.

Islam, Saiful, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. 2014. “Quantitative Single-Cell RNA-Seq with Unique Molecular Identifiers.” *Nature Methods* 11 (2): 163–66.

Jaitin, Diego Adhemar, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, et al. 2014. “Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types.” *Science* 343 (6172): 776–79.

Kivioja, Teemu, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. 2012. “Counting Absolute Numbers of Molecules Using Unique Molecular Identifiers.” *Nature Methods* 9 (1): 72–74.

Klein, Allon M., Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. 2015. “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells.” *Cell* 161 (5): 1187–1201.

Kolodziejczyk, Aleksandra A., Jong Kyoung Kim, Valentine Svensson, John C. Marioni, and Sarah A. Teichmann. 2015. “The Technology and Biology of Single-Cell RNA

Sequencing." *Molecular Cell* 58 (4): 610–20.
Kolodziejczyk, Aleksandra A., Jong Kyoung Kim, Jason C. H. Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N. Natarajan, Alex C. Tuck, et al. 2015. "Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation." *Cell Stem Cell* 17 (4): 471–85.
Kulpa, D., R. Topping, and A. Telesnitsky. 1997. "Determination of the Site of First Strand Transfer during Moloney Murine Leukemia Virus Reverse Transcription and Identification of Strand Transfer-Associated Reverse Transcriptase Errors." *The EMBO Journal* 16 (4): 856–65.
Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." *Genome Biology* 15 (2): R29.
Li, E., T. H. Bestor, and R. Jaenisch. 1992. "Targeted Mutation of the DNA Methyltransferase Gene Results in Embryonic Lethality." *Cell* 69 (6): 915–26.
Lun, Aaron T. L., Karsten Bach, and John C. Marioni. 2016a. "Pooling across Cells to Normalize Single-Cell RNA Sequencing Data with Many Zero Counts." *Genome Biology* 17 (April): 75.
———. 2016b. "Pooling across Cells to Normalize Single-Cell RNA Sequencing Data with Many Zero Counts." *Genome Biology* 17 (April): 75.
Macaulay, Iain C., and Thierry Voet. 2014. "Single Cell Genomics: Advances and Future Perspectives." *PLoS Genetics* 10 (1): e1004126.
Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161 (5): 1202–14.
Parekh, Swati, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. 2016. "The Impact of Amplification on Differential Expression Analyses by RNA-Seq." *Scientific Reports* 6 (May): 25533.
———. 2017. "zUMIs: A Fast and Flexible Pipeline to Process RNA Sequencing Data with UMIs." *bioRxiv*. doi:10.1101/153940.
Petropoulos, Sophie, Daniel Edsgärd, Björn Reinius, Qiaolin Deng, Sarita Pauliina Panula, Simone Codeluppi, Alvaro Plaza Reyes, Sten Linnarsson, Rickard Sandberg, and Fredrik Lanner. 2016. "Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos." *Cell* 0 (0). Elsevier. doi:10.1016/j.cell.2016.03.023.
Pettitt, Stephen J., Qi Liang, Xin Y. Rairdan, Jennifer L. Moran, Haydn M. Prosser, David R. Beier, Kent C. Lloyd, Allan Bradley, and William C. Skarnes. 2009. "Agouti C57BL/6N Embryonic Stem Cells for Mouse Genetic Resources." *Nature Methods* 6 (7): 493–95.
Phipson, Belinda, Luke Zappia, and Alicia Oshlack. 2017. "Gene Length and Detection Bias in Single Cell RNA Sequencing Protocols." *F1000Research* 6 (April). doi:10.12688/f1000research.11290.1.
Picelli, Simone, Åsa K. Björklund, Omid R. Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. 2013. "Smart-seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells." *Nature Methods* 10 (11): 1096–98.
Pollen, Alex A., Tomasz J. Nowakowski, Joe Shuga, Xiaohui Wang, Anne A. Leyrat, Jan H. Lui, Nianzhen Li, et al. 2014. "Low-Coverage Single-Cell mRNA Sequencing Reveals Cellular Heterogeneity and Activated Signaling Pathways in Developing Cerebral Cortex." *Nature Biotechnology*, August. doi:10.1038/nbt.2967.
Quail, Michael A., Thomas D. Otto, Yong Gu, Simon R. Harris, Thomas F. Skelly, Jacqueline A. McQuillan, Harold P. Swerdlow, and Samuel O. Oyola. 2012. "Optimal Enzymes for Amplifying Sequencing Libraries." *Nature Methods* 9 (1): 10–11.
Ramsköld, Daniel, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R. Faridani,

Gregory A. Daniels, et al. 2012. “Full-Length mRNA-Seq from Single-Cell Levels of RNA and Individual Circulating Tumor Cells.” *Nature Biotechnology* 30 (8): 777–82.
Risso, Davide, John Ngai, Terence P. Speed, and Sandrine Dudoit. 2014. “Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples.” *Nature Biotechnology* 32 (9): 896–902.
Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47.
Rosenberg, Alexander B., Charles Roco, Richard A. Muscat, Anna Kuchina, Sumit Mukherjee, Wei Chen, David J. Peeler, et al. 2017. “Scaling Single Cell Transcriptomics through Split Pool Barcoding.” *bioRxiv*. doi:10.1101/105163.
Sasagawa, Y., H. Danno, H. Takada, and M. Ebisawa. 2017. “Quartz-Seq2: A High-Throughput Single-Cell RNA-Sequencing Method That Effectively Uses Limited Sequence Reads.” *bioRxiv*. biorxiv.org. http://www.biorxiv.org/content/early/2017/07/05/159384.abstract.
SEQC/MAQC-III Consortium. 2014. “A Comprehensive Assessment of RNA-Seq Accuracy, Reproducibility and Information Content by the Sequencing Quality Control Consortium.” *Nature Biotechnology* 32 (9): 903–14.
Seyednasrollah, Fatemeh, Krista Rantanen, Panu Jaakkola, and Laura L. Elo. 2015. “ROTS: Reproducible RNA-Seq Biomarker Detector—prognostic Markers for Clear Cell Renal Cell Cancer.” *Nucleic Acids Research*. Oxford Univ Press, gkv806.
Sheng, Kuanwei, Wenjian Cao, Yichi Niu, Qing Deng, and Chenghang Zong. 2017. “Effective Detection of Variation in Single-Cell Transcriptomes Using MATQ-Seq.” *Nature Methods* 14 (3): 267–70.
Soumillon, Magali, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S. Mikkelsen. 2014. “Characterization of Directed Differentiation by High-Throughput Single-Cell RNA-Seq.” *bioRxiv*, March. doi:10.1101/003236.
Stegle, Oliver, Sarah A. Teichmann, and John C. Marioni. 2015. “Computational and Analytical Challenges in Single-Cell Transcriptomics.” *Nature Reviews. Genetics* 16 (3): 133–45.
Svensson, Valentine, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J. Miragaia, Charlotte Labalette, Iain C. Macaulay, Ana Cvejic, and Sarah A. Teichmann. 2017. “Power Analysis of Single-Cell RNA-Sequencing Experiments.” *Nature Methods*, March. doi:10.1038/nmeth.4220.
Tung, Po-Yuan, John D. Blischak, Chiaowen Joyce Hsiao, David A. Knowles, Jonathan E. Burnett, Jonathan K. Pritchard, and Yoav Gilad. 2017. “Batch Effects and the Effective Design of Single-Cell Gene Expression Studies.” *Scientific Reports* 7 (January): 39921.
Vickovic, Sanja, Patrik L. Ståhl, Fredrik Salmén, Sarantis Giatrellis, Jakub Orzechowski Westholm, Annelie Mollbrink, José Fernández Navarro, et al. 2016. “Massive and Parallel Expression Profiling Using Microarrayed Single-Cell Sequencing.” *Nature Communications* 7 (October): 13182.
Vieth, Beate, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. 2017. “powsimR: Power Analysis for Bulk and Single Cell RNA-Seq Experiments.” *Bioinformatics* , July. doi:10.1093/bioinformatics/btx435.
Wagner, Allon, Aviv Regev, and Nir Yosef. 2016. “Revealing the Vectors of Cellular Identity with Single-Cell Genomics.” *Nature Biotechnology* 34 (11): 1145–60.
Wu, Angela R., Norma F. Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E. Rothenberg, Francis M. Mburu, et al. 2014. “Quantitative Assessment of Single-Cell RNA-Sequencing Methods.” *Nature Methods* 11 (1): 41–46.
Young, Lesley, Julia Sung, Glyn Stacey, and John R. Masters. 2010. “Detection of Mycoplasma in Cell Cultures.” *Nature Protocols* 5 (5): 929–34.

Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. “Massively Parallel Digital Transcriptional Profiling of Single Cells.” *Nature Communications* 8 (January): 14049.
Zhu, Y. Y., E. M. Machleder, A. Chenchik, R. Li, and P. D. Siebert. 2001. “Reverse Transcriptase Template Switching: A SMART Approach for Full-Length cDNA Library Construction.” *BioTechniques* 30 (4): 892–97.
Ziegenhain, Christoph, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. 2017. “Comparative Analysis of Single-Cell RNA Sequencing Methods.” *Molecular Cell* 65 (4): 631–43.e4.
Zimmerman, S. B., and B. H. Pheiffer. 1983. “Macromolecular Crowding Allows Blunt-End Ligation by DNA Ligases from Rat Liver or Escherichia Coli.” *Proceedings of the National Academy of Sciences of the United States of America* 80 (19): 5852–56.

Author contributions

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

Acknowledgments

We thank Ines Bliesener for expert technical assistance. We are grateful to Magali Soumillon and Tarjei Mikkelsen for providing the SCRB-seq protocol and to Stefan Krebs and Helmut Blum for sequencing.

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

Competing interests

The authors declare no competing interests.

Data availability

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

Analysis code to reproduce all main Figures can be found at:

https://github.com/cziegenhain/Bagnoli_2017.

Figure Legends

Figure 1: Schematic overview.
A) Low amounts of universal human reference RNA (UHRR) were used in optimization experiments. We assessed library preparation components affecting reverse transcription and PCR amplification in regards to cDNA yield, cDNA quality and sensitivity. At the stage of reverse transcription, 9 reverse transcriptase enzymes, reaction enhancers and primer modifications were investigated. At the stage of DNA amplification, we benchmarked 12 high-fidelity polymerases (see methods).
B) Overview of the mcSCRB-seq protocol workflow. Single cells are isolated via FACS in multiwell plates containing lysis buffer comprising barcoded oligo-dT primers and Proteinase K. Reverse transcription and template-switching is carried out in the presence of PEG 8000 to induce molecular crowding. After pooling of barcoded cDNA using magnetic SPRI beads, optimized PCR amplification is performed.
C) Sequencing data of mcSCRB-seq libraries is processed using the zUMIs pipeline (Parekh et al. 2017). After filtering of cells, we benchmark the protocol's performance in terms of sensitivity and power to detect differential gene expression (Vieth et al. 2017).

Figure 2: Optimizing reverse transcription sensitivity.
A) cDNA yield (ng) after reverse transcription and amplification using oligo-dT primers already in the lysis buffer ("in Lysis"; blue) or separately added before reverse transcription ("in RT"). Each dot represents a replicate and each box represents the median and first and third quartiles.
B) cDNA yield dependent on the absence (grey) or presence of 7.5 % PEG 8000 (blue) during reverse transcription. Each dot represents a replicate. Lines represent a linear model fit of the data.
C) cDNA yield (ng) dependent on UHRR input using 9 different RT enzymes. Each dot represents a replicate and fit lines were created using local regression of data points.

Figure 3: Molecular crowding increases sensitivity.
A-D) RNA-seq libraries were generated from 10 pg of UHRR using four protocol variants (see Supplementary Table 1).
A) cDNA yield (ng) after PCR amplification per method. Each dot represents a replicate and each box represents the median and first and third quartiles per method.
B) Number of genes detected (>1 UMI) per replicate. Each dot represents a replicate and each box represents the median and first and third quartiles per method.
C) Number of unique molecular identifiers per replicate. Each dot represents a replicate and each box represents the median and first and third quartiles per method.
D) Gene dropout probability (0-1) over all replicates for each method.

Figure 4: mcSCRB-seq detects large fractions of the cellular transcriptome.
A) Relative increase in the median of detected UMIs dependent on raw sequencing depth (reads) using mcSCRB-seq compared to SCRB-seq. Each point represents the median over all cells at the given sequencing depth. The size of each point depicts the number of cells that were considered to calculate the median. The 95 % confidence interval of a local regression model is depicted by the shaded area.

B) Percentage of cellular mRNA content than can be detected with SCRB-seq (green) or mcSCRB-seq (blue) dependent on the sequencing depth (reads). Each box represents the median and first and third quartiles per sequencing depth and method.

Figure 5: mcSCRB-seq is the most sensitive protocol using ERCC spike-ins.
A-C) Detection of ERCC spike-in transcripts was modeled using a binomial logistic regression relative to the input molecule number.
A) Shown is the detection of the 92 ERCC transcripts in an average mcSCRB-seq cell at 2 million reads coverage. Points and solid line represent the ERCC genes with their logistic regression model. Dashed lines and label indicate the number of ERCC molecules required for a detection probability of 50 %.
B) Number of ERCC molecules required for 50 % detection probability dependent on the sequencing depth (reads) for mcSCRB-seq. Each dot represents an outlier and each box represents the median and first and third quartiles of cells per sequencing depth. A non-linear asymptotic fit is depicted as a solid black line.
C) Number of ERCC molecules required for 50 % detection probability for various library preparation protocols. Per-cell distributions are shown using violin plots, vertical lines and labels depict the median per protocol.

Figure 6: mcSCRB-seq is highly powerful and efficient.
A) Power simulations were performed using the powsimR package (Vieth et al. 2017). For SCRB-seq and mcSCRB-seq, we simulated *n*-cell two-group differential gene expression experiments with 10% differentially expressed genes. Shown are true positive rate (“TPR”) and false discovery rate (“FDR”) for sample sizes n = 24, n = 48, n = 96, n = 192 and n = 384 per group. Boxplots represent the median and first and third quartiles of 25 simulations.
B) Library preparation costs per cell were calculated for 96-well or 384-well scenarios. Colors indicate the consumable type. (see Supplementary Table 3)
C) Library preparation time for one plate of mcSCRB-seq libraries were measured for bench times (“Hands-on”) and incubation times (“Hands-off”). Colors indicate the library preparation step. The total time was 7.5 hours. (see Supplementary Table 4)

Figure 1

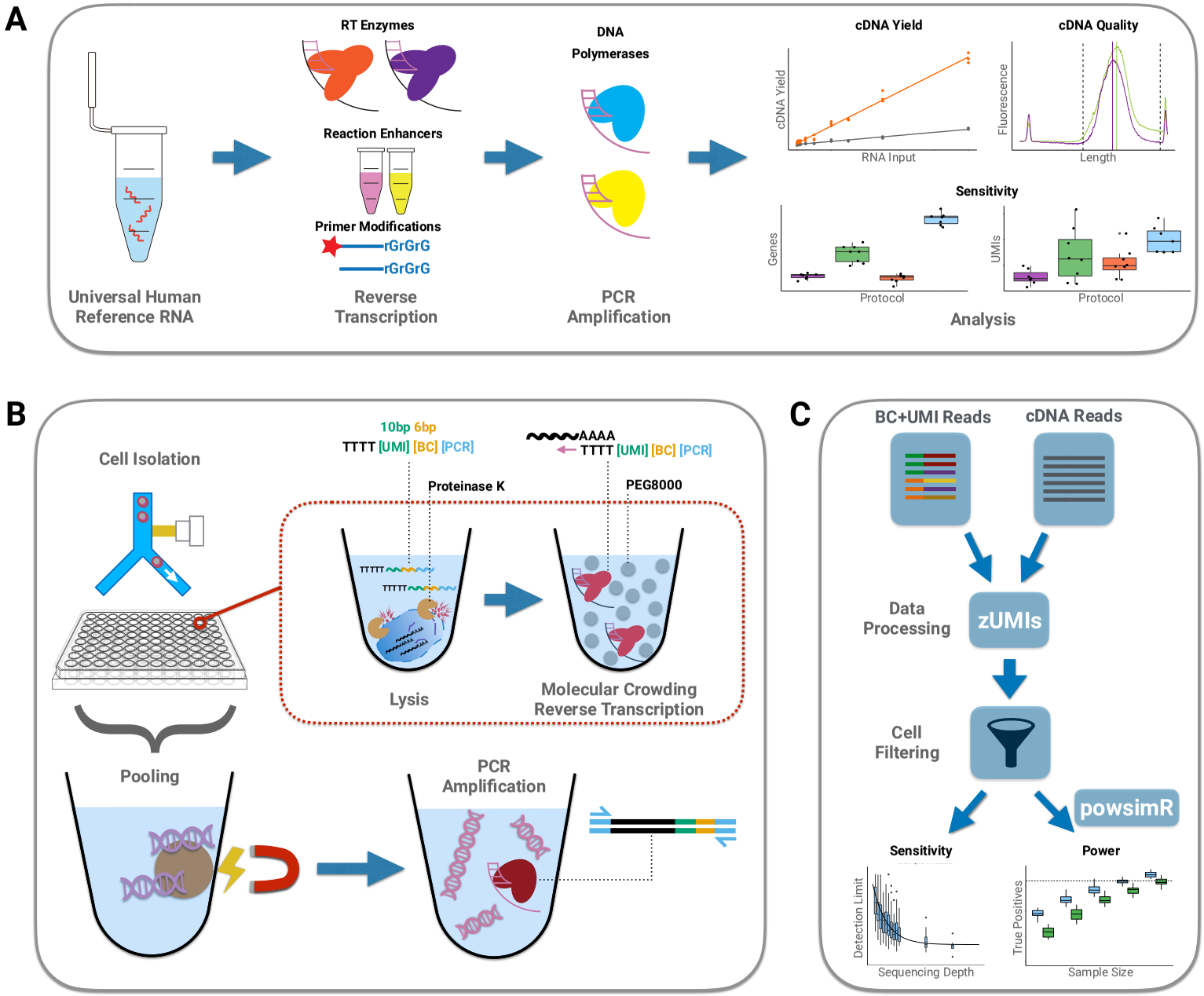

A
RT Enzymes
Reaction Enhancers
Primer Modifications
rGrGrG
rGrGrG
DNA
Polymerases
cDNA Yield
cDNA Quality
cDNA Yield
RNA Input
Fluorescence
Length
Sensitivity
Genes
UMIs
Protocol
Protocol
Universal Human
Reference RNA
Reverse
Transcription
PCR
Amplification
Analysis
B
Cell Isolation
10bp 6bp
TTTT [UMI] [BC] [PCR]
AAAA
TTTT [UMI] [BC] [PCR]
Proteinase K
PEG8000
Lysis
Molecular Crowding
Reverse Transcription
Pooling
PCR
Amplification
C
BC+UMI Reads
cDNA Reads
Data
Processing
zUMIs
Cell
Filtering
powsimR
Sensitivity
Power
Detection Limit
Sequencing Depth
True Positives
Sample Size

Figure 2

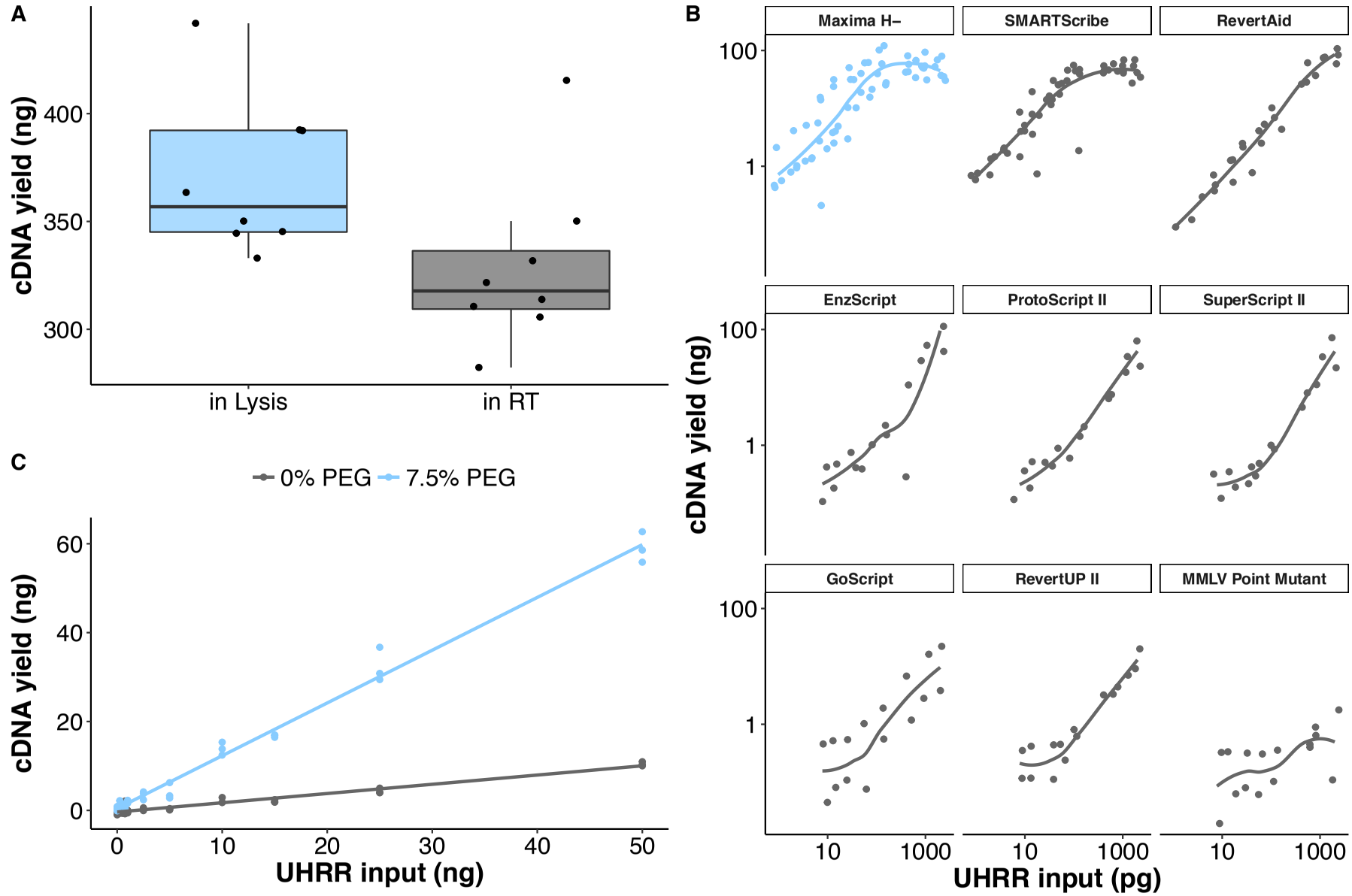

A
cDNA yield (ng)
400
350
300
in Lysis
in RT
B
Maxima H–
SMARTScribe
RevertAid
EnzScript
ProtoScript II
SuperScript II
GoScript
RevertUP II
MMLV Point Mutant
cDNA yield (ng)
100
1
10
1000
UHRR input (pg)
C
0% PEG
7.5% PEG
cDNA yield (ng)
60
40
20
0
0
10
20
30
40
50
UHRR input (ng)

Figure 3

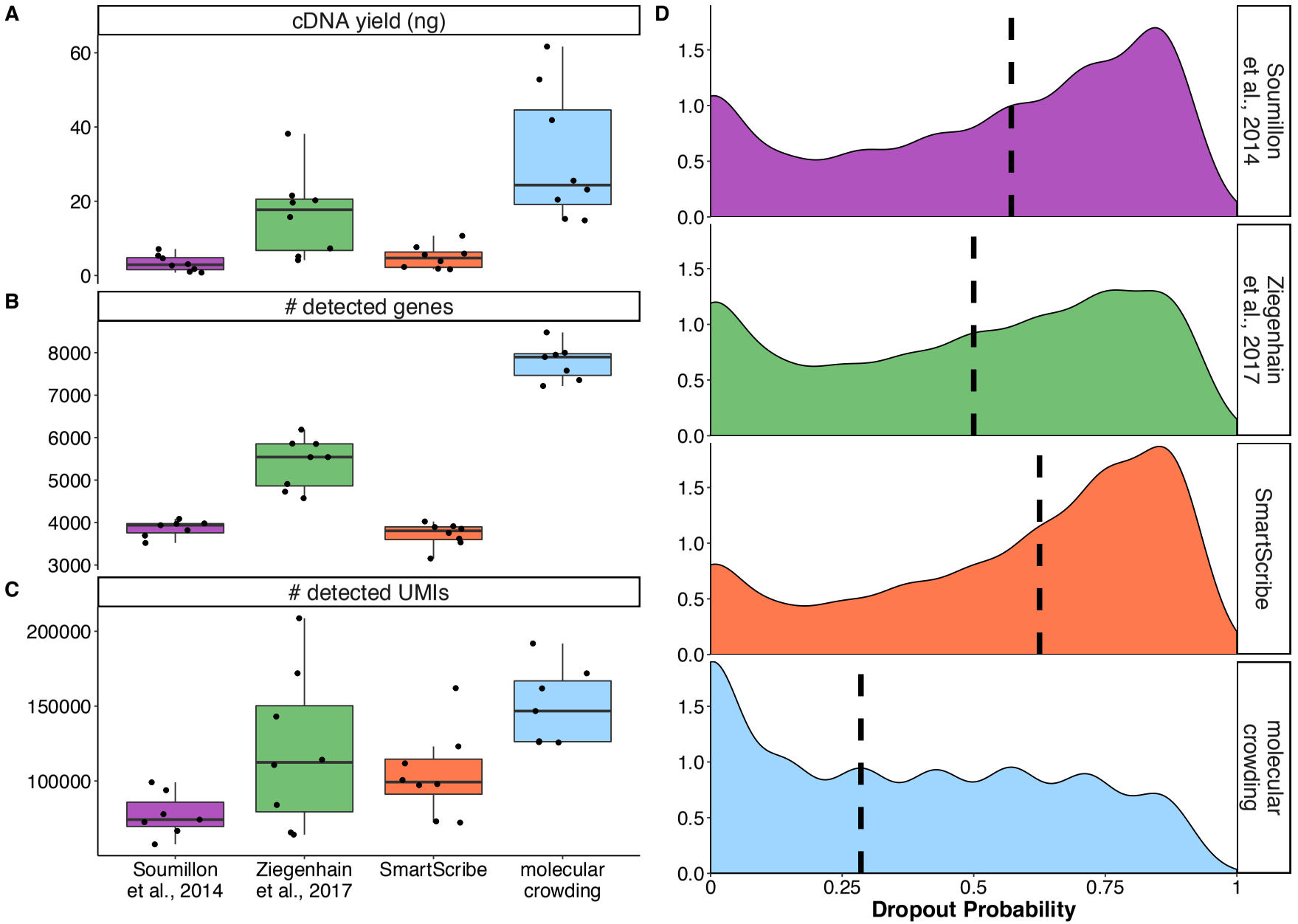

A
cDNA yield (ng)
60
40
20
0
B
detected genes
8000
7000
6000
5000
4000
3000
C
detected UMIs
200000
150000
100000
Soumillon et al., 2014
Ziegenhain et al., 2017
SmartScribe
molecular crowding
D
1.5
1.0
0.5
0.0
Soumillon et al., 2014
Ziegenhain et al., 2017
SmartScribe
molecular crowding
0
0.25
0.5
0.75
1
Dropout Probability

Figure 4

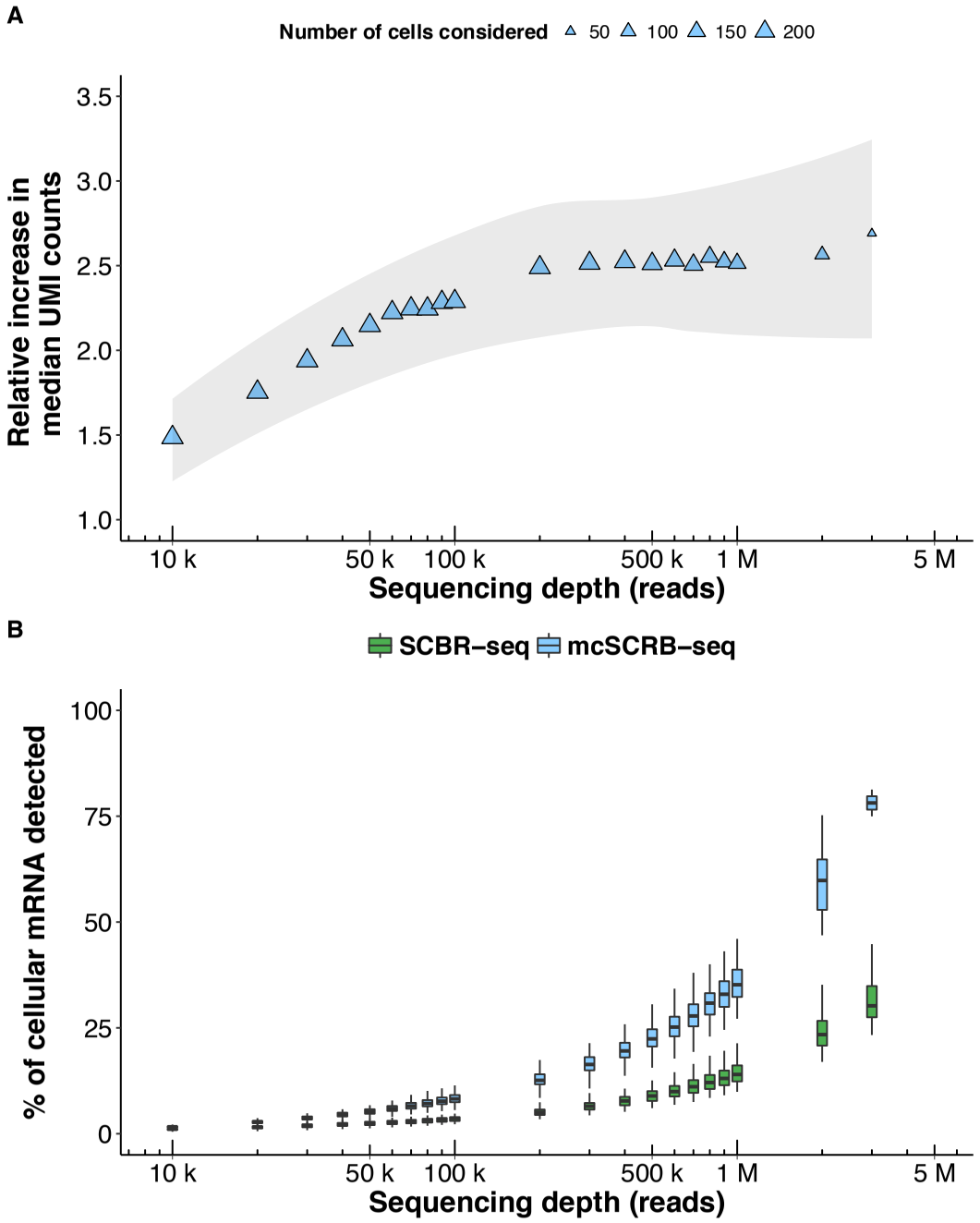

A
Number of cells considered 50 100 150 200
Relative increase in median UMI counts
3.5
3.0
2.5
2.0
1.5
1.0
10 k
50 k
100 k
500 k
1 M
5 M
Sequencing depth (reads)
B
SCBR−seq
mcSCRB−seq
% of cellular mRNA detected
100
75
50
25
0
10 k
50 k
100 k
500 k
1 M
5 M
Sequencing depth (reads)

Figure 5

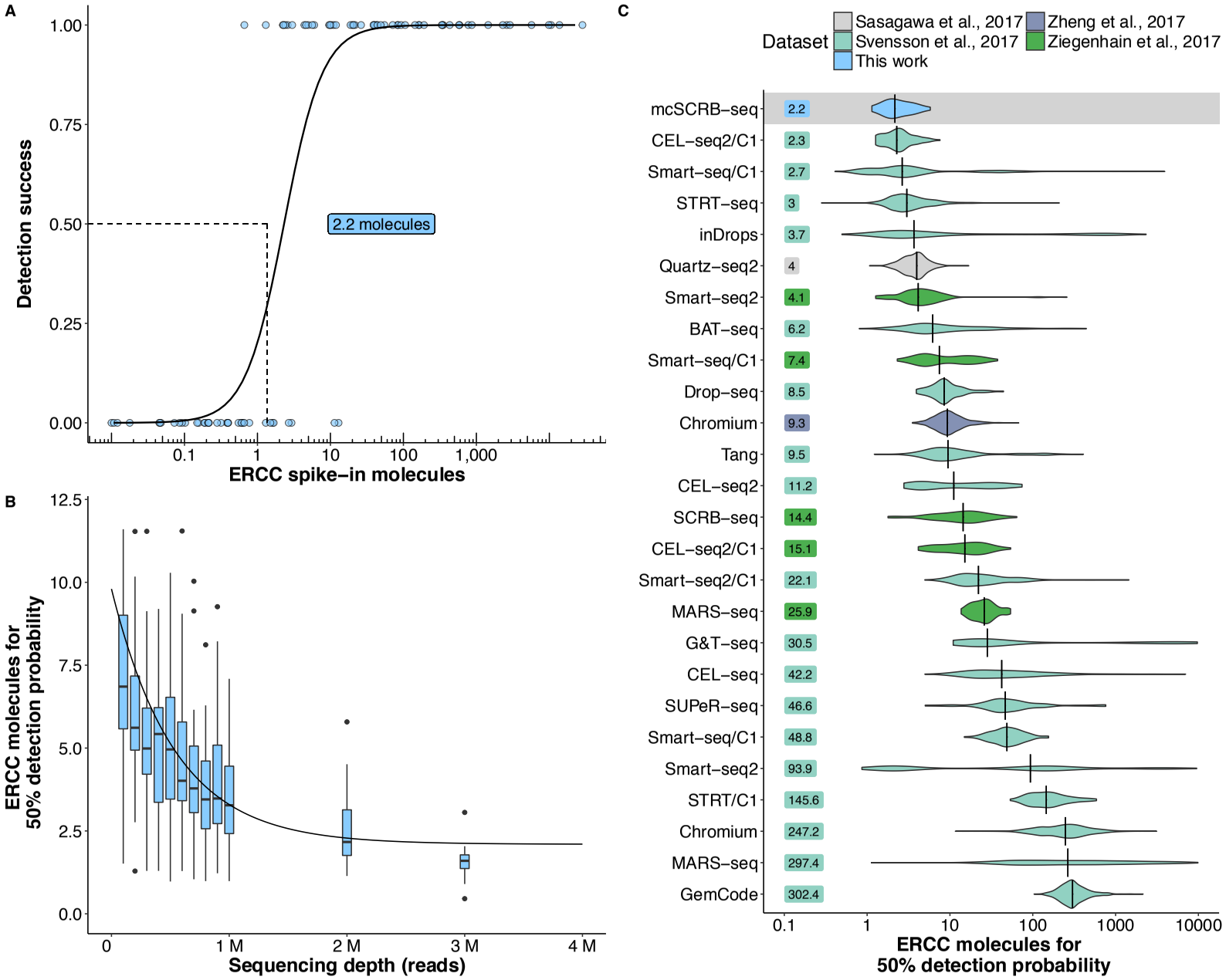
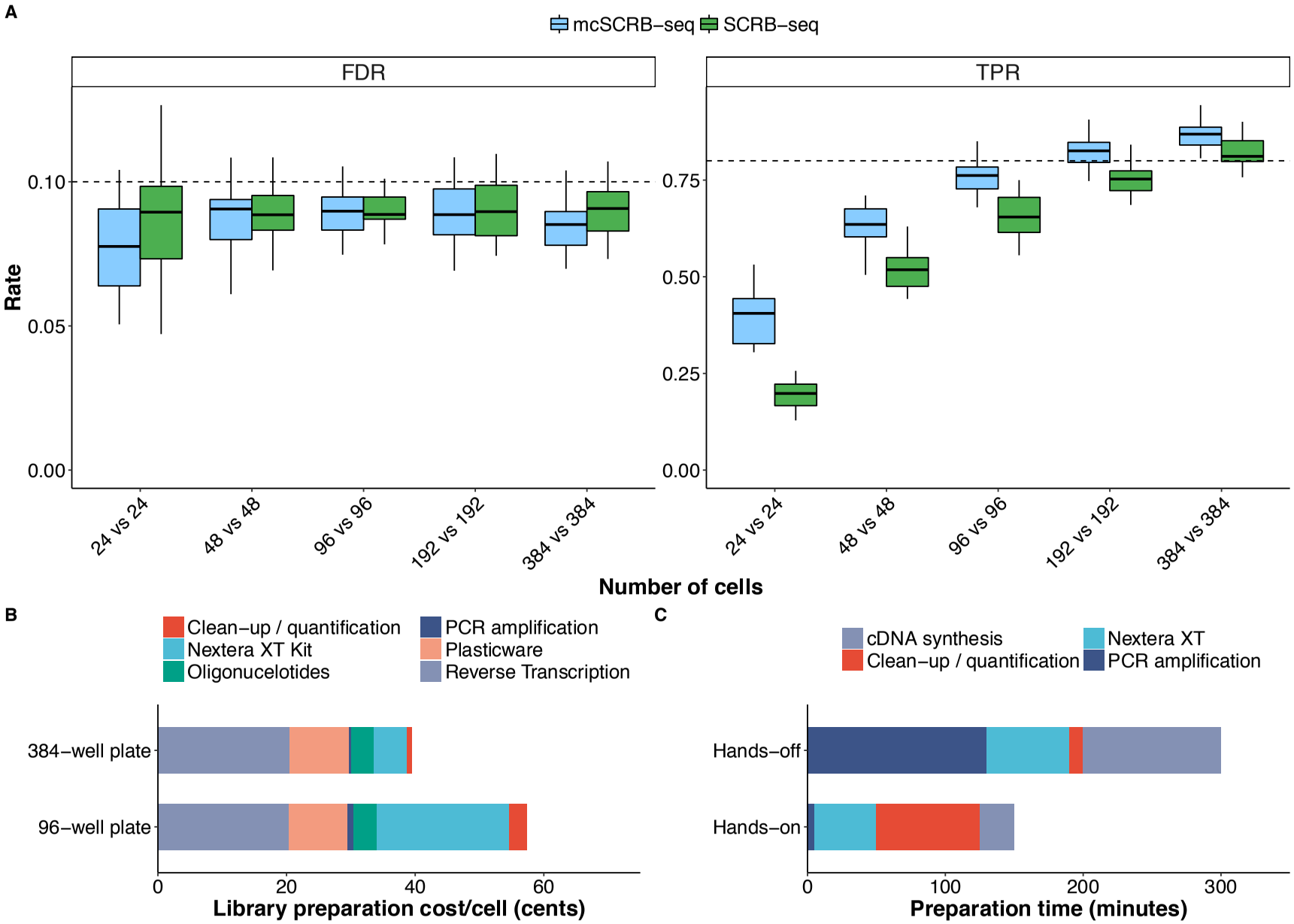
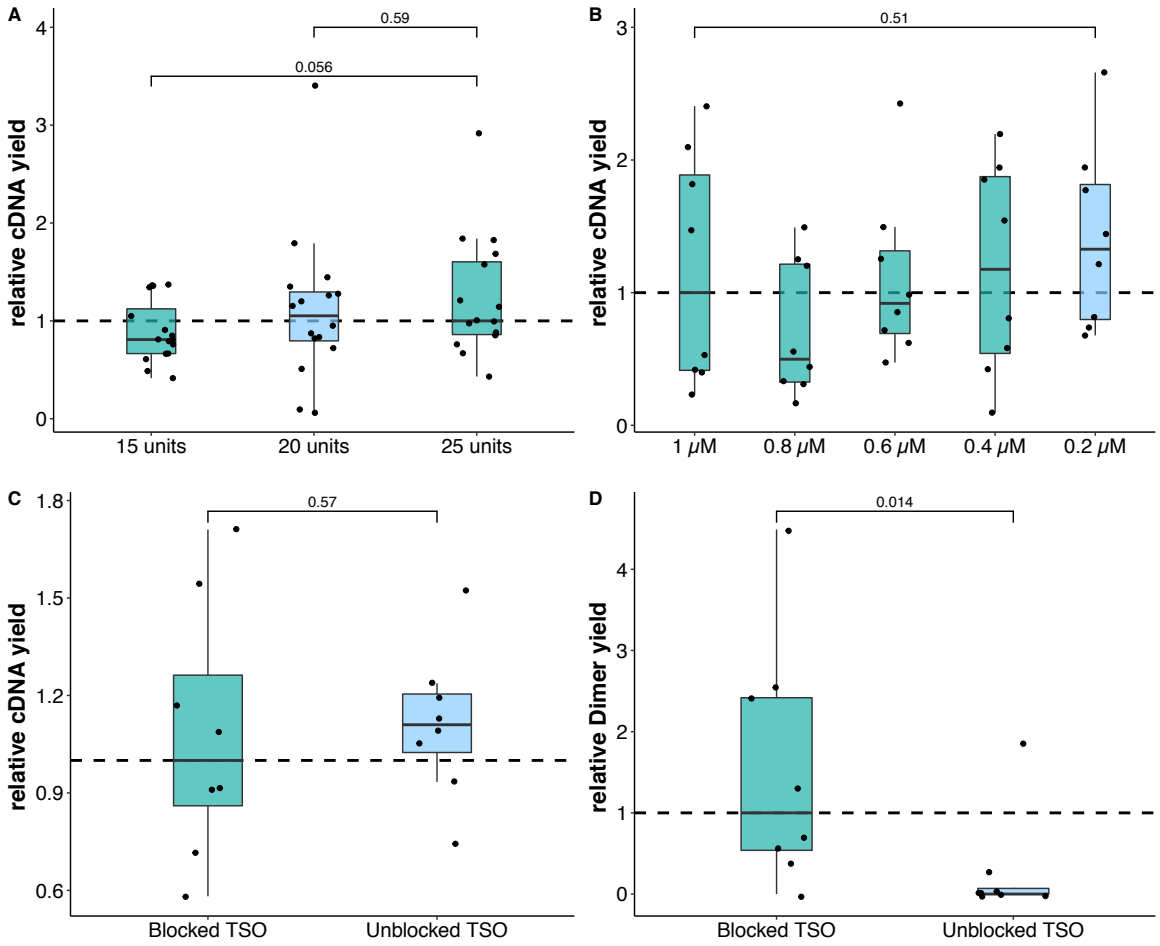

A
Detection success
1.00
0.75
0.50
0.25
0.00
2.2 molecules
0.1
1
10
100
1,000
ERCC spike−in molecules
B
ERCC molecules for 50% detection probability
12.5
10.0
7.5
5.0
2.5
0.0
0
1 M
2 M
3 M
4 M
Sequencing depth (reads)
C
Dataset
Sasagawa et al., 2017
Svensson et al., 2017
This work
Zheng et al., 2017
Ziegenhain et al., 2017
mcSCRB−seq 2.2
CEL−seq2/C1 2.3
Smart−seq/C1 2.7
STRT−seq 3
inDrops 3.7
Quartz−seq2 4
Smart−seq2 4.1
BAT−seq 6.2
Smart−seq/C1 7.4
Drop−seq 8.5
Chromium 9.3
Tang 9.5
CEL−seq2 11.2
SCRB−seq 14.4
CEL−seq2/C1 15.1
Smart−seq2/C1 22.1
MARS−seq 25.9
G&T−seq 30.5
CEL−seq 42.2
SUPeR−seq 46.6
Smart−seq/C1 48.8
Smart−seq2 93.9
STRT/C1 145.6
Chromium 247.2
MARS−seq 297.4
GemCode 302.4
0.1
1
10
100
1000
10000
ERCC molecules for 50% detection probability

Figure 6


A
mcSCRB−seq
SCRB−seq
FDR
TPR
Rate
0.10
0.05
0.00
0.75
0.50
0.25
0.00
24 vs 24
48 vs 48
96 vs 96
192 vs 192
384 vs 384
Number of cells
B
Clean−up / quantification
Nextera XT Kit
Oligonucelotides
PCR amplification
Plasticware
Reverse Transcription
384−well plate
96−well plate
0
20
40
60
Library preparation cost/cell (cents)
C
cDNA synthesis
Clean−up / quantification
Nextera XT
PCR amplification
Hands−off
Hands−on
0
100
200
300
Preparation time (minutes)

Supplementary Figure 1



Optimization of reverse transcription conditions

A-C) Shown are relative cDNA yields after reverse transcription and PCR amplification of 1 ng UHRR per replicate using:

A) varying amounts of reverse transcriptase enzyme (15-25 units, Maxima H-),

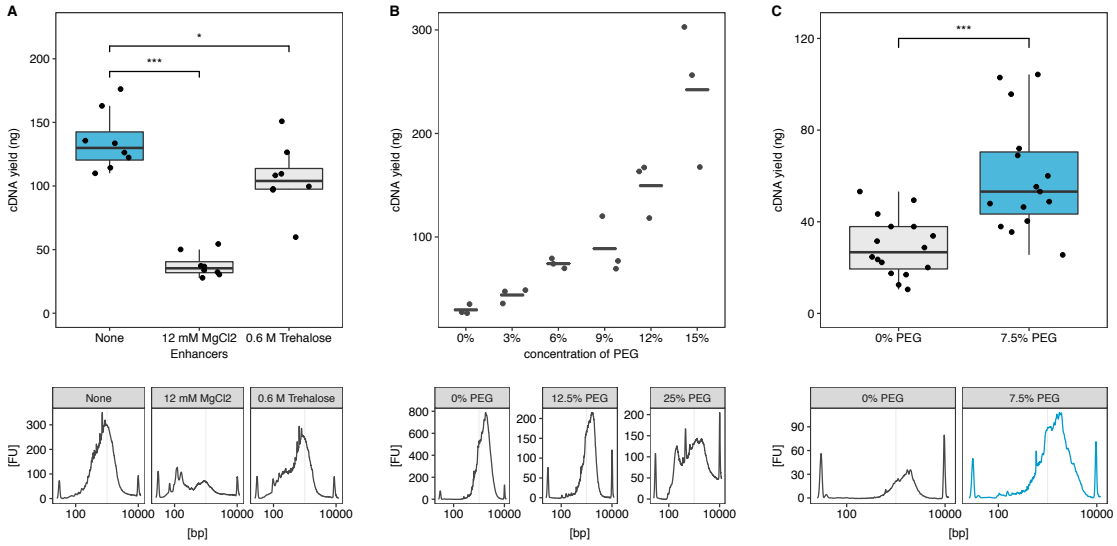
B) varying amounts of oligo-dT primer (E3V6),

C) blocked or unblocked Template switching oligo (TSO, E5V6).

D) Relative primer dimer yield using blocked or unblocked Template switching oligo (TSO, E5V6).

All values are relative to the median of the condition used in the original SCRB-seq protocol (Soumillon et al. 2014), which is indicated by a dashed horizontal line. Each dot represents a replicate and each box represents the median and first and third quartiles method. Numbers above boxes indicated p values (Welch Two Sample t-test). Conditions selected for the mcSCRB-seq protocol are marked in blue.

Supplementary Figure 2



Reverse transcription yield is increased by molecular crowding

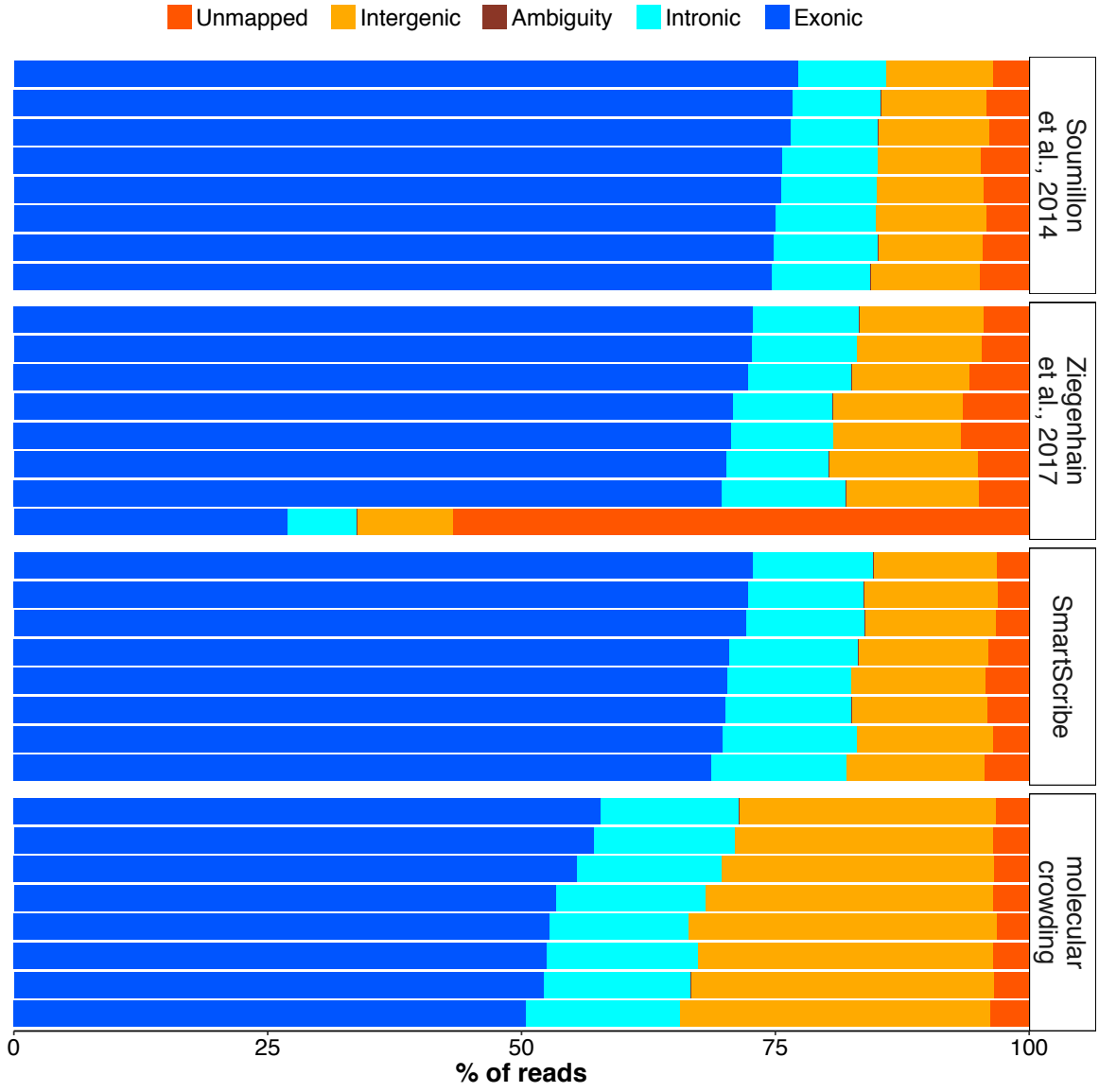
A-C) cDNA yield (top) as well as representative length distributions (capillary gel electrophoresis, bottom) using different indicated reverse transcription enhancers. Each dot represents a replicate.

A) Influence of $MgCl_2$ and Trehalose on yield was investigated. Boxes represent median and first and third quartiles per condition. Numbers above boxes indicate p-values (Welch Two Sample t-test).

B) Concentration-dependant influence of PEG on cDNA yield was investigated. Lines represent the median per concentration.

C) Effect of 7.5 % PEG8000 in reverse transcription was investigated. Boxes represent median and first and third quartiles per condition. Numbers above boxes indicate p-values (Welch Two Sample t-test).

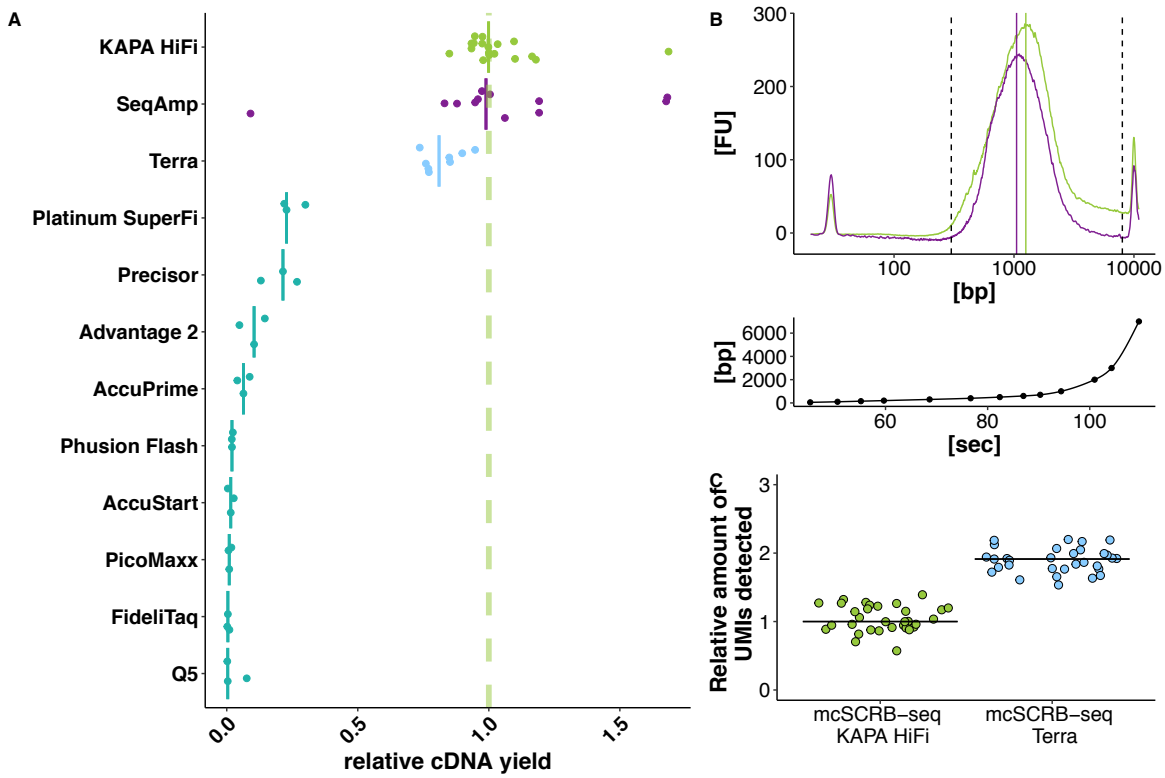
Supplementary Figure 3



Sequencing of UHRR samples

Libraries were generated from 10 pg of UHRR input in four protocol variants (Supplementary Table 1). Shown are the percentage of sequencing reads that cannot be mapped to the human genome (red), mapped to ambiguous genes (brown), mapped to intergenic regions (orange), inside introns (teal) or inside exons (blue).

Supplementary Figure 4



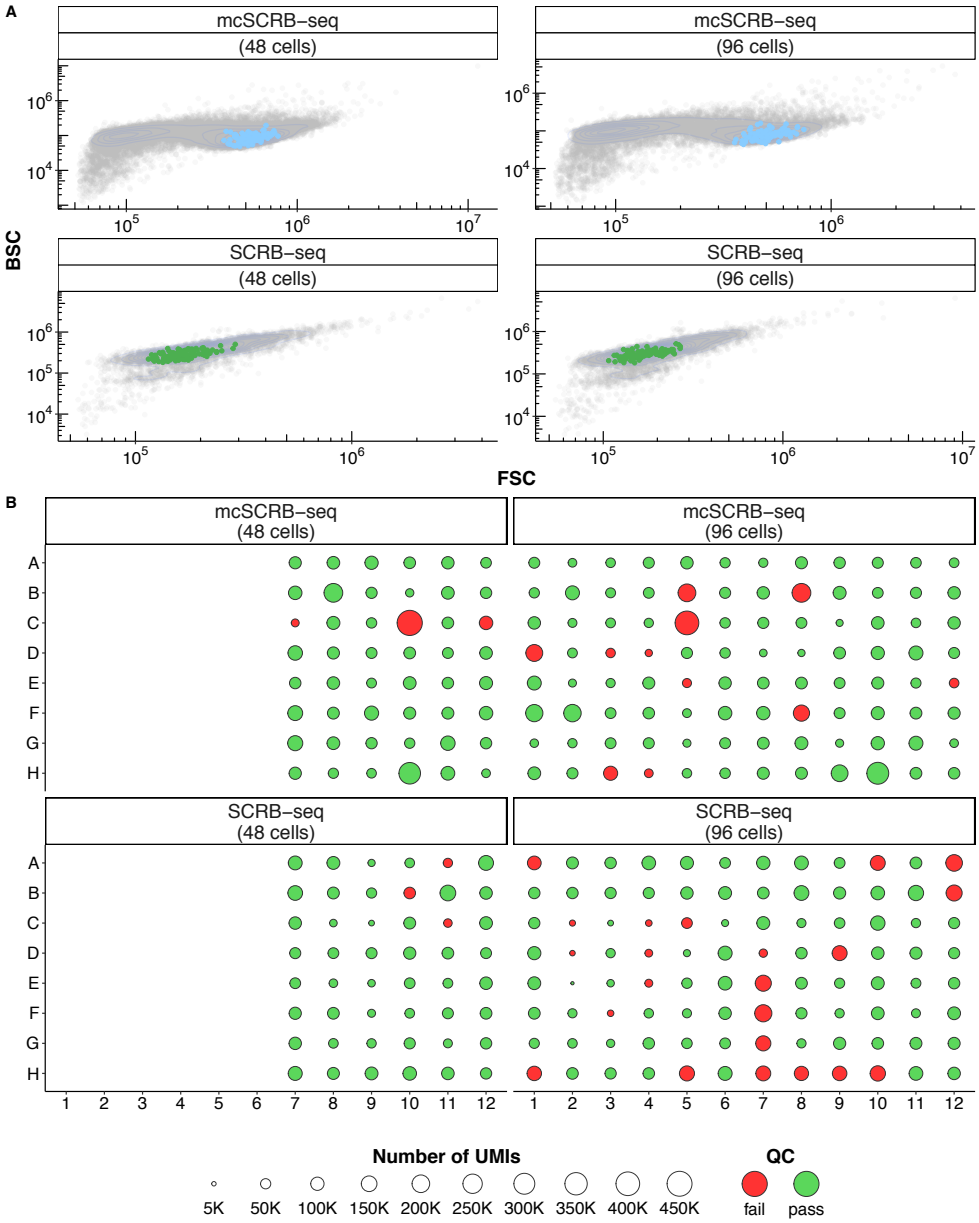
Optimization of PCR amplification

A) Relative cDNA yield after reverse transcription and amplification using different polymerase enzymes or ready mixes. All values are relative to the median of KAPA HiFi which is indicated by a dashed vertical line, as this was used in the SCRB-seq protocol variant of Ziegenhain et al., 2017. Solid vertical lines indicate the median for each polymerase.

B) Top: Representative length quantification of cDNA libraries amplified with Kapa HiFi (green) or SeqAmp (purple) as quantified by capillary gel electrophoresis (Agilent Bioanalyzer). Solid vertical lines depict the ranked mean length for each library within the region marked with dashed vertical lines. Bottom: Depiction of time length model (spline fit) used to analyze capillary gel electrophoresis via the ladder. Each dot represents a ladder peak with known length (bp) and measurement time (sec).

C) Relative amount of detected UMIs using KAPA-HiFi or Terra for cDNA amplification. For both conditions, molecular crowding reverse transcription was used. Each dot represents a replicate and each horizontal line indicates the median per polymerase.

Supplementary Figure 5

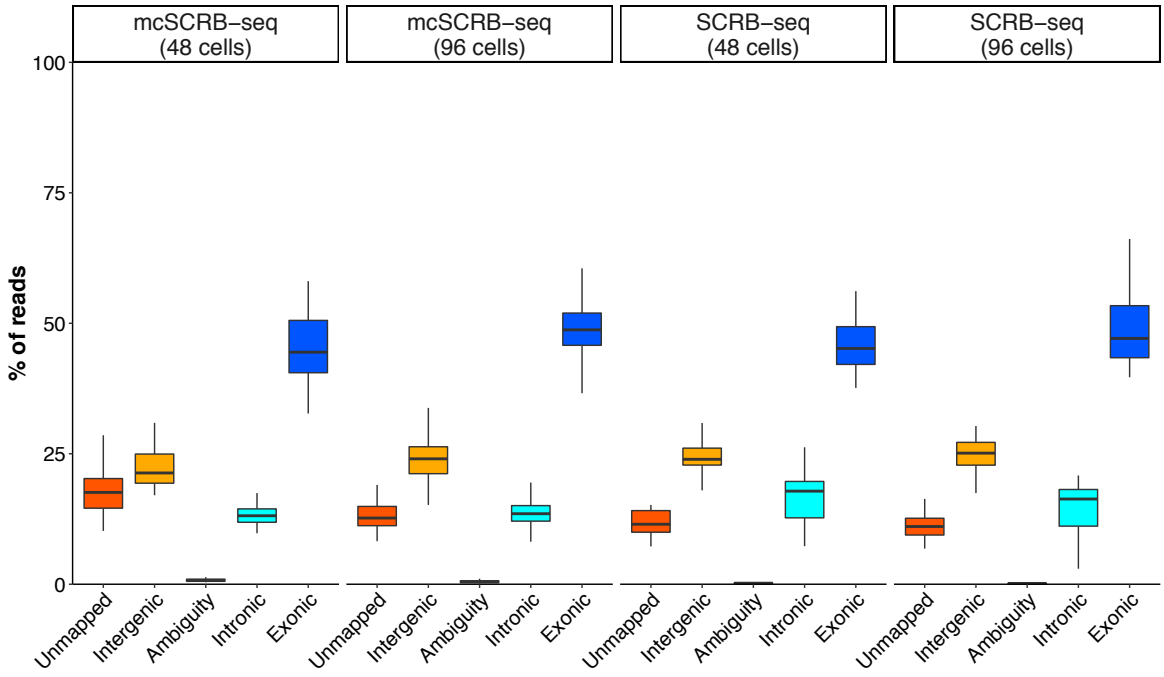


Library quality of JM8 mESC samples

A) Scatter plots showing FACS data with forward (FSC) and backward (BSC) scatter intensities of JM8 mESCs. Each point represents an event. Coloured points represent events that were sorted for scRNA-seq libraries. Library batches are depicted as facets.

B) UMI counts for each cell by method (SCRB-seq/mcSCRB-seq) and replicate (48 cells/96 cells) are shown in their respective position in 96-well plates. Point sizes indicate the number of detected UMIs. Colouring indicates whether a cell passed (green) or failed (red) the Quality Control (QC) as described (see Methods).

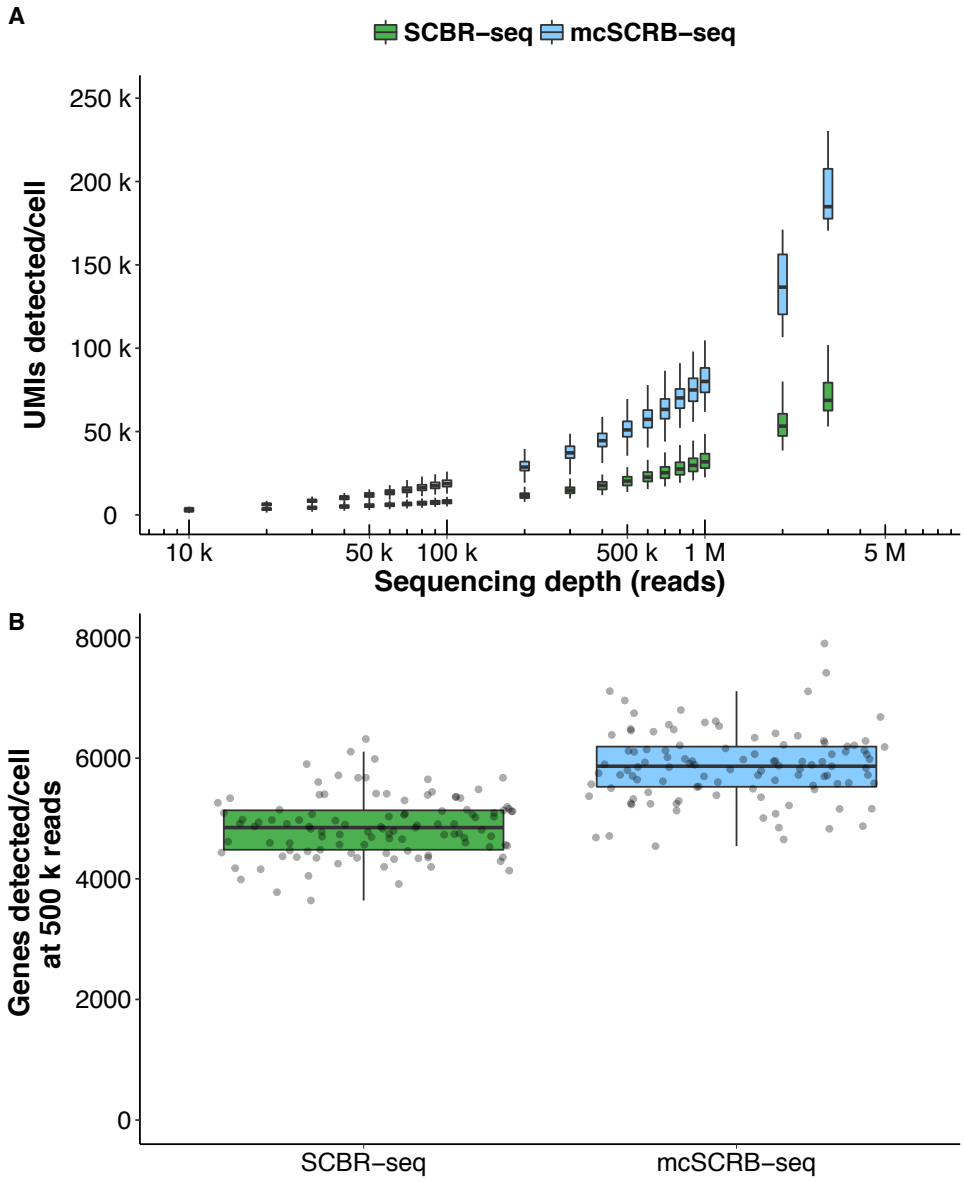
Supplementary Figure 6



Feature distribution of JM8 mESC samples

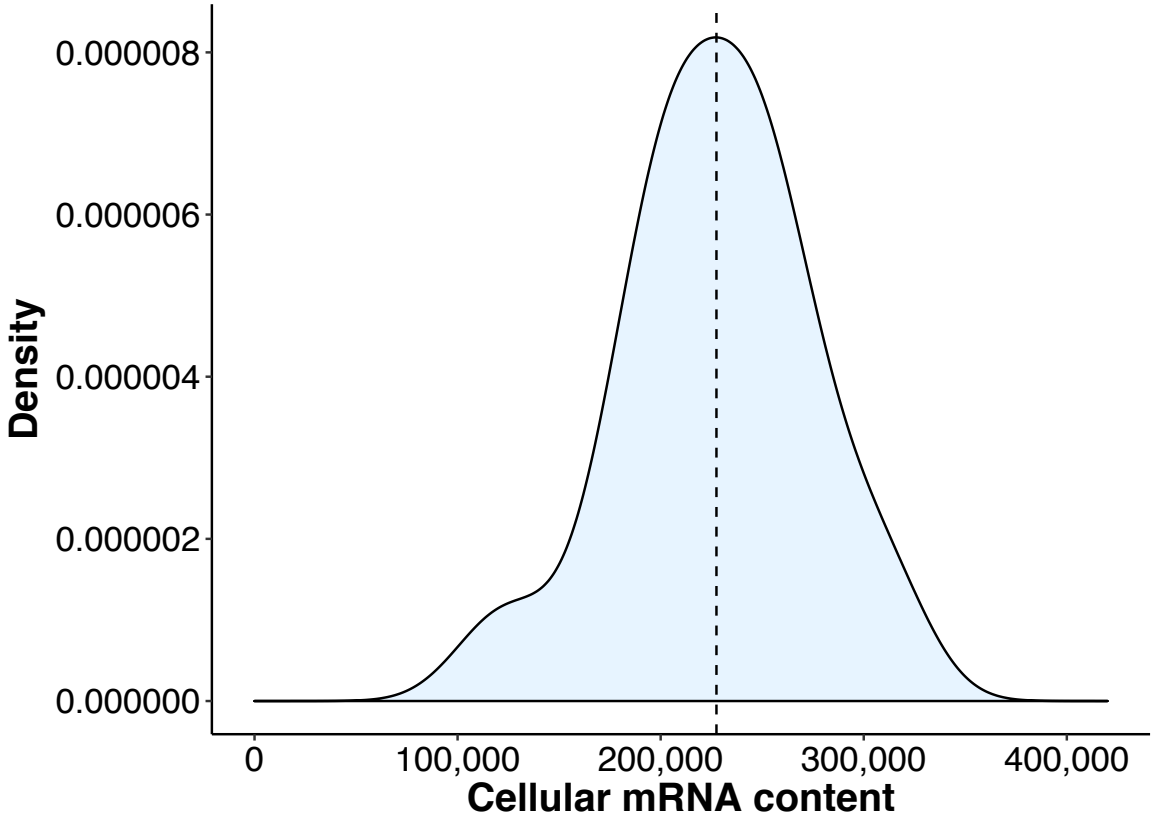
Percentage of reads from mESC based experiments that cannot be mapped to the human genome (red) are mapped ambiguously(brown), are mapped to intergenic regions (orange), inside introns (teal) or inside exons (blue). Each box represents the median and first and third quartiles of cells per method.

Supplementary Figure 7



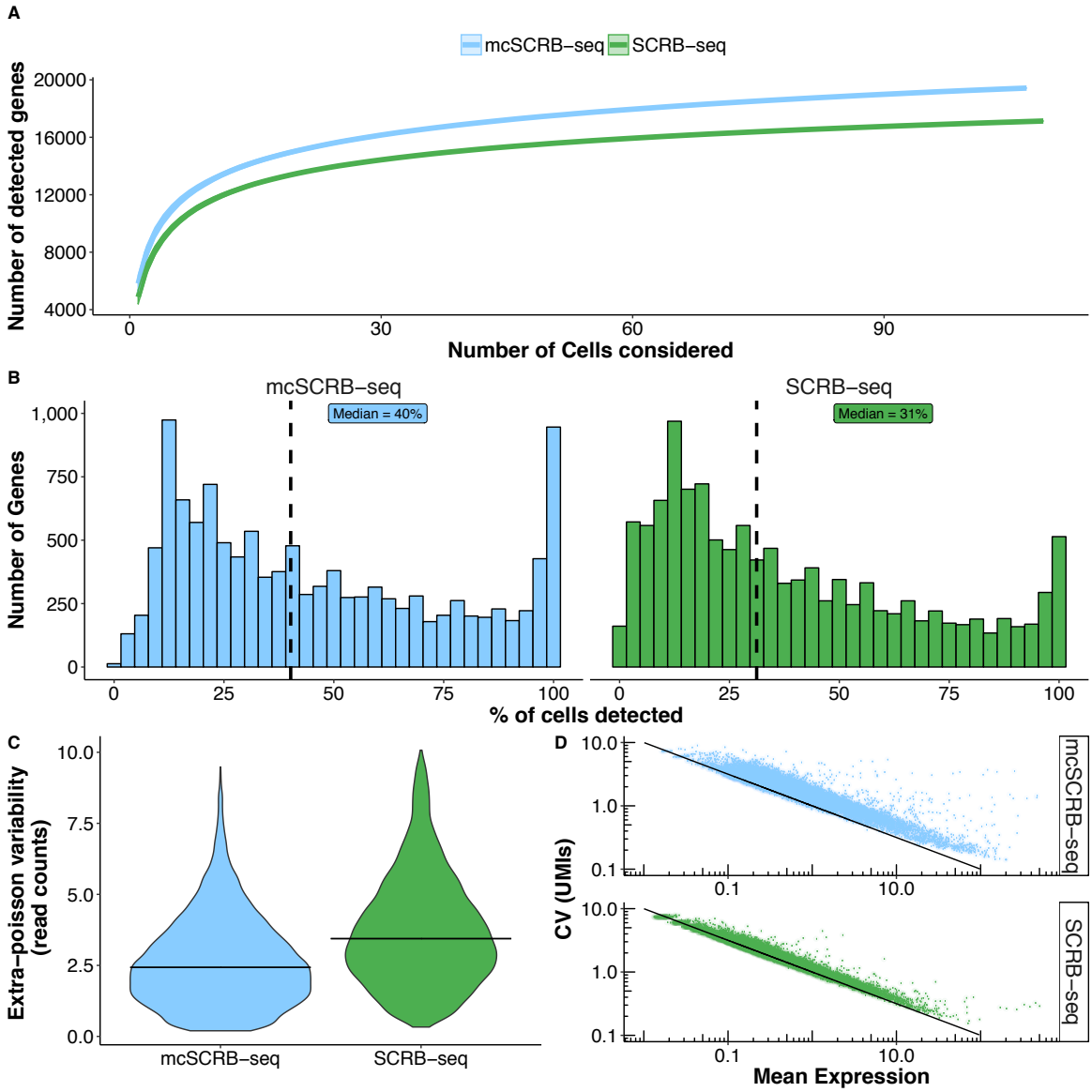
Sensitivity of SCRB-seq and mcSCRB-seq protocols
A) Number of detected UMIs dependent on sequencing depth (reads). Each box represents the median and first and third quartiles per sequencing depth and method. Sequencing depths are scaled logarithmically (base 10).
B) Number of detected genes per cell and method (SCRB-seq/mcSCRB-seq) at a sequencing depth of 500,000 reads per cell (downsampled). Each dot represents a cell and each box represents the median and first and third quartiles.

Supplementary Figure 8



Estimation of cellular mRNA content
For each cell, cellular mRNA content was estimated using ERCC spike-ins (see methods). Shown is the distribution of estimated mRNA counts over all cells.

Supplementary Figure 9



Descriptive statistics for mESC-based experiments at 500,000 reads

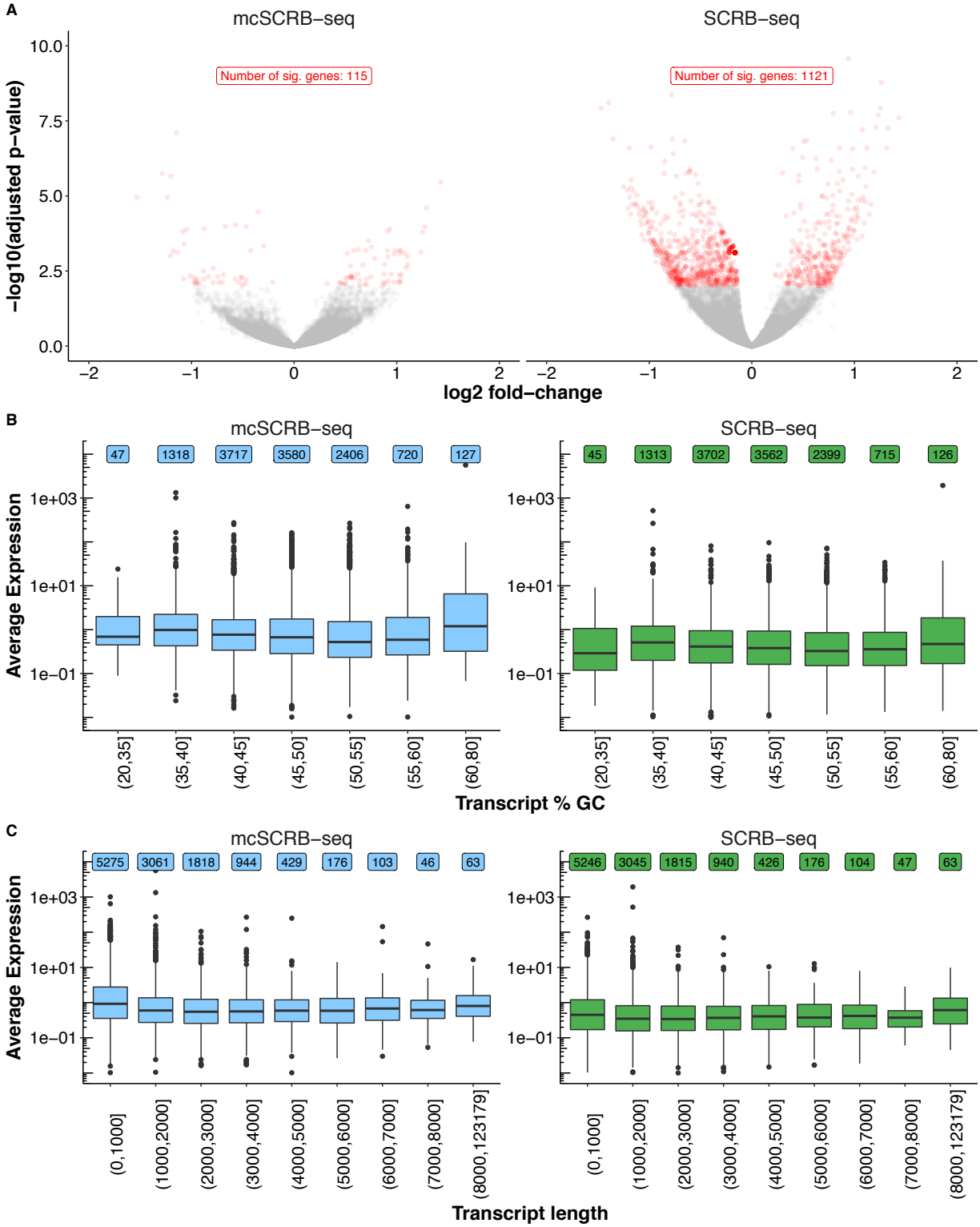
A) Number of detected genes dependent on the number of cells considered.

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

C) Extra Poisson variability across 12086 reliably detected genes (detected in > 10% of cells) was calculated by subtracting the expected amount of variation due to Poisson sampling from the coefficient of variation (CV) measured in read-count quantification. Distributions are shown as violin plots and medians are shown as bars.

D) Gene-wise mean and coefficient of variation from all cells are shown as scatterplots for all methods based on UMI counts. The black line indicates variance according to the poisson distribution.

Supplementary Figure 10



Minimal batch effects and biases in mcSCRB-seq.
A) Volcano plots show differentially expressed genes between plates for each method. Points in red depict significantly differentially expressed genes (limma-voom; FDR < 0.01). Red labels show the number of differentially expressed genes between batches.
B-C) Each dot represents an outlier and each box represents the median and first and third quartiles.
B) Average detected gene-wise expression levels (log normalized UMI) dependent on GC content of each transcript. Transcripts are grouped in 7 bins of GC content.
C) Average detected gene-wise expression levels (log normalized UMI) dependent on transcript length. Transcripts are grouped in 7 bins of length.

Supplementary Table 1

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

Table S1 (related to Figure 3): Overview of used enzymes and enhancers in UHRR based experiments.

Supplementary Table 2

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

Table S2 (related to Figure 4/5/6): Overview of the key differences between SCRB-seq as used in Ziegenhain et al., 2017 and mcSCRB-seq (this work).

Supplementary Table 3

consumable	price/unit	# 96 plates	# 384 plates	price/96 plate	price/384 plate
Barcode oligo-dT	24.000,00 €	20000	5000	1,20 €	4,80 €
TSO E5V6unblocked	453,40 €	200	50	2,27 €	9,07 €
Maxima RT	355,00 €	20	5	17,75 €	71,00 €
Exonuclease I	310,00 €	1000	1000	0,31 €	0,31 €
Clontech Terra	500,00 €	800	800	0,63 €	0,63 €
Nextera XT	1.900,00 €	96	96	19,79 €	19,79 €
dNTPs	927,00 €	500	125	1,85 €	7,42 €
Beads	20,00 €	10	10	2,00 €	2,00 €
Picogreen	233,00 €	400	400	0,58 €	0,58 €
PCR Seal	375,00 €	1000	1000	0,38 €	0,38 €
PCR Plate/96	116,00 €	25	0	4,64 €	0,00 €
PCR Plate/384	162,00 €	0	25	0	6,48 €
Tips/96	36,50 €	10	0	3,65 €	0,00 €
Robotic tips/384	290,00 €	0	10	0	29,00 €
Total				55,05 €	151,45 €
Total/cell				**0,57 €**	**0,39 €**

Table S3 (related to Figure 6): Detailed overview of costs for mcSCRB-seq.

Supplementary Table 4

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

Table S4 (related to Figure 6): Detailed overview of hands-on and hands-off time necessary to create a sequenceable mcSCRB-seq library from one single cell plate.

Applying Single-Cell RNA Sequencing

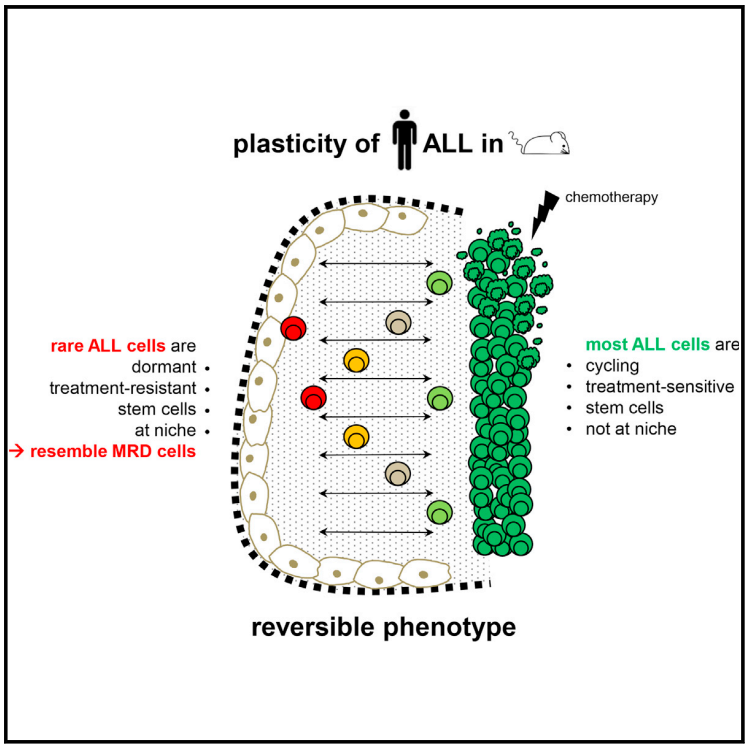
Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia

Article

Cancer Cell

Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia

Graphical Abstract



Authors

Sarah Ebinger, Erbey Ziya Özdemir, Christoph Ziegenhain, ..., Wolfgang Enard, Olivier Gires, Irmela Jeremias

Correspondence

irmela.jeremias@helmholtz-muenchen.de

In Brief

Ebinger et al. identify a rare subpopulation of acute lymphoblastic leukemia (ALL) cells that have the combined properties of long-term dormancy, treatment resistance, and leukemia initiation. RNA sequencing results show that these cells are similar to ALL cells isolated from patients at minimal residual disease.

Highlights

- Patients' ALL cells growing in mice contain a rare unfavorable subpopulation
- Unfavorable cells display treatment resistance, dormancy, and stemness
- Unfavorable cells mimic patients' primary cells at minimal residual disease
- Retrieving unfavorable cells from their environment sensitizes them for treatment

Accession Numbers

GSE83142

Ebinger et al., 2016, Cancer Cell *30*, 849–862
December 12, 2016
http://dx.doi.org/10.1016/j.ccell.2016.11.002

CellPress

Cancer Cell
Article

Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia

Sarah Ebinger,[1,15] Erbey Ziya Özdemir,[1,15] Christoph Ziegenhain,[2,15] Sebastian Tiedt,[1,15] Catarina Castro Alves,[1,15] Michaela Grunert,[1] Michael Dworzak,[3] Christoph Lutz,[4] Virginia A. Turati,[5] Tariq Enver,[5] Hans-Peter Horny,[6] Karl Sotlar,[6] Swati Parekh,[2] Karsten Spiekermann,[7,8] Wolfgang Hiddemann,[7,8] Aloys Schepers,[1] Bernhard Polzer,[9] Stefan Kirsch,[9] Martin Hoffmann,[9] Bettina Knapp,[10] Jan Hasenauer,[10,11] Heike Pfeifer,[12] Renate Panzer-Grümayer,[3] Wolfgang Enard,[2] Olivier Gires,[13] and Irmela Jeremias[1,8,14,16,*]

[1]Department of Gene Vectors, Helmholtz Zentrum München, German Center for Environmental Health (HMGU), 81377 Munich, Germany
[2]Anthropology and Human Genomics, Department Biology II, Faculty of Biology, Ludwig-Maximilians-Universität München, 82152 Martinsried, Germany
[3]Children's Cancer Research Institute and St. Anna Kinderspital, Department of Pediatrics, Medical University of Vienna, 1090 Vienna, Austria
[4]Department of Medicine V, University of Heidelberg, 69120 Heidelberg, Germany
[5]University College London Cancer Institute, London WC1E, UK
[6]Institute of Pathology, Ludwig-Maximilians-Universität München, 80337 Munich, Germany
[7]Department of Internal Medicine III, University Hospital Grosshadern, Ludwig-Maximilians-Universität München, 81377 Munich, Germany
[8]German Consortium for Translational Cancer Research (DKTK), Partnering Site, Munich, 81377 Munich, Germany
[9]Project Group Personalized Tumor Therapy, Fraunhofer Institute for Toxicology and Experimental Medicine ITEM, 93053 Regensburg, Germany
[10]Institute of Computational Biology, Helmholtz Zentrum München, German Center for Environmental Health (HMGU), 85764 Neuherberg, Germany
[11]Department of Mathematics, Technische Universität München (TUM), 85748 Munich, Germany
[12]Department of Medicine, Hematology and Oncology, Goethe University, 60590 Frankfurt, Germany
[13]Department of Otorhinolaryngology, Head and Neck Surgery, Grosshadern Medical Center, Ludwig-Maximilians-Universität München, 81377 Munich, Germany
[14]Department of Pediatrics, Dr. von Hauner Children's Hospital, Ludwig Maximilians University München, 80337 Munich, Germany
[15]Co-first author
[16]Lead Contact
*Correspondence: irmela.jeremias@helmholtz-muenchen.de
http://dx.doi.org/10.1016/j.ccell.2016.11.002

SUMMARY

Tumor relapse is associated with dismal prognosis, but responsible biological principles remain incompletely understood. To isolate and characterize relapse-inducing cells, we used genetic engineering and proliferation-sensitive dyes in patient-derived xenografts of acute lymphoblastic leukemia (ALL). We identified a rare subpopulation that resembled relapse-inducing cells with combined properties of long-term dormancy, treatment resistance, and stemness. Single-cell and bulk expression profiling revealed their similarity to primary ALL cells isolated from pediatric and adult patients at minimal residual disease (MRD). Therapeutically adverse characteristics were reversible, as resistant, dormant cells became sensitive to treatment and started proliferating when dissociated from the in vivo environment. Our data suggest that ALL patients might profit from therapeutic strategies that release MRD cells from the niche.

Significance

After initially successful chemotherapy, relapse frequently jeopardizes the outcome of cancer patients. To improve the prognosis of ALL patients, treatment strategies that eliminate tumor cells at minimal residual disease (MRD) and prevent relapse are required. Toward a better understanding of the underlying biology, we established preclinical mouse models mimicking MRD and relapse in patients. Primary and surrogate MRD cells shared major similarities in expression profiles, demonstrating the suitability of our model. MRD cells revealed major functional plasticity in vivo and treatment resistance was reversible; MRD cells became sensitive toward treatment once released from their in vivo environment. Effective therapeutic strategies might aim at dissociating persistent cells from their protective niche to prevent relapse in ALL patients.

INTRODUCTION

Relapse represents a major threat for patients with cancer. After initially successful treatment, rare tumor cells might survive and re-initiate the malignant disease with dismal outcome. Acute lymphoblastic leukemia (ALL) is associated with poor prognosis in infants and adult patients and is the most frequent malignancy in children (Inaba et al., 2013). In many patients, the majority of ALL cells respond to chemotherapy but a minority display resistance, survive therapy, and cause relapse with poor outcome (Gokbuget et al., 2012).

Despite its clinical importance, basic biologic conditions underlying relapse remain partially elusive. For example, it is unclear whether relapse-inducing cells exist before onset of treatment or develop as result of therapy, and whether permanent or reversible characteristics determine relapse-inducing cells (Kunz et al., 2015). Of translational importance, understanding basic mechanisms opens perspectives for effective therapies to eradicate relapse-inducing cells.

Relapse-inducing cells, by their clinical definition, self-renew and give rise to entire tumors indicating tumor-initiating potential, a typical characteristic of cancer stem cells (Essers and Trumpp, 2010). In numerous tumor entities including acute myeloid leukemia, cancer stem cells were identified as a biologically distinct subpopulation that displays specific surface markers, has leukemia-inducing potential in mice, and gives rise to a hierarchy of descendant cells that lack such properties (Bonnet and Dick, 1997; Visvader and Lindeman, 2008). In ALL, however, many different subpopulations display stem cell properties; neither a stem cell hierarchy nor phenotypic markers defining stem cells could be identified (Kong et al., 2008; le Viseur et al., 2008; Rehe et al., 2013). Thus, up to now, stemness represents an insufficient criterion to define the subpopulation of relapse-inducing cells in ALL.

An additional feature of relapse-inducing cells is their treatment resistance, as, again by definition, they survive chemotherapy and eventually give rise to relapse with decreased chemosensitivity. Resistance against chemotherapy is closely related to dormancy as chemotherapy mainly targets proliferation-associated processes that are inactive in dormant cells (Clevers, 2011; Zhou et al., 2009). Dormant cells, by definition, do not divide or divide very slowly over prolonged periods of time, might survive chemotherapy, persist in minimal residual disease (MRD), and give rise to relapse (Schillert et al., 2013; Schrappe, 2014). Indeed, an increased frequency of non-dividing tumor cells has been described in patients after chemotherapy for defined subtypes of ALL (Lutz et al., 2013).

So far, technical obstacles have hampered characterizing phenotypic and functional features of relapse-inducing cells in ALL in detail. Established ALL cell lines represent inappropriate models as they display continuous proliferation. In patients, relapse-inducing cells are very rare and defining cell surface markers that reliably identify these rare ALL cells from the multiplicity of normal bone marrow cells remains intricate, at least in certain ALL subtypes (Hong et al., 2008; Ravandi et al., 2016). Moreover, primary ALL cells do not grow ex vivo, disabling their amplification in culture.

An attractive possibility to experimentally study patients' tumor cells in vivo is the patient-derived xenograft (PDX) model, which uses immuno-compromised mice to expand tumor cells from patients (Kamel-Reid et al., 1989). As shown previously, PDX ALL cells retain important characteristics of primary ALL cells (Castro Alves et al., 2012; Schmitz et al., 2011; Terziyska et al., 2012). While PDX models are mostly used for preclinical treatment trials (Gao et al., 2015; Townsend et al., 2016), we used them here to study relapse-inducing cells in ALL.

RESULTS

To characterize the challenging subpopulation of relapse-inducing cells in ALL, we used the individualized xenograft mouse model as a preclinical model, molecular cell marking as an unbiased approach, and in vivo dormancy as a functional benchmark. To mimic the heterogeneity of ALL, samples from nine different ALL patients were studied including children and adults, B cell precursor-ALL and T-ALL, first diagnosis, and relapse (Table S1).

Molecular Marking Allows Unbiased, Sensitive Isolation of Rare PDX ALL Cells

To study ALL growth starting very early after disease onset in the PDX mouse model, the technical challenge consisted in reliably enriching very low numbers of human ALL cells from mouse bone marrow. As expression levels of endogenous surface antigens across potentially relevant, but yet undefined, subpopulations are unknown, we used lentiviral transduction for unbiased molecular marking and in vivo imaging (Figure 1A).

PDX ALL cells were lentivirally transduced to express a luciferase for in vivo imaging (Terziyska et al., 2012), an artificial antigen (truncated nerve growth factor receptor [NGFR]) for magneto-activated cell sorting (Fehse et al., 1997) and a red fluorochrome for cell sorting by flow cytometry (Figures S1A and S1B). Transgenes allowed effective and reliable enrichment of minute numbers of PDX cells from mouse bone marrow in this two-step procedure. Quantification of PDX cells isolated with the magnetic-activated cell sorting (MACS)/fluorescence-activated cell sorting approach closely correlated with other methods monitoring leukemic proliferation, such as in vivo imaging and flow cytometry-based quantification of leukemia cells (Figure S1C). Quality controls showed that the procedure was highly efficient and reliable with minor cell loss (Table S2).

The procedure enabled addressing basic questions with translational potential in ALL biology. Homing capacity of PDX cells to mouse bone marrow differed by more than two orders of magnitude between the nine samples studied (Figure 1B). Homing efficiency decreased significantly when smaller cell numbers were injected (Figure S1D). These data argue in favor of sample-specific characteristics determining homing, and against the presence of a preformed, fixed number of leukemia homing sites within the niche. Spontaneous growth of PDX ALL cells in mouse bone marrow was logarithmic over the first 2 weeks of in vivo growth (Figures 1C and S1C). Growth slowed down thereafter and as early as at 10% blasts in bone marrow, when space restriction appears unlikely to be causative. Model selection indicated overall logistic growth which is typical for insufficient nutrient supply (Figure S1E). Thus, PDX ALL cells show sample-specific homing followed by logistic growth in mouse bone marrow.

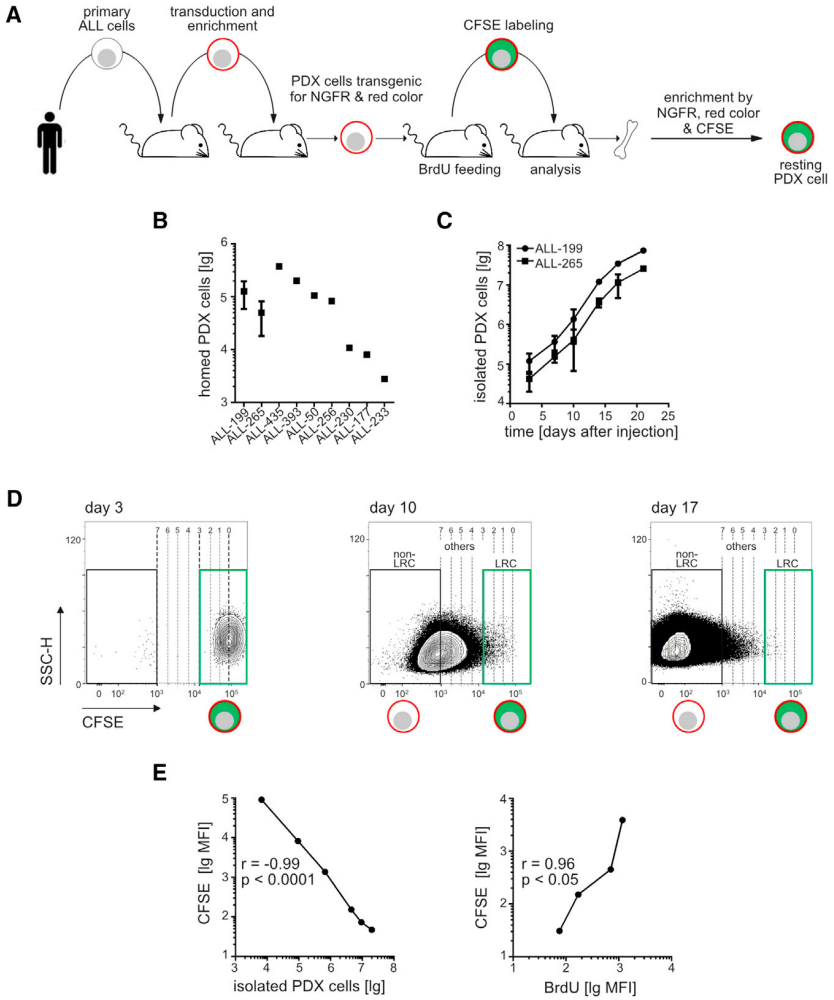


Figure 1. CFSE Staining Allows Reliable Monitoring of PDX ALL Growth in Mice

(A) Experimental procedure of generating PDX ALL cells expressing several transgenes, staining with CFSE, and enriching rare transgenic, CFSE-stained PDX cells from mouse bone marrow.

(B) Of each PDX sample, 10^7 triple transgenic PDX cells were injected intravenously into mice and re-isolated from the bone marrow 3 days later; each dot represents data from one mouse, except that a mean of eight mice plus SE is shown for samples ALL-199 and ALL-265.

(C) 10^7 CFSE-stained PDX cells/mouse were injected and PDX cells were quantified in up to 11 mice per time point; shown is mean and SE.

(D) Gating strategy defining LRC, non-LRC, and others. MFI of CFSE at the start of the experiment (3 days after cell injection) was divided by factor 2 to model bisections; upon no more than three bisections, cells were considered as LRC, upon more than seven bisections as non-LRC; intermediate cells were considered as others.

(E) Similar experiment as in (C), except that the donor mouse was fed with BrdU in the last 7 days before cell harvesting. Each dot represents data from one mouse.

See also Figure S1, Tables S1, and Table S2.

CFSE Staining Allows Reliable Monitoring of PDX ALL Growth in Mice

Proliferation-dependent dyes such as bromodeoxyuridine (BrdU) and carboxyfluorescein diacetate succinimidyl ester (CFSE) remain stable in mice over several months, enabling the characterization of a heterogeneous growth pattern in normal hematopoiesis (Takizawa et al., 2011). We adapted the use of these dyes in PDX tumor models. As BrdU staining requires the permeabilization and destroying of cells, fluorescent CFSE was mainly used as it allows flow cytometric enrichment of living cells for functional experiments including re-transplantation. Loss of CFSE was used to distinguish subpopulations of slowly and rapidly growing cells (Figures 1D and S1F) that were called label-retaining cells (LRC) and non-label-retaining cells (non-LRC), respectively (Takizawa et al., 2011). LRC were defined as those cells that had undergone at most three CFSE bisections resembling cell divisions (see the Supplemental Experimental Procedures for details). Loss of CFSE tightly correlated with increase in PDX cell numbers and loss of BrdU (Figures 1E and S1G) and confirmed that PDX ALL cells grow in vivo, but not ex vivo (Figure S1H). Thus, CFSE staining represents a reliable approach to monitor proliferation of PDX ALL cells in mice.

A Rare, Long-Term Dormant Subpopulation Exists in ALL PDX Cells

Importantly, CFSE staining disclosed the existence of a rare fraction of PDX ALL cells that hardly divided over prolonged periods of time (Figure 2A). LRC, by definition, had undergone no more than three cell divisions within 21 days, during which the leukemia burden had risen by several orders of magnitude so that mice would succumb to leukemia within a few days. In all nine PDX ALL samples studied, LRC were identified after prolonged periods of leukemic growth; (Figures 2B and S2A).

Thus, similarly to normal hematopoiesis (Trumpp et al., 2010), PDX ALL contains a rare subpopulation of LRC. LRC might resemble the dormant tumor cells described in ALL patients (Figure S2B) (Lutz et al., 2013). As an advantage over work with primary cells, our preclinical approach allows repetitive work on pure, vivid LRC, which gave us the chance to functionally and phenotypically characterize this interesting population.

LRC Localize to the Endosteum, but Are Not Enriched for Stem Cells

Both normal hematopoietic stem cells and leukemia stem cells were reported to preferentially localize close to the endosteum, where a supportive niche might exist (Morrison and Spradling, 2008). We also found that LRCs preferentially localized close to the endosteum (Figures 3A–3C and S3), suggesting that they might use the same niche as normal hematopoietic stem cells and cancer stem cells.

We therefore asked whether LRC might resemble cancer stem cells. To compare leukemia-initiating potential between LRC and

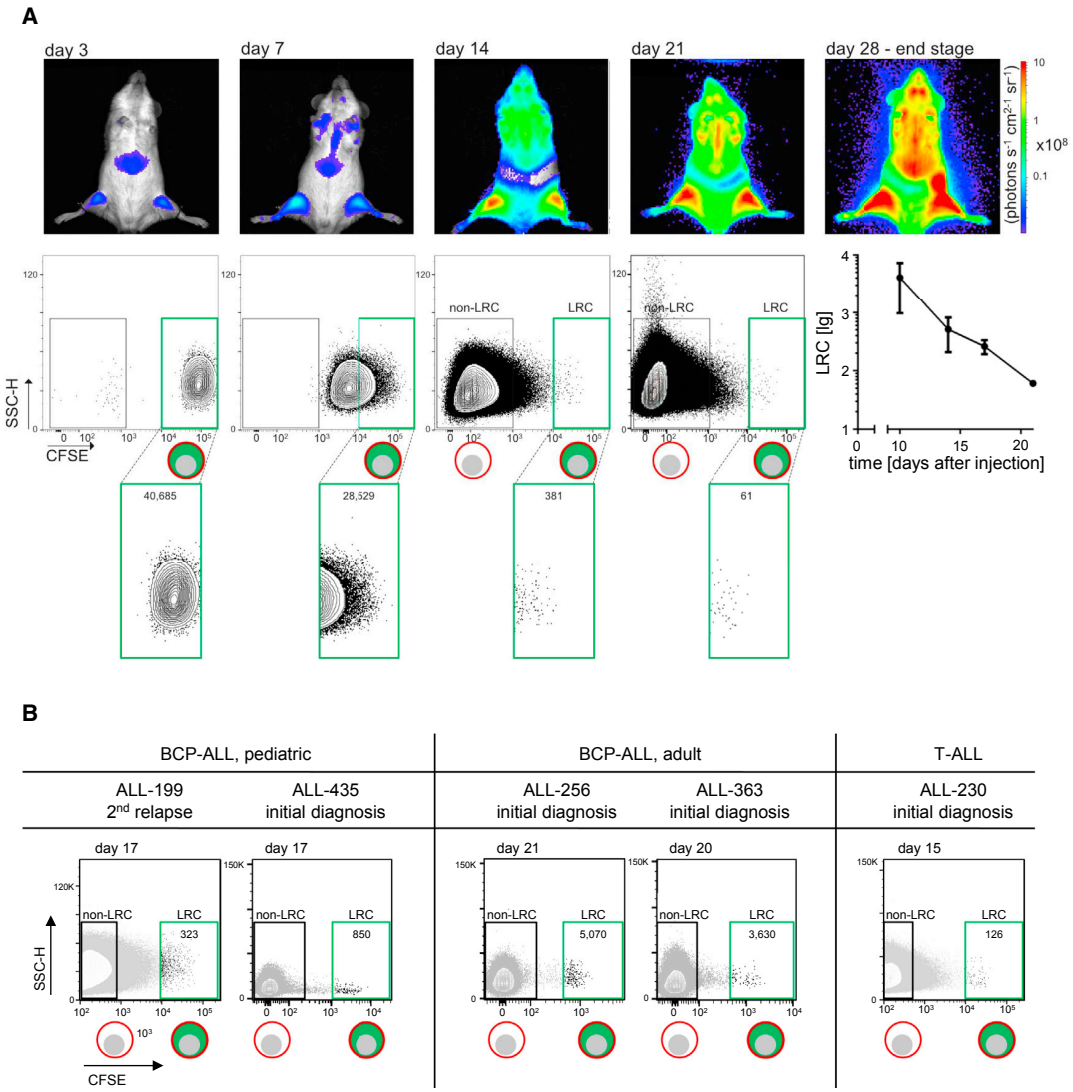


Figure 2. A Rare, Long-Term Dormant Subpopulation Exists in ALL PDX Cells

(A) 10^7 CFSE-stained PDX ALL-265 cells were injected into each of six mice; bioluminescence in vivo imaging was performed prior to quantifying LRC in one mouse per time point; LRC numbers are indicated and summarized in the line graph as a mean of up to ten mice ± SE.

(B) Identification of LRC in PDX cells from all different ALL patients. Experiments were performed as in (A).

See also Figure S2.

non-LRC, we performed limiting dilution transplantation assays and monitored engraftment by bioluminescence in a total of 83 mice (Table S3). To our surprise, we found highly similar stem cell frequencies in LRC and non-LRC and similar engraftment rates after transplantation of, e.g., ten cells per mouse (Figure 3D). The 95% confidence interval of the estimated frequency of leukemia-inducing cells ranged between 1/19 and 1/84 cells for LRC and between 1/40 and 1/179 cells in non-LRC of ALL-265 (Table S3). Similar findings were obtained for ALL-199 (Table S3). Thus, although only LRC display typical characteristics of stem cells such as reduced proliferation rate and localization close to the endosteum, LRC and non-LRC exhibited similar leukemia-initiating potential.

LRC Survive Systemic Drug Treatment In Vivo

Dormant cells are known for their resistance against drug treatment, complicating elimination by anti-cancer therapy (Essers and Trumpp, 2010). We compared in vivo drug response of LRC and non-LRC by transplanting CFSE-labeled PDX ALL cells, treating mice with systemic chemotherapy on day 7 and analyzing surviving LRC and non-LRC on day 10 (Figure 4A). Chemotherapy reduced the overall leukemic burden by over 90% (Figures 4B and S4A) and eradicated most non-LRC. As a prominent difference, most LRC survived chemotherapy so that LRC increased in relative proportions (Figures 4C–4E and S4B–S4D). A 10- to 100-fold less efficient elimination of LRC compared with non-LRC became obvious across all PDX ALL

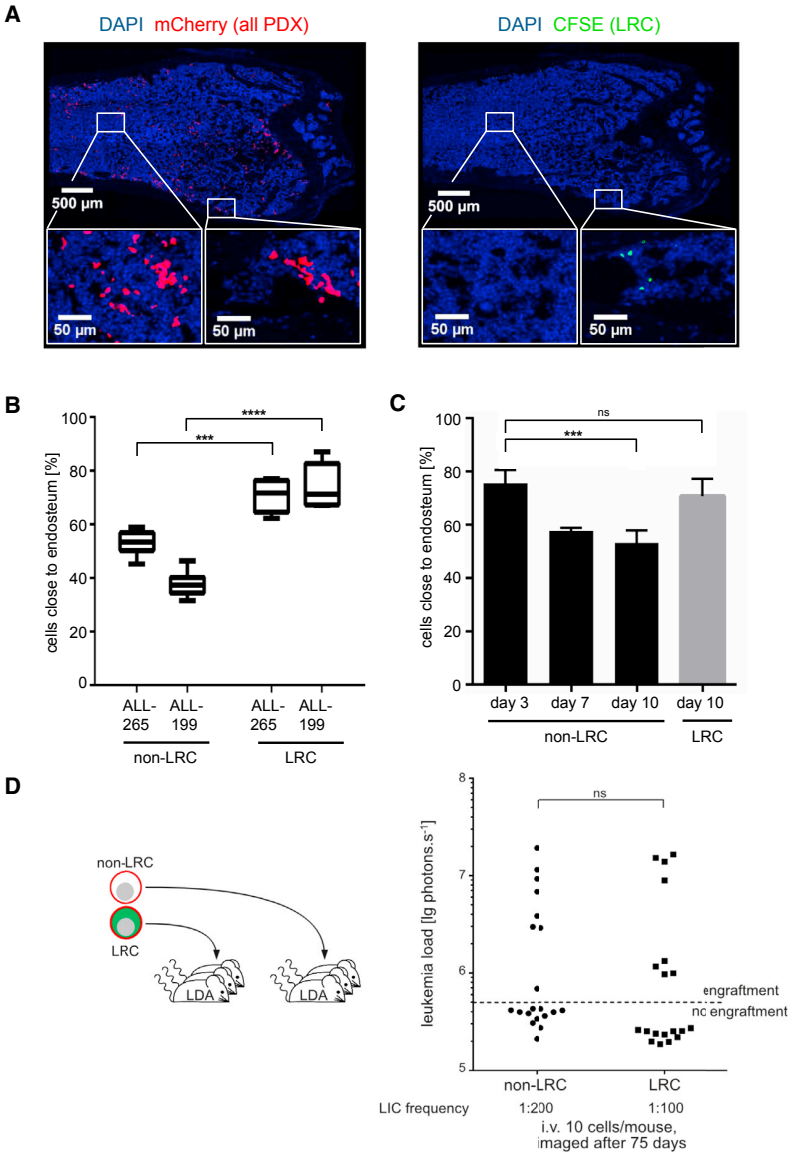


Figure 3. LRC Localize to the Endosteum, but Are Not Enriched for Stem Cells

(A) Immunohistochemistry of consecutive mouse bone marrow femur sections 10 days after injection of CFSE-stained PDX ALL-265 cells; mCherry (red; left panel) indicates all PDX cells, CFSE (green; right panel) indicates LRC.

(B) All sections from day 10 were quantified defining the endosteal region as less than 100 µm from bone matrix; shown is the median with upper/lower quartile and maximum/minimum of two to three sections from two femurs in two mice per data point; ***$p < 0.001$, ****$p < 0.0001$ by two-tailed unpaired t test.

(C) Kinetic for ALL-265 as mean ± SE; ***$p < 0.01$ by two-tailed unpaired t test.

(D) Ten LRC or non-LRC were injected into each of 39 mice and engraftment was determined by in vivo imaging at day 75; each dot represents one mouse; dashed line represents detection threshold (5×10^5 photons s^{-1}); ns: not significant as determined by two-tailed unpaired t test.

See also Figure S3 and Table S3.

samples tested that were derived from either primary disease or relapse, suggesting that this phenomenon is not restricted to a certain disease stage. Treatment-surviving LRC harbored leukemia-initiating potential as they gave rise to leukemias upon retransplantation at a kinetic similar to that of untreated LRC (Figures 4F and S4E).

Taken together, LRC share the most important functional features that impede the cure of cancer: (1) dormancy, (2) in vivo drug resistance, and (3) leukemia-initiating potential. LRC might thus serve as preclinical surrogate for relapse-inducing cells in ALL.

Expression Profile of LRC Shows Distinct Changes to Non-LRC

We then evaluated whether LRC adequately resemble challenging cells in patients. For a broad, unbiased comparison between LRC and non-LRC, RNA sequencing (RNA-seq) was performed on single cells and bulk populations (Figure 5A). Data from single cells correlated with data from bulk populations and different ALL PDX samples showed similar expression profiles (Figures S5A and S5B). Preliminary expression arrays on pools of 40 LRC and non-LRC showed mainly similar results (data not shown).

Single LRC differed consistently from single non-LRC as revealed by clustering differently expressed genes (Figures 5B and Table S4) and by a principle component analysis of the most variable genes (Figure 5C). Single LRC also had an overall reduced RNA content (Figure S5C), indicating a less active metabolism that is a prerequisite of dormant cells. We combined single-cell and bulk data of all six sample pairs to identify differently expressed genes (Table S5). Enrichment analysis revealed that genes expressed less in LRC were most strongly enriched in cell cycle and DNA replication and that genes more expressed in LRC were most strongly enriched in cell adhesion (Figures 5D, S5D, and Table S6). Hence, expression profiling of single cells and in bulk confirmed the quiescent state of LRC and an LRC signature of at least 2-fold differently expressed genes ranked by their significance (Figures 5E and Table S5) was used for further comparisons.

LRC Resemble MRD Cells in the PDX Mouse Model

Relapse often results from treatment-resistant tumor cells that survive chemotherapy and persist at MRD. MRD cells contain a major fraction of dormant tumor cells (Lutz et al., 2013). Here, we hypothesized that LRC might represent surrogates for MRD cells.

To experimentally test this hypothesis, we established a preclinical model of MRD for ALL-265 and ALL-199. When untreated

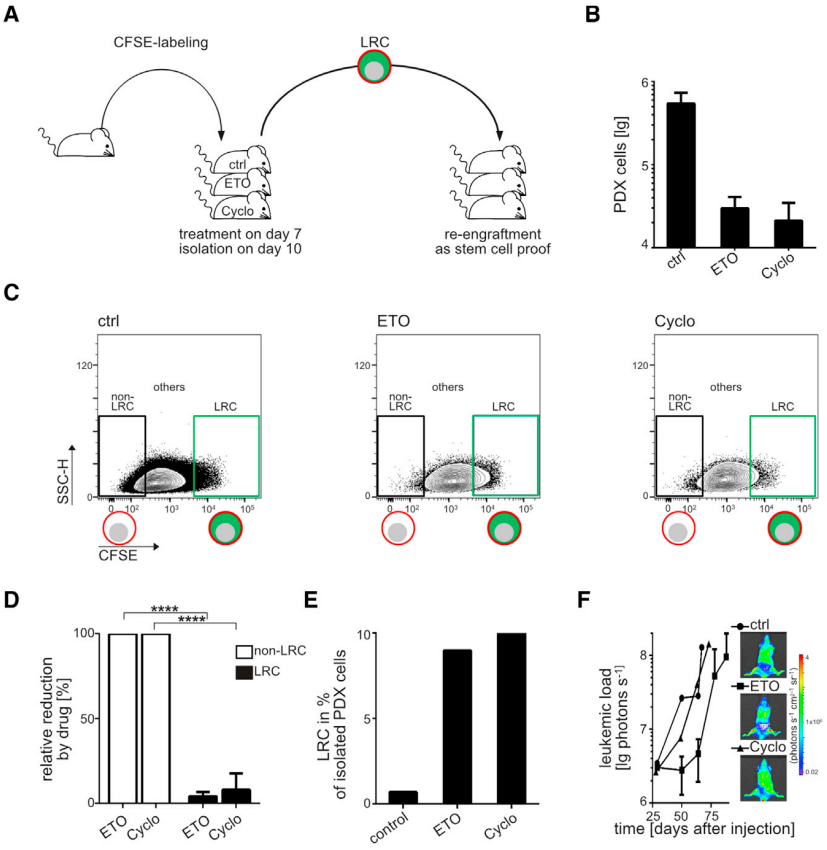


Figure 4. LRC Survive Systemic Drug Treatment In Vivo

(A) Each mouse was injected with 10^7 CFSE-stained ALL-265 PDX cells and treated with buffer, etoposide (ETO, 50 mg/kg, intraperitoneally [i.p.]), or cyclophosphamide (Cyclo, 150 mg/kg, i.p.) on day 7. Mice were euthanized on day 10; LRC were analyzed and re-transplanted into secondary recipients.

(B) Living PDX cells from mice in (A) were quantified and presented as mean of each group (n = 4–5) ± SE.

(C) Original data for one representative mouse per treatment.

(D) Mean of all four to five mice per treatment, depicted as relative drug effect on LRC compared with non-LRC (100%) ± SE; ****p < 0.0001 by two-tailed unpaired t test.

(E) Mean relative proportion of LRC of total PDX cells.

(F) LRC isolated were re-transplanted and mice monitored by in vivo imaging; mean of each group (n = 1–2) ± SE.

See also Figure S4.

control samples were harvested at advanced leukemia, they contained a leukemic burden of ~30% human blasts in mouse bone marrow, mimicking the situation at diagnosis. Remaining mice received a systemic treatment with conventional chemotherapeutic drugs over 2–3 weeks (Figure 6A), which needs careful dosing as supportive therapy is mainly unfeasible in mice. A combination treatment of vincristine and cyclophosphamide reduced tumor burden substantially according to in vivo imaging (Figures 6B, 6C, and S6A). Postmortem analysis revealed that chemotherapy had reduced leukemic burden by more than two orders of magnitude to ~0.1% leukemia cells in bone marrow. This resembled not only complete morphologic, but also complete molecular remission criteria (Figures 6D and S6B). MRD cells revealed relapse-inducing potential as they re-grew in mice when treatment was stopped (Figure S6C).

MRD cells were isolated from mouse bone marrow using expressed transgenes as above, and RNA sequencing of single cells and bulk samples was performed. Resulting transcriptomes showed marked differences between MRD and untreated control cells (Figure S6D). Enrichment analysis revealed significantly reduced expression of MYC and E2F target genes in MRD compared with untreated cells. Genes expressed less in MRD cells were most strongly enriched in cell cycle and DNA replication, while genes expressed more in MRD cells were most strongly enriched in cell adhesion (Figures 6E and S6E). This suggests a dormant phenotype of MRD cells similar to the dormant phenotype seen in LRC (Figure 5D). KEGG pathway analysis highlighted that MRD cells were of dormant nature and expressed increased adhesion molecules (Figure S6E). Indeed, single MRD cells clustered together with single LRC in a principal component analysis separated from non-LRC and cells from untreated mice (Figure S6F). Accordingly, the LRC signature (Figure 5E and Table S5) was strongly enriched in MRD cells and genes in MRD and LRC cells were similarly regulated compared with their respective controls (Figure 6F). This suggests that LRC mimic MRD cells in our preclinical mouse model.

LRC Resemble Primary MRD Cells from Patients

To relate these findings to the clinical situation, expression profiles from primary tumor cells from five children and two adults with B cell precursor (BCP) ALL were profiled at diagnosis and at MRD (Figure 7A and Table S7). Children and adults were treated according to the BFM-2009 and GMALL-0703 protocols, respectively, and MRD cells were enriched by flow cytometry at days 33 and 71 of treatment, respectively. In adults, we chose BCR-ABL-positive ALL and enriched the subpopulation of StemB cells at MRD, as Lutz et al. (2013) had shown that these cells exhibit a dormant phenotype. As dormancy in StemB cells might have persisted for a long period during treatment in patients, LRC might especially resemble StemB cells at MRD. We could obtain single-cell transcriptomes from one patient and one bulk transcriptome from another patient. K-means clustering and principal component analysis revealed that single StemB cells clustered together with single LRC and MRD cells, while single non-LRC clustered together with single untreated control cells (Figures 7B and 7C). The bulk StemB sample was distinct from diagnostic tumor cells of untreated adult patients with BCR-ABL-positive ALL (Figure S7A). Although limited by small cell and sample numbers, the data indicate that LRC resemble

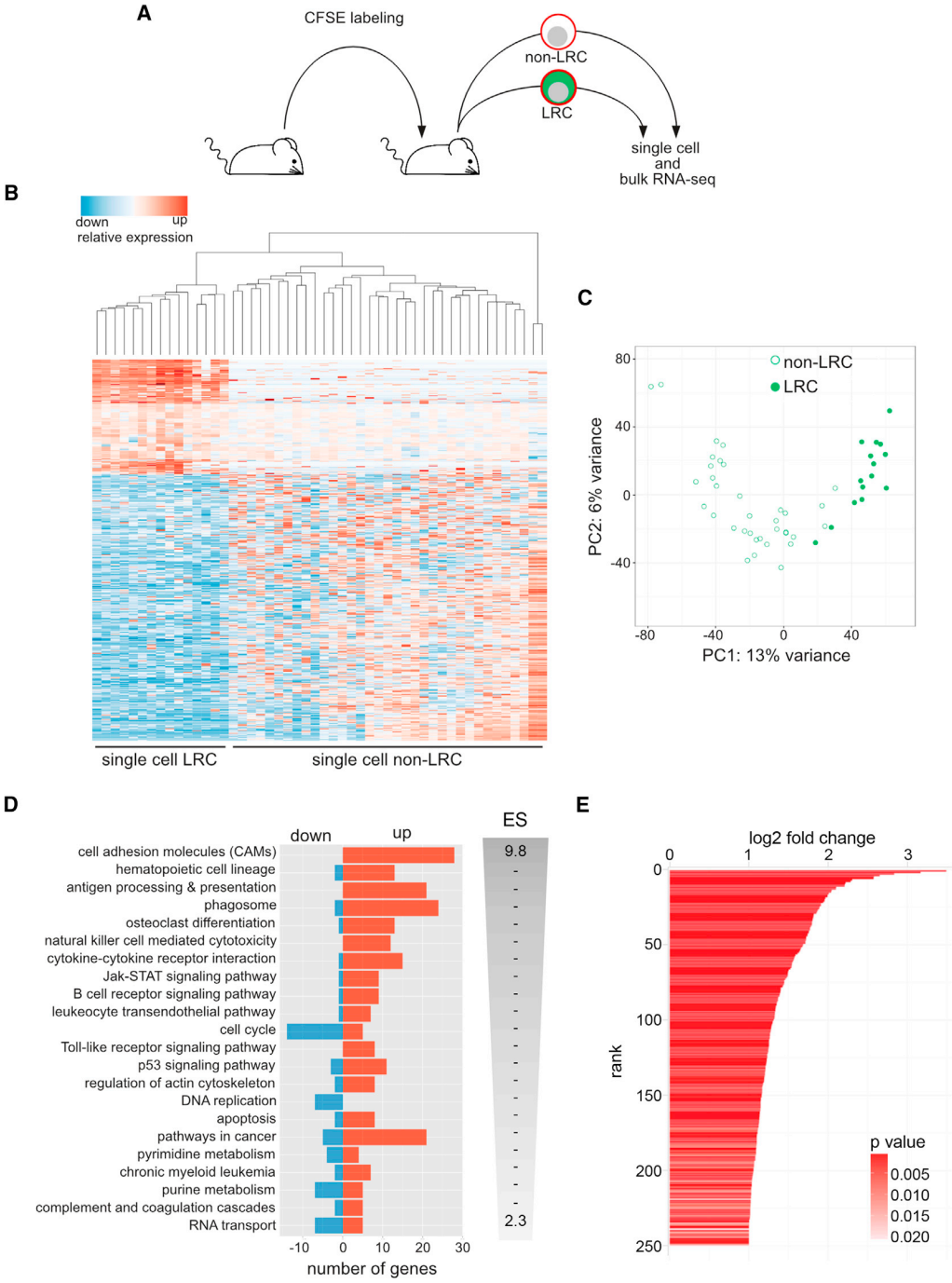


Figure 5. Expression Profile of LRC Shows Distinct Changes to Non-LRC

(A) Fifteen days after transplantation, ALL-265 LRC or non-LRC were isolated and single-cell mRNA-seq was performed in 15 LRC and 35 non-LRC.

(B) Hierarchical clustering and gene expression heatmap across the 500 most differentially expressed genes (false discovery rate [FDR] <0.01) in 15 LRC and 35 non-LRC single cells. Values are plotted relative to the average of non-LRC.

(C) Principal component analysis of the 500 most variable genes in all 50 single cells.

(D) Significantly enriched KEGG pathways (FDR <0.05) as determined by fixed network enrichment analysis (FNEA); bars show the number of significantly up- or downregulated genes in the corresponding pathway and are ordered according to the enrichment score (ES).

(E) LRC signature genes (FDR < 0.05 and log2 fold-change >1) were derived from integrated bulk and single-cell RNA-seq analysis from six animals carrying either ALL-265 or ALL-199 and are shown ranked by fold-change and colored by significance.

See also Figure S5, Tables S4, S5, and S6.

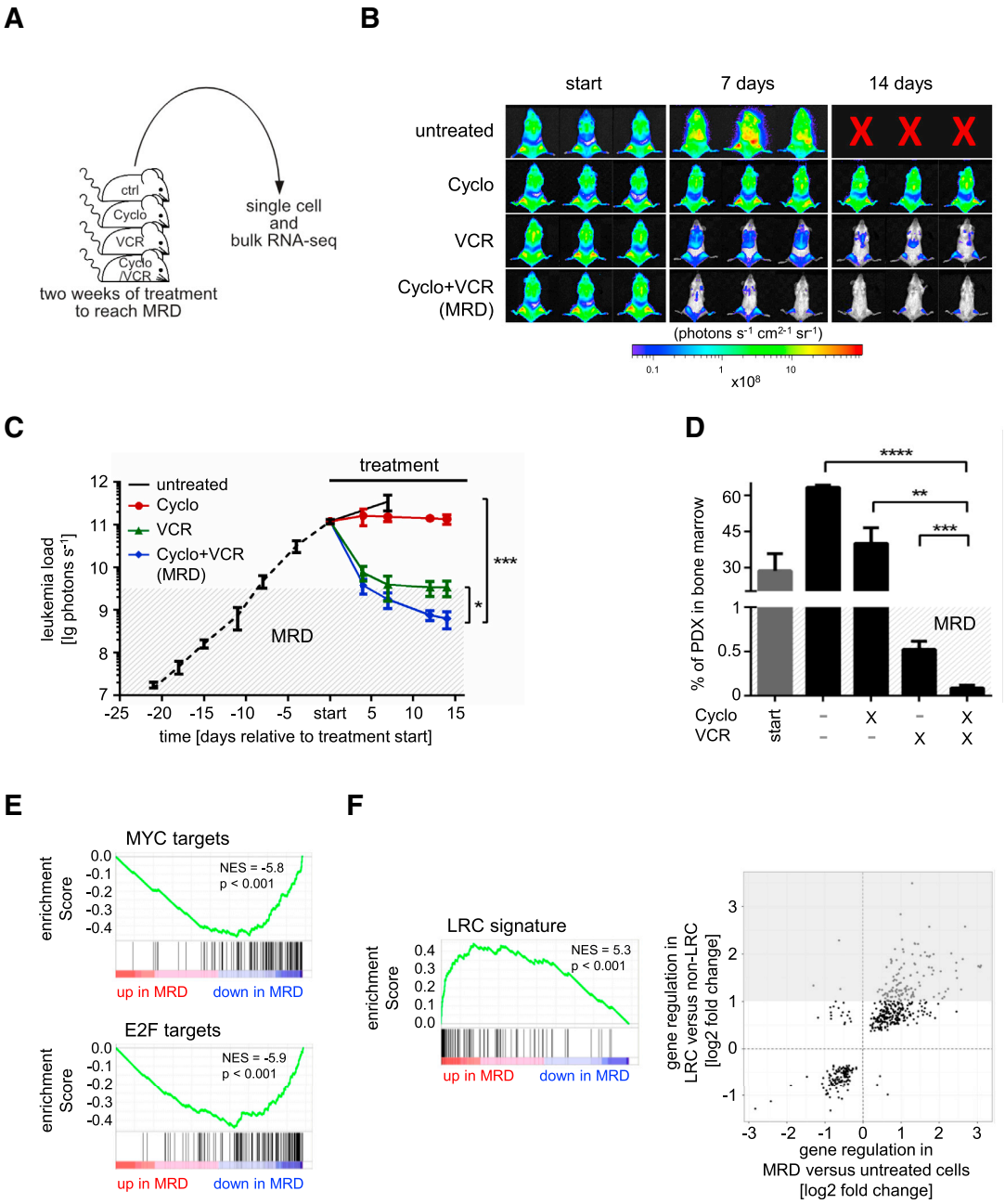


Figure 6. LRC Resemble MRD Cells in the PDX Mouse Model

(A) 10^7 ALL-199 cells were injected into 19 mice; when 30% of bone marrow cells were human, PDX cells were enriched from five mice and used as untreated control samples; cells of one mouse were subjected to single-cell sequencing; the remaining mice received buffer, vincristine (VCR, 0.25 mg/kg; n = 5), cyclophosphamide (Cyclo, 100 mg/kg; n = 3), or a combination thereof (VCR + Cyclo; n = 6) weekly for 2 weeks; when VCR + Cyclo combination treatment had reduced tumor burden to MRD (<1% human cells in bone marrow), PDX cells were enriched and cells of one VCR + Cyclo mouse were subjected to single cell mRNA-seq.

(B) In vivo imaging data of three representative mice per group.

(C) Mean of each group ± SE; *p < 0.05, ***p < 0.001 by two-tailed unpaired t test; mice receiving buffer had to be euthanized after 1 week of treatment due to end-stage leukemia.

(D) Percentage of PDX ALL cells in mouse bone marrow as determined by flow cytometry postmortem as mean ± SE; **p < 0.01, ***p < 0.001, ****p < 0.0001 by two-tailed unpaired t test.

(E) MRD cells show reduced expression of MYC- and E2F-target genes in gene set enrichment analysis (GSEA) (Liberzon et al., 2015).

(legend continued on next page)

the dormant subpopulation of StemB cells in adult ALL patients at MRD.

This is also supported when comparing the LRC profiles with further published transcriptomes. Genes differently expressed in CD34-positive chronic myeloid leukemia cells (Graham et al., 2007), in leukemia stem cells (Saito et al., 2010), in hematopoietic stem cells (Eppert et al., 2011; Georgantas et al., 2004), as well as in pediatric ALL cells with high risk of relapse (Kang et al., 2010) were all significantly enriched in LRC versus non-LRC cells (Figures 7D, S7B, and S7C).

To further analyze the similarity of LRC to MRD cells from patients, we generated bulk transcriptomes of primary samples from five children with BCP-ALL before the onset of treatment and three matched MRD samples collected 33 days after the onset of treatment. Expression profiles differed significantly between diagnosis and MRD (Figure 7E and Table S8) and MRD cells regulated genes in the same direction as LRC compared with their respective controls, as revealed by a significant overlap of up- and downregulated genes (hypergeometric test, $p = 1.9 \times 10^{-23}$) and by a significant enrichment of the LRC signature ($p < 0.001$; Figure 7F). Finally, we combined these transcriptomes with all bulk samples isolated from the LRC and MRD mouse models and analyzed them unsupervised in a principal component analysis (Figure 7G). The first principal component separated all dormant and drug-resistant cells (PDX-LRC, PDX-MRD, and primary MRD) from all control cells (PDX-non-LRC, PDX untreated, and primary diagnosis).

In summary, we show that a distinct subpopulation of LRC exists in our ALL PDX model that combines the unfavorable characteristics of stemness, drug resistance, and dormancy. These LRC show high similarities to MRD cells in our mouse model and to MRD cells in ALL patients. Hence, LRC might represent preclinical surrogates for relapse-inducing cells in patients and could be used to develop therapeutic strategies to prevent relapse.

Release from the Environment Induces Proliferation in LRC

As the first step toward therapies, we studied whether unfavorable drug resistance and dormancy represented permanent or reversible features in LRC. Dormancy and drug resistance might exist as genuine, constant biological characteristics of a special ALL subpopulation or as reversible functional phenotypes of putatively every ALL cell depending on the context.

To address this question, LRC and non-LRC were dissociated from their environment, isolated, and re-transplanted into recipient mice (Figures 8A and S8A). When non-LRC were re-stained with CFSE and re-transplanted at high numbers, they gave rise to an identical LRC population as re-transplanted bulk cells (Figures 8B and S8A); transplantation of high cell numbers of LRC was impossible, as only low numbers of LRC can be recovered from mice. When low cell numbers were re-transplanted, LRC, non-LRC, and bulk cells initiated identical leukemic growth in mice as monitored by bioluminescence in vivo imaging (Figures 8C and S8A). These data indicate that dormancy represents a reversible feature of LRC, as LRC lose their dormant nature once they are retrieved from their specific environment and transferred into a different surrounding.

Release from the Environment Sensitizes LRC and MRD Cells for Drug Treatment

As dormancy emerged as a reversible phenotype, we asked whether drug resistance might be equally reversible. Isolated LRC and non-LRC or MRD and previously untreated cells from the PDX mouse model were treated ex vivo with common ALL chemotherapy drugs or drug controls. Here, the technical challenge lay in the very minor cell numbers of LRC and MRD that can be isolated from mice and used for ex vivo experiments (Figure S8B). Co-culture with feeder cells resembling bone marrow stroma reduced drug response in all samples, suggesting the influence of the bone marrow environment on drug resistance (Figures S8C–S8F) (Tesfai et al., 2012). Ex vivo, neither LRC nor MRD cells displayed increased drug resistance compared with their respective controls (Figures 8D and S8G).

Taken together, LRC and MRD cells showed a marked gain in drug sensitivity ex vivo compared with in vivo after isolation from the bone marrow environment. Both LRC and MRD cells lost their enhanced drug resistance, distinguishing them from non-LRC or untreated cells, once they were retrieved from their in vivo environment and cultured ex vivo (Figure 8E). Dormancy was reversible in LRC and drug resistance was reversible in both LRC and MRD cells. As LRC might represent surrogates for relapse-inducing cells in patients, our data suggest that the interaction between LRC and their environment represents an attractive therapeutic target for preventing relapse. Relapse-inducing cells might gain sensitivity toward treatment once mobilized from their in vivo environment.

DISCUSSION

The present work aimed at a better understanding of the cells that induce relapse in ALL and thereby limit prognosis of patients. We identified a rare, long-term dormant subpopulation termed LRC exhibiting the adverse characteristics of dormancy, in vivo drug resistance, and leukemia-initiating properties. LRC highly resemble primary MRD cells from adult and pediatric patients with ALL. MRD cells require preferential eradication by anti-leukemia treatment. LRC in preclinical models can now be used as surrogates for relapse-inducing cells in patients for developing therapies to prevent relapse. Upon removal from their in vivo environment, LRC lost dormancy and drug resistance, suggesting a reversible nature of adverse characteristics and an important role for the interaction between ALL and the environment. The data suggest that drug resistance and dormancy are linked and represent an acquired stem-like phenotype. Our data imply developing treatment approaches that dissociate ALL cells from their protective niche to sensitize them toward anti-leukemia treatment.

(F) GSEA was performed comparing LRC signature with transcriptomes of MRD versus untreated cells (mean of data for ALL-199; left panel). Scatterplot of fold-changes for genes differentially expressed (FDR < 0.05) between both LRC versus non-LRC and MRD versus untreated control cells; grey area indicates LRC signature (right panel).
See also Figure S6.

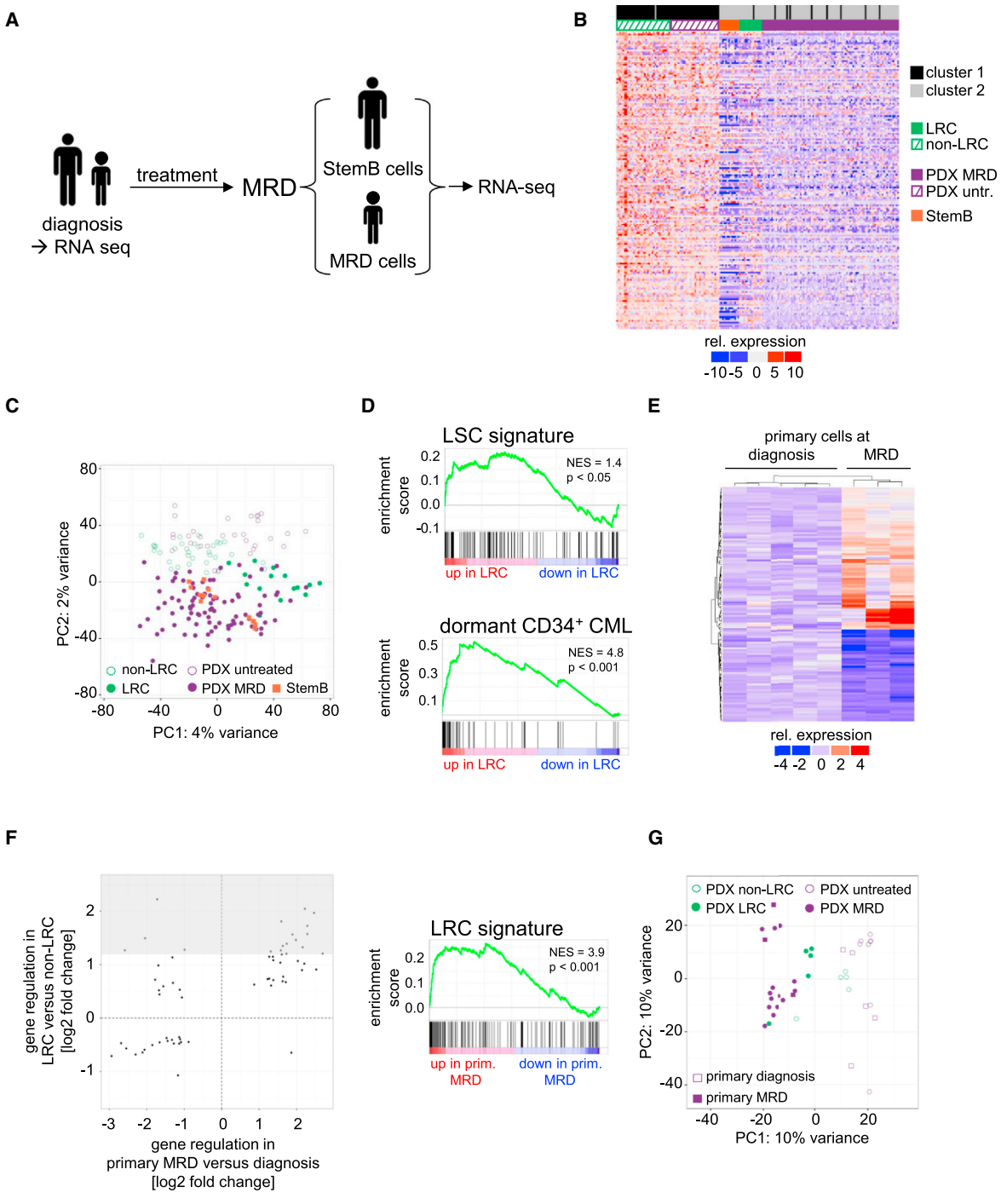


Figure 7. LRC Resemble Primary MRD Cells from Patients

(A) Adult or pediatric ALL patients were treated according to GMALL-0703 or BFM-2009 protocols for 71 or 33 days, respectively; at MRD, the subgroup of StemB cells (in samples from adults) or all remaining ALL cells (in samples from children) were enriched out of normal bone marrow; cells at diagnosis and at MRD were subjected to RNA-seq.

(B) K-means clustering of gene expression values of 167 highly differentially expressed genes (FDR < 0.001) of all data from single cells.

(C) Principal-component analysis (PCA) of single cell transcriptomes using all shared expressed genes; each symbol indicates a single cell.

(D) GSEA comparing the LRC signature with signatures of leukemia stem cells (Saito et al., 2010) and dormant CD34-positive chronic myeloid leukemia (CML) (Graham et al., 2007).

(legend continued on next page)

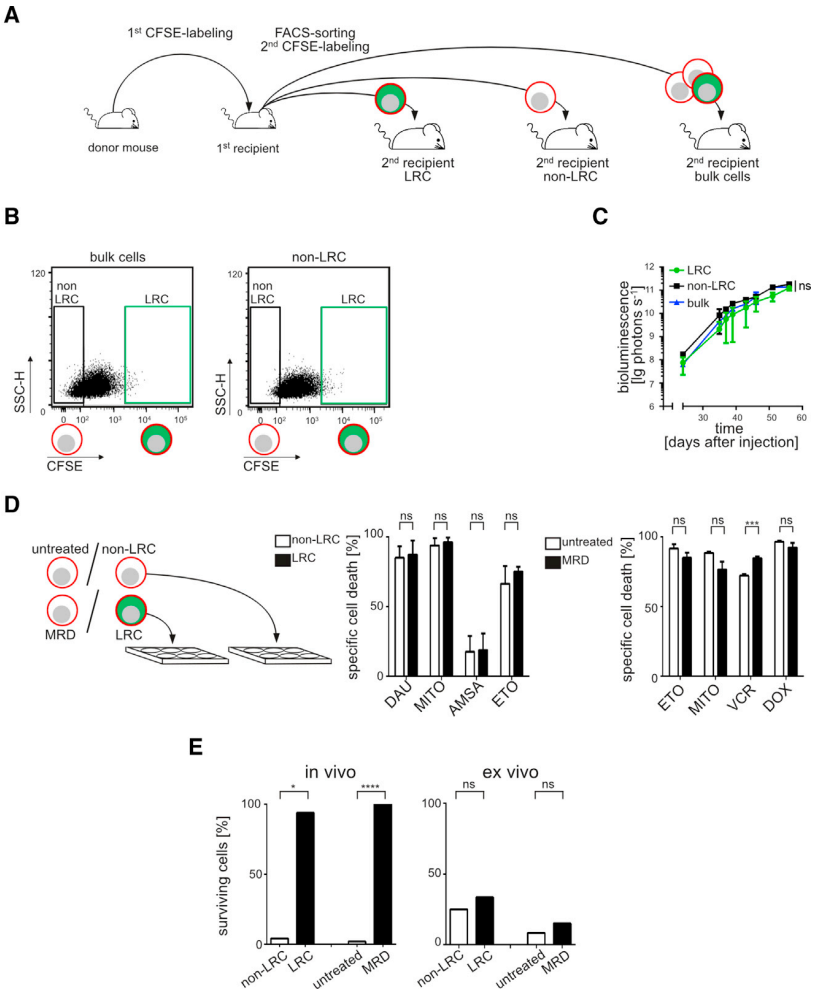


Figure 8. Release from the Environment Induces Proliferation in LRC and Sensitizes LRC and MRD Cells toward Drug Treatment

(A) From a first recipient mouse carrying CFSE-stained ALL-199 cells, LRC, non-LRC, and bulk cells were obtained at day 10; bulk cells and non-LRC were re-labeled with CFSE, re-transplanted in second recipient mice at high numbers, and re-analyzed at day 10 using flow cytometry; bulk cells, LRC, and non-LRC were re-transplanted at low numbers into groups of mice and leukemia growth was monitored over time.

(B) CFSE staining at day 10 in secondary recipient mice receiving high cell numbers.

(C) Growth curve in secondary recipients; mean ± SE; ns, no statistical significance by Kruskal-Wallis test and Dunn's multiple comparison test. One out of two independent experiments is shown.

(D) Fourteen days after transplantation, LRC or non-LRC were isolated and 500–800 cells treated ex vivo for 48 hr with daunorubicin (DAU; 250 nM), mitoxantrone (MITO; 675 nM), amsacrine (AMSA; 18 nM), or etoposide (ETO; 300 nM). Spontaneous cell death in the absence of cytotoxic drugs was 60%; a mean of eight data points from three independent experiments in triplicates or duplicates is shown for DAU and MITO and one experiment in triplicates is shown for AMSA and ETO. Four thousand untreated cells and MRD cells were treated ex vivo for 48 hr with 15 μM ETO, 450 μM MITO, 300 nM VCR, or 500 nM DOX. Cell death was measured by flow cytometry; spontaneous cell death in the absence of cytotoxic drugs was 33%; shown is one experiment in triplicate; mean ± SE; ns, not significant, ***$p < 0.001$ by two-tailed unpaired t test.

(E) Summary of ALL-265 data from Figure 4C (n = 5), S6 (n = 3), and 8D (n = 3); ns, not significant, *$p < 0.05$ and ****$p < 0.0001$ by two-tailed unpaired t test.

Here, we provide a preclinical tool to study dormant human ALL cells in vivo and show that long-term resting cells exist in ALL. This fact was previously unknown, as primary patients' samples allow quantifying non-cycling cells in a snapshot at a given moment, but fail to distinguish between short- and long-term resting cells (Lutz et al., 2013). As monitoring functionally defined cellular subpopulations such as LRC in longitudinal studies is still impossible in patients, our preclinical model enables the gaining of insights into ALL biology that cannot be obtained in patients: here the presence of long-term resting cells in ALL. Beyond its use in preclinical treatment trials, PDX models harbor major potential in basic research and enable unique insights into disease biology.

The emergence of relapse is a complex process involving genetic and non-genetic factors. Early relapse might be caused by a putatively pre-existing clone with additional mutations responsible for drug resistance, especially in adult patients. The genetic stability of most cases of ALL suggests that many relapses may not be mediated by mutational mechanisms. Late relapse might be caused by persisting, dormant tumor cells in the absence of additional mutations, and relapse cells often respond to the identical drugs used to treat the primary disease. LRC represent surrogates for late relapse and relapse in the absence of additional mutations, as often seen in children.

The fact that LRC exist might explain why ALL patients benefit from maintenance therapy, even in prognostically favorable,

(E) All genes differentially expressed (padj < 0.05) between primary samples from five children before onset of treatment to three matched MRD samples 33 days after onset of treatment.

(F) Scatterplot of fold-changes for genes differentially expressed between both LRC versus non-LRC and primary MRD versus primary diagnostic cells, grey area indicates LRC signature (left panel); GSEA comparing the LRC signature with differentially expressed genes between primary MRD and primary diagnostic cells (right panel).

(G) PCA of bulk samples transcriptomes using all shared expressed genes; each symbol indicates a single sample.

See also Figure S7, Tables S7, and S8.

chemo-sensitive ALL subtypes. ALL patients are routinely treated with oral low-dose chemotherapy from end of intensive chemotherapy until, e.g., 2 years after diagnosis, and maintenance therapy improves patients' prognosis (Schrappe et al., 2000). Low-dose maintenance therapy might act by removing LRC-type ALL cells with relapse-inducing potential that remained quiescent over prolonged periods of time and turned on their cell cycle at late time points in the months following intensive chemotherapy.

Tumor cells often display both dormancy and drug resistance. It is unclear whether either dormancy or drug resistance is pivotal in respect to the other, so that dormancy is a consequence of resistance or vice versa (Blatter and Rottenberg, 2015). Our two complementary mouse models show that LRC were defined by their dormant nature and displayed drug resistance, while MRD cells were defined by their ability to survive drug treatment and displayed a dormant phenotype. Thus, both characteristics might be equally sufficient to determine each other and coincide interdependently.

Our study shows that ALL consists of functionally heterogeneous cells regarding proliferation rate and drug resistance, similar to the functional heterogeneity shown in other tumor entities (Kreso et al., 2013). As LRC did not substantially participate in proliferation during growth of leukemia over weeks, in our model LRC existed before onset of therapy and were not developed as a consequence of treatment. As both LRC and non-LRC contain similar amounts of stem cells, but show different sensitivity toward drug treatment in vivo, our data imply that stemness and drug resistance are not directly connected in ALL.

So how does a rare subpopulation acquire the three clinically challenging features dormancy, resistance, and stemness? LRC might represent a cell subpopulation with genuinely different biology harboring distinct intrinsic, constant characteristics, or being an LRC might represent a reversible, temporary, functional phenotype depending on circumstances. In the first case, LRC and non-LRC might be organized in a hierarchical way similar to the known stem cell hierarchy existing in many tumors including AML (Kreso and Dick, 2014). In the second case, ALL cells might mimic the phenotypic reversibility of normal hematopoiesis, where long-term dormant hematopoietic stem cells start cycling in response to stress for a defined period of time and turn back into dormancy later (Trumpp et al., 2010).

Our data favor the second scenario as LRC exhibit their specific characteristics as reversible, temporary, transient functional phenotypes. Re-transplantation experiments showed that formerly dormant LRC started proliferating as soon as they were dissociated from their in vivo environment and transferred into next recipient mice. Upon re-transplantation, LRC converted into non-LRC, while certain non-LRC converted into LRC. Both LRC and non-LRC thus harbored plasticity to switch between slow and rapid proliferation depending on the current context. This fact might explain the area of overlap between LRC and non-LRC detected in single-cell RNA sequencing.

Besides proliferation, drug resistance also proved to be a transient characteristic. Drug-treatment experiments showed that LRC lost their in vivo drug resistance upon ex vivo culture. The discrepancy between drug sensitivity ex vivo and in vivo might at least partly explain the limited predictability of ex vivo drug-screening tests for the outcome of cancer patients (Wilding and Bodmer, 2014). Thus, localization of LRC to the bone marrow niche influences both dormancy and drug resistance.

These insights have translational implications. For diagnostics, as LRC lose their clinically relevant characteristics upon release from their niche, rapid sample processing might be critical for reliable profiling, which represents a challenge in clinical routine (Bacher et al., 2010). Our data at least in part explain the limited power of in vitro assays using, e.g., proliferating cell lines, for studies on MRD cells or primary leukemia cells for drug testing in the absence of feeders. Most importantly for putative treatment strategies, the transient nature of the adverse characteristics of LRC suggests aiming at removing MRD cells from their protective environment to sensitize them toward treatment (Essers et al., 2009; Essers and Trumpp, 2010). The interaction between MRD cells and their bone marrow niche represents a promising target for therapeutic approaches to prevent relapse. Beyond the tumor cell itself, its interaction with the environment represents a suitable therapeutic target. As a caveat, a persistent therapeutic inhibition of the bone marrow niche might be required over prolonged periods of time, as in principle each and every remaining non-LRC ALL cell could convert into a drug-resistant LRC, as soon as it gets access to the protective niche.

At this point, we can only speculate which signals might determine whether an ALL cell behaves like an LRC or a non-LRC. In theory, external as well as internal factors or conditions might be influential; stimuli might be sent or received either stochastically or within a well-regulated process. As our studies were restricted to bone marrow, the bone marrow niche is a likely candidate for a regulatory function and requires investigatory work (Raaijmakers, 2011). Further research is required to address these important questions. Obvious candidates for therapeutic intervention are cell surface molecules expressed on LRC, the inhibition of which might release cells from their environment. Similarly, niche cells could be targeted to aim at reducing environmental support.

Our study shows that ALL growing in vivo contains a rare subpopulation of LRC that exhibits typical challenging adverse characteristics of relapse induction, which proved to be of a reversible nature. Our model might help to develop future anti-leukemia treatment strategies allowing the eradication of the precarious subpopulation of drug-resistant stem cells to prevent relapse and improve the prognosis of patients with ALL.

EXPERIMENTAL PROCEDURE

Ethical Statements

Written informed consent was obtained from all patients and from parents/carers in the cases where the patients were minors. The study was performed in accordance with the ethical standards of the responsible committee on human experimentation (written approval by Ethikkommission des Klinikums der Ludwig-Maximilians-Universität München, Ethikkommission@med.unimuenchen.de, April 15, 2008, number 068-08) and with the Helsinki Declaration of 1975, as revised in 2000.

All animal trials were performed in accordance with the current ethical standards of the official committee on animal experimentation under the written approvals by Regierung von Oberbayern, poststelle@reg-ob.bayern.de, May 10, 2007 number 55.2-1-54-2531-2-07 and August 8, 2010 number 55.2-1-54-2531-95-10.

Enriching and Quantifying PDX and LRC from Mouse Bone Marrow

PDX ALL cells were genetically engineered as described using lentiviruses (Terziyska et al., 2012; Vick et al., 2015) to express the transgenes' truncated NGFR, a red fluorochrome, and luciferase; cells were stained with BrdU and/or CFSE before retransplantation of fresh cells into mice.

For determining the fraction of dormant PDX ALL cells, mouse bone marrow was harvested from numerous bones and enriched for human PDX ALL cells using NGFR for MACS and the red fluorochrome for flow cytometry cell sorting. LRC were discriminated from non-LRC using CFSE staining as shown in Figure 1D. CFSE mean fluorescence intensity (MFI) was measured at day 3 after injection, when bleaching had ceased, and defined cells before the onset of proliferation ("0 divisions"). Day 3 CFSE MFI was divided by factor 2 to calculate CFSE bisections mimicking cell divisions. Seven CFSE MFI bisections or more were defined as entire loss of the CFSE signal characterizing non-LRC. The LRC gate was set to include all cells harboring high CFSE signal of below three bisections of the maximum CFSE MFI (Schillert et al., 2013) (Figure 1D).

PDX Single-Cell RNA-Seq Library Construction

Single cells were isolated at 4°C and processed on the Fluidigm C1 platform. In brief, 500 cells were loaded on the 10–17 μm mRNA-seq IFC (Fluidigm) with External RNA Controls Consortium spike-in controls. Cell lysis, reverse transcription, and pre-amplification of cDNA was done on-chip using the SMARTer Ultra Low RNA Kit for C1 (Clontech). Harvested cDNA libraries of the samples (2.5 μL) were used as input for tagmentation with the Nextera XT Sample Preparation Kit (Illumina) at half the volume of Illumina's protocol. Barcoding PCR was performed for 12 cycles. Equal amounts of libraries were pooled.

RNA-Seq

Single-cell Smart-seq and bulk Smart-seq2 libraries were sequenced at 1 × 50 bases on an Illumina HiSeq1500. SCRB-seq and UMI-seq libraries were sequenced paired-end with 16 cycles on the first read to decode sample barcodes and unique molecular identifiers and 50 cycles on the second read into the cDNA fragment.

ACCESSION NUMBERS

RNA-seq data reported in this paper have been deposited in the NCBI's GEO database and are accessible through the GEO Series accession number GEO: GSE83142.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, eight figures, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.ccell.2016.11.002.

AUTHOR CONTRIBUTIONS

S.E. and E.Z.Ö. planned, performed, and analyzed the experiments and designed the data presentation; C.Z., E.Z.Ö., S.P., and W.E. generated and analyzed RNA-seq data, W.E. participated in writing the manuscript; S.T. and C.C.A. established the mouse model, first detected LRC, and started their characterization; M.G. started establishing the MRD PDX model; A.S. guided the work of E.Z.Ö.; M.D. performed enrichment of pediatric MRD cells provided by R.P.G.; C.L., V.A.T., and T.E. performed enrichment of adult StemB cells provided by H.P.; H.P.H. and K.So. performed immunohistochemistry of primary bone marrow biopsies; K.Sp. and W.H. provided primary adult samples of Figure S2; B.P., S.K., M.H., and B.K. performed and analyzed gene expression array data; J.H. performed the mathematical analysis of Figure S1E; O.G. participated in designing the experiments, guiding the study, and writing the manuscript; I.J. initiated and guided the study and wrote the manuscript.

ACKNOWLEDGMENTS

We thank Jean Pierre Bourquin and Beat Bornhäuser for providing engrafted sample ALL-265 and Cornelia Eckert and the I-BFM study group for providing the clinical data on sample ALL-265. We thank Markus G. Manz for helpful scientific discussions, Hitoshi Takizawa for help in establishing CFSE staining, Susanne Suhendra for sorting pediatric MRD cells, Volker Eckstein, Panagiotis Gitsioudis, and Linda Manta for their help in sorting stemB cells, Michael Hagemann and his team for excellent animal care, Lothar Strobl for help in array analysis, Kai Höfig for help with qPCR, Andreas Sendelhofert for establishing immunohistochemistry, Annette Frank and Volker Groiβ for help in mice experiments, Liliana Mura and Fabian Klein for technical support, and Michela Carlet, Cornelia Finkenzeller, and Binje Vick for helpful discussions. The work was supported by ERC Consolidator Grant 681524; Deutsche José Carreras Leukämie-Stiftung (R 10/26); the Collaborative Research Centers 684 Molecular Mechanisms of Normal and Malignant Hematopoiesis, project A22 and 1243 Genetic and Epigenetic Evolution of Hematopoetic Neoplasms, project A05; the German Consortium for Translational Cancer Research (DKTK); Bettina Bräu-Stiftung; and Dr. Helmut Legerlotz Stiftung (all to I.J.); Collaborative Research Center 1243, project A14, to W.E.; project 07 to K.S., and project 08 to W.H.; the German Federal Ministry of Education and Research (BMBF) within the Virtual Liver Project (grant no. 0315766) to J.H.; and the German Research Foundation grants SPP1395/InKoMBio Busch 900/6-1 to B.K. and DFG GI-540-3/1 to O.G.

Received: August 19, 2015
Revised: June 11, 2016
Accepted: October 31, 2016
Published: December 1, 2016

REFERENCES

Bacher, U., Kohlmann, A., and Haferlach, T. (2010). Gene expression profiling for diagnosis and therapy in acute leukaemia and other haematologic malignancies. Cancer Treat Rev. *36*, 637–646.

Blatter, S., and Rottenberg, S. (2015). Minimal residual disease in cancer therapy – small things make all the difference. Drug Resist Updat *21-22*, 1–10.

Bonnet, D., and Dick, J.E. (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. Nat. Med. *3*, 730–737.

Castro Alves, C., Terziyska, N., Grunert, M., Gündisch, S., Graubner, U., Quintanilla-Martinez, L., and Jeremias, I. (2012). Leukemia-initiating cells of patient-derived acute lymphoblastic leukemia xenografts are sensitive toward TRAIL. Blood *119*, 4224–4227.

Clevers, H. (2011). The cancer stem cell: premises, promises and challenges. Nat. Med. *17*, 313–319.

Eppert, K., Takenaka, K., Lechman, E.R., Waldron, L., Nilsson, B., van Galen, P., Metzeler, K.H., Poeppl, A., Ling, V., Beyene, J., et al. (2011). Stem cell gene expression programs influence clinical outcome in human leukemia. Nat. Med. *17*, 1086–1093.

Essers, M.A.G., and Trumpp, A. (2010). Targeting leukemic stem cells by breaking their dormancy. Mol. Oncol. *4*, 443–450.

Essers, M.A., Offner, S., Blanco-Bose, W.E., Waibler, Z., Kalinke, U., Duchosal, M.A., and Trumpp, A. (2009). IFNalpha activates dormant haematopoietic stem cells in vivo. Nature *458*, 904–908.

Fehse, B., Uhde, A., Fehse, N., Eckert, H.G., Clausen, J., Ruger, R., Koch, S., Ostertag, W., Zander, A.R., and Stockschlader, M. (1997). Selective

immunoaffinity-based enrichment of CD34+ cells transduced with retroviral vectors containing an intracytoplasmatically truncated version of the human low-affinity nerve growth factor receptor (deltaLNGFR) gene. Hum. Gene Ther. *8*, 1815–1824.

Gao, H., Korn, J.M., Ferretti, S., Monahan, J.E., Wang, Y., Singh, M., Zhang, C., Schnell, C., Yang, G., Zhang, Y., et al. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat. Med. *21*, 1318–1325.

Georgantas, R.W., 3rd, Tanadve, V., Malehorn, M., Heimfeld, S., Chen, C., Carr, L., Martinez-Murillo, F., Riggins, G., Kowalski, J., and Civin, C.I. (2004). Microarray and serial analysis of gene expression analyses identify known and novel transcripts overexpressed in hematopoietic stem cells. Cancer Res. *64*, 4434–4441.

Gokbuget, N., Stanze, D., Beck, J., Diedrich, H., Horst, H.A., Huttmann, A., Kobbe, G., Kreuzer, K.A., Leimer, L., Reichle, A., et al. (2012). Outcome of relapsed adult lymphoblastic leukemia depends on response to salvage chemotherapy, prognostic factors, and performance of stem cell transplantation. Blood *120*, 2032–2041.

Graham, S.M., Vass, J.K., Holyoake, T.L., and Graham, G.J. (2007). Transcriptional analysis of quiescent and proliferating CD34+ human hemopoietic cells from normal and chronic myeloid leukemia sources. Stem Cells *25*, 3111–3120.

Hong, D., Gupta, R., Ancliff, P., Atzberger, A., Brown, J., Soneji, S., Green, J., Colman, S., Piacibello, W., Buckle, V., et al. (2008). Initiating and cancer-propagating cells in TEL-AML1-associated childhood leukemia. Science *319*, 336–339.

Inaba, H., Greaves, M., and Mullighan, C.G. (2013). Acute lymphoblastic leukaemia. Lancet *381*, 1943–1955.

Kamel-Reid, S., Letarte, M., Sirard, C., Doedens, M., Grunberger, T., Fulop, G., Freedman, M., Phillips, R., and Dick, J. (1989). A model of human acute lymphoblastic leukemia in immune-deficient SCID mice. Science *246*, 1597–1600.

Kang, H., Chen, I.M., Wilson, C.S., Bedrick, E.J., Harvey, R.C., Atlas, S.R., Devidas, M., Mullighan, C.G., Wang, X., Murphy, M., et al. (2010). Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. Blood *115*, 1394–1405.

Kong, Y., Yoshida, S., Saito, Y., Doi, T., Nagatoshi, Y., Fukata, M., Saito, N., Yang, S.M., Iwamoto, C., Okamura, J., et al. (2008). CD34+CD38+CD19+ as well as CD34+CD38-CD19+ cells are leukemia-initiating cells with self-renewal capacity in human B-precursor ALL. Leukemia *22*, 1207–1213.

Kreso, A., and Dick, J.E. (2014). Evolution of the cancer stem cell model. Cell Stem Cell *14*, 275–291.

Kreso, A., O'Brien, C.A., van Galen, P., Gan, O.I., Notta, F., Brown, A.M., Ng, K., Ma, J., Wienholds, E., Dunant, C., et al. (2013). Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. Science *339*, 543–548.

Kunz, J.B., Rausch, T., Bandapalli, O.R., Eilers, J., Pechanska, P., Schuessele, S., Assenov, Y., Stutz, A.M., Kirschner-Schwabe, R., Hof, J., et al. (2015). Pediatric T-cell lymphoblastic leukemia evolves into relapse by clonal selection, acquisition of mutations and promoter hypomethylation. Haematologica *100*, 1442–1450.

le Viseur, C., Hotfilder, M., Bomken, S., Wilson, K., Röttgers, S., Schrauder, A., Rosemann, A., Irving, J., Stam, R.W., Shultz, L.D., et al. (2008). In childhood acute lymphoblastic leukemia, blasts at different stages of immunophenotypic maturation have stem cell properties. Cancer Cell *14*, 47–58.

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst *1*, 417–425.

Lutz, C., Woll, P.S., Hall, G., Castor, A., Dreau, H., Cazzaniga, G., Zuna, J., Jensen, C., Clark, S.A., Biondi, A., et al. (2013). Quiescent leukaemic cells account for minimal residual disease in childhood lymphoblastic leukaemia. Leukemia *27*, 1204–1207.

Morrison, S.J., and Spradling, A.C. (2008). Stem cells and niches: mechanisms that promote stem cell maintenance throughout life. Cell *132*, 598–611.

Raaijmakers, M.H. (2011). Niche contributions to oncogenesis: emerging concepts and implications for the hematopoietic system. Haematologica *96*, 1041–1048.

Ravandi, F., Jorgensen, J.L., O'Brien, S.M., Jabbour, E., Thomas, D.A., Borthakur, G., Garris, R., Huang, X., Garcia-Manero, G., Burger, J.A., et al. (2016). Minimal residual disease assessed by multi-parameter flow cytometry is highly prognostic in adult patients with acute lymphoblastic leukaemia. Br. J. Haematol. *172*, 392–400.

Rehe, K., Wilson, K., Bomken, S., Williamson, D., Irving, J., den Boer, M.L., Stanulla, M., Schrappe, M., Hall, A.G., et al. (2013). Acute B lymphoblastic leukaemia-propagating cells are present at high frequency in diverse lymphoblast populations. EMBO Mol. Med. *5*, 38–51.

Saito, Y., Kitamura, H., Hijikata, A., Tomizawa-Murasawa, M., Tanaka, S., Takagi, S., Uchida, N., Suzuki, N., Sone, A., Najima, Y., et al. (2010). Identification of therapeutic targets for quiescent, chemotherapy-resistant human leukemia stem cells. Sci. Transl Med. *2*, 17ra19.

Schillert, A., Trumpp, A., and Sprick, M.R. (2013). Label retaining cells in cancer – the dormant root of evil? Cancer Lett. *341*, 73–79.

Schmitz, M., Breithaupt, P., Scheidegger, N., Cario, G., Bonapace, L., Meissner, B., Mirkowska, P., Tchinda, J., Niggli, F.K., Stanulla, M., et al. (2011). Xenografts of highly resistant leukemia recapitulate the clonal composition of the leukemogenic compartment. Blood *118*, 1854–1864.

Schrappe, M. (2014). Detection and management of minimal residual disease in acute lymphoblastic leukemia. Hematology *2014*, 244–249.

Schrappe, M., Reiter, A., Zimmermann, M., Harbott, J., Ludwig, W.D., Henze, G., Gadner, H., Odenwald, E., and Riehm, H. (2000). Long-term results of four consecutive trials in childhood ALL performed by the ALL-BFM study group from 1981 to 1995. Berlin-Frankfurt-Munster. Leukemia *14*, 2205–2222.

Takizawa, H., Regoes, R.R., Boddupalli, C.S., Bonhoeffer, S., and Manz, M.G. (2011). Dynamic variation in cycling of hematopoietic stem cells in steady state and inflammation. J. Exp. Med. *208*, 273–284.

Terziyska, N., Alves, C.C., Groiss, V., Schneider, K., Farkasova, K., Ogris, M., Wagner, E., Ehrhardt, H., Brentjens, R.J., zur Stadt, U., et al. (2012). In vivo imaging enables high resolution preclinical trials on patients' leukemia cells growing in mice. PLoS One *7*, e52798.

Tesfai, Y., Ford, J., Carter, K.W., Firth, M.J., O'Leary, R.A., Gottardo, N.G., Cole, C., and Kees, U.R. (2012). Interactions between acute lymphoblastic leukemia and bone marrow stromal cells influence response to therapy. Leuk. Res. *36*, 299–306.

Townsend, E.C., Murakami, M.A., Christodoulou, A., Christie, A.L., Koster, J., DeSouza, T.A., Morgan, E.A., Kallgren, S.P., Liu, H., Wu, S.C., et al. (2016). The public repository of xenografts enables discovery and randomized phase II-like trials in mice. Cancer Cell *29*, 574–586.

Trumpp, A., Essers, M., and Wilson, A. (2010). Awakening dormant haematopoietic stem cells. Nat. Rev. Immunol. *10*, 201–209.

Vick, B., Rothenberg, M., Sandhofer, N., Carlet, M., Finkenzeller, C., Krupka, C., Grunert, M., Trumpp, A., Corbacioglu, S., Ebinger, M., et al. (2015). An advanced preclinical mouse model for acute myeloid leukemia using patients' cells of various genetic subgroups and in vivo bioluminescence imaging. PLoS One *10*, e0120925.

Visvader, J.E., and Lindeman, G.J. (2008). Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. Nat. Rev. Cancer *8*, 755–768.

Wilding, J.L., and Bodmer, W.F. (2014). Cancer cell lines for drug discovery and development. Cancer Res. *74*, 2377–2384.

Zhou, B.-B.S., Zhang, H., Damelin, M., Geles, K.G., Grindley, J.C., and Dirks, P.B. (2009). Tumour-initiating cells: challenges and opportunities for anticancer drug discovery. Nat. Rev. Drug Discov. *8*, 806–823.

Cancer Cell, Volume *30*

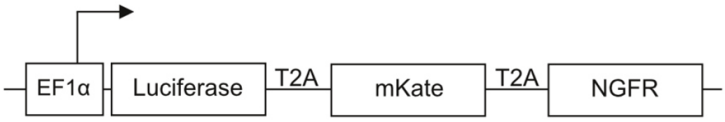
Supplemental Information

Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia

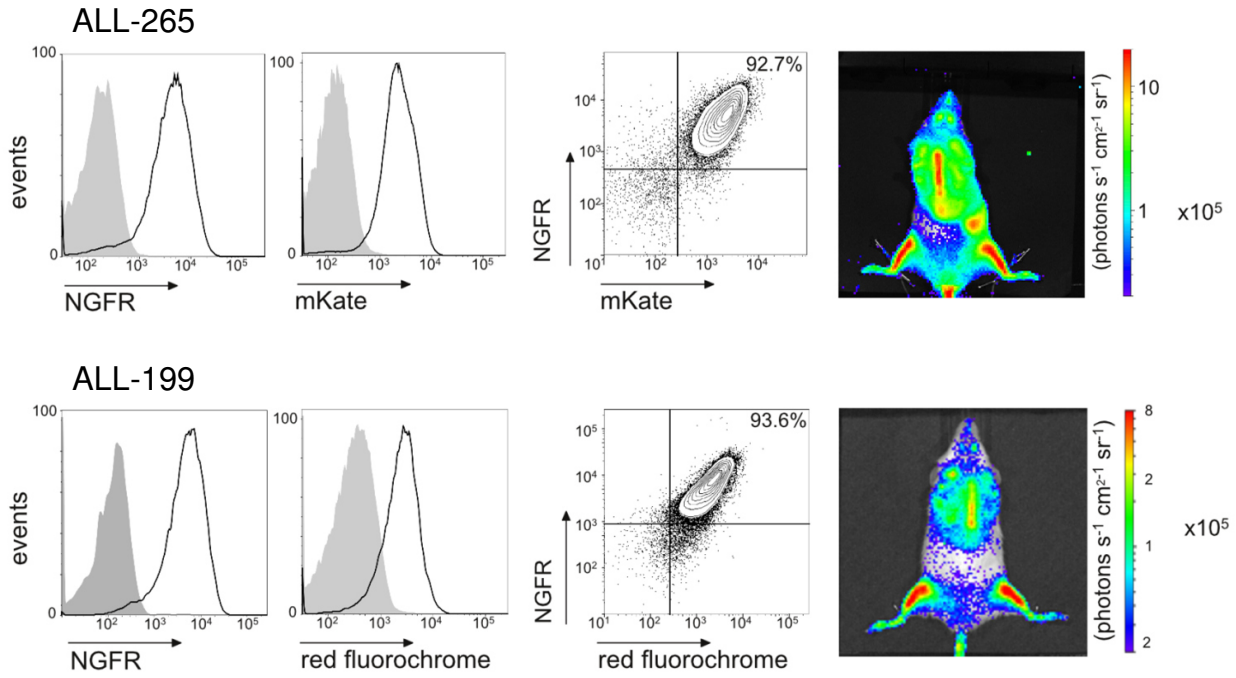
Sarah Ebinger, Erbey Ziya Özdemir, Christoph Ziegenhain, Sebastian Tiedt, Catarina Castro Alves, Michaela Grunert, Michael Dworzak, Christoph Lutz, Virginia A. Turati, Tariq Enver, Hans-Peter Horny, Karl Sotlar, Swati Parekh, Karsten Spiekermann, Wolfgang Hiddemann, Aloys Schepers, Bernhard Polzer, Stefan Kirsch, Martin Hoffmann, Bettina Knapp, Jan Hasenauer, Heike Pfeifer, Renate Panzer-Grümayer, Wolfgang Enard, Olivier Gires, and Irmela Jeremias

Supplemental Data

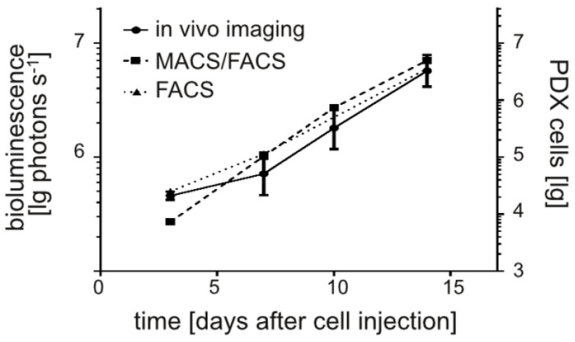
A



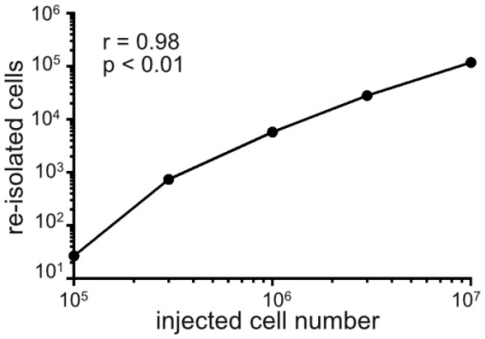
B



C



D



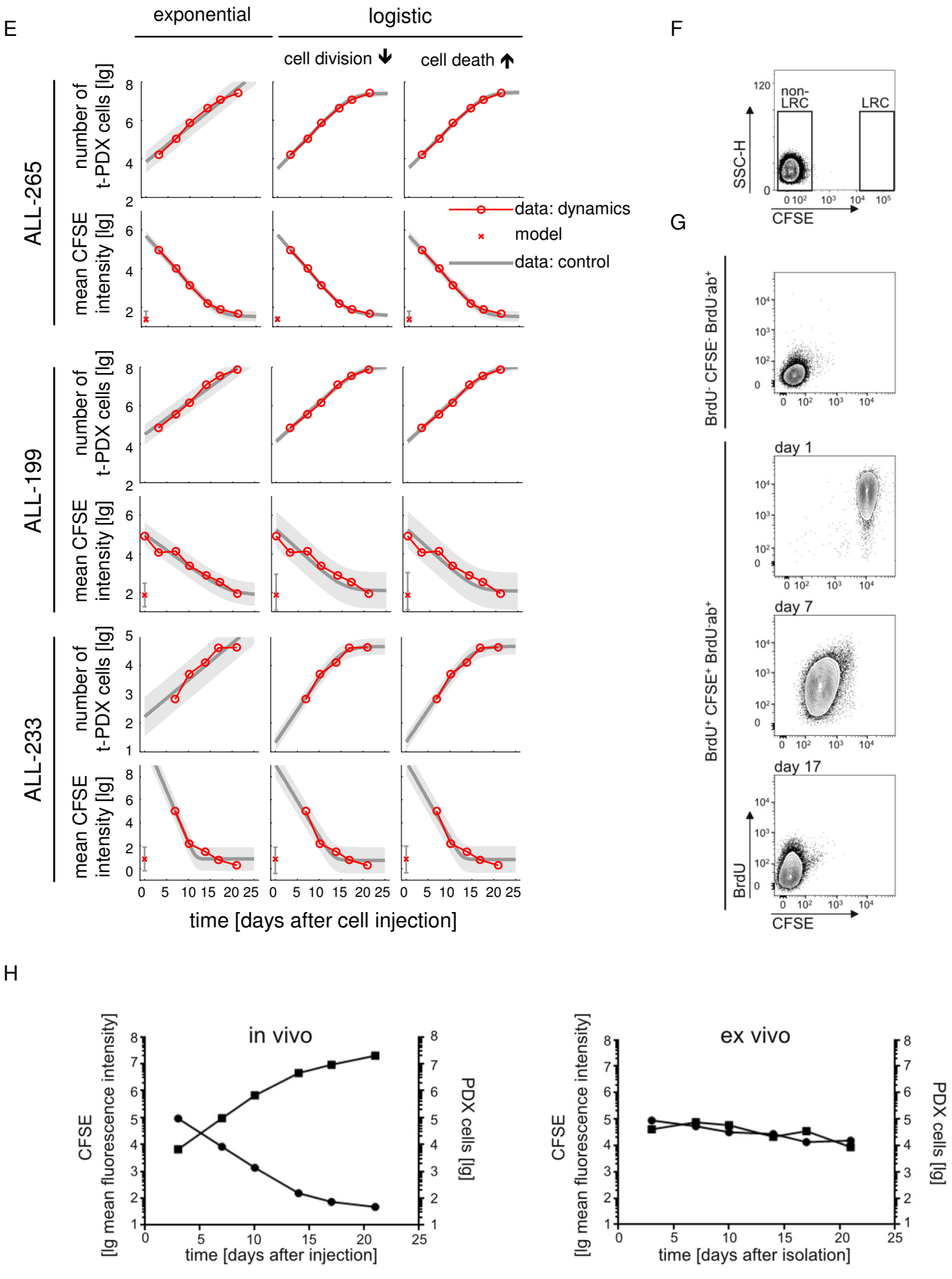

E
exponential
logistic
cell division
cell death
ALL-265
ALL-199
ALL-233
number of t-PDX cells [lg]
mean CFSE intensity [lg]
data: dynamics
model
data: control
time [days after cell injection]
F
non-LRC
LRC
SSC-H
CFSE
G
BrdU⁻ CFSE⁻ BrdU⁻ab⁺
BrdU⁺ CFSE⁺ BrdU⁻ab⁺
day 1
day 7
day 17
BrdU
CFSE
H
in vivo
CFSE [lg mean fluorescence intensity]
PDX cells [lg]
time [days after injection]
ex vivo
time [days after isolation]

Figure S1, related to Figure 1.
CFSE staining allows reliable monitoring of PDX ALL growth in mice.

(A) Lentiviral construct for equimolar expression of 3 transgenes; arrow indicates start of transcription; EF1α = elongation factor 1-alpha promoter; mKate = red fluorescent protein cloned from sea anemone Entacmaea quadricolor; NGFR = human low affinity nerve growth factor receptor lacking the intracellular signaling domain.
(B) Quality controls on enriched transgenic PDX ALL-265 or ALL-199 cells by flow cytometry or bioluminescence in vivo imaging.
(C) 10^7 ALL-265 cells were injected into groups of mice and one mouse was sacrificed at each time point. In vivo imaging was performed directly before cell harvesting and quantifying PDX cells by flow cytometry with and without prior MACS selection; mean of each group +/- standard error.
(D) Different cell numbers of ALL-199 cells were injected in mice at and re-isolated after 3 days; each dot indicates data from one animal.
(E) The measured numbers of PDX cells and the measured mean fluorescence intensities of CFSE were fitted with three mechanistic ordinary differential equation models assuming: exponential growth; logistic growth caused by a decreased rate of cell division at higher cell densities; and logistic growth caused by a increased rate of cell death at higher cell densities. The measured data (red circles and crosses), the best fit (gray line) and the noise related uncertainty intervals (gray shaded area) are depicted.
(F) No cells devoid of CFSE labeling are found in the LRC gate; flow cytometry analysis at day 0 of unlabeled ALL-265 PDX cells.
(G) Controls for BrdU and CFSE stainings; BrdU indicates feeding of mice and cells with BrdU; BrdU-ab indicates that cells were stained with the anti-BrdU antibody; “+” and “-” indicate that the procedures were performed or not, respectively.
(H) To compere behavior of PDX cells in vivo and ex vivo, 10^7 ALL-265 cells were injected into groups of mice and one mouse was sacrificed at each time point to isolate PDX cells (left panel); 10^7 fresh CFSE labeled ALL-265 PDX cells per ml were cultured on MS-5 feeder cells ex vivo (right panel).

Table S1, related to Figure 1.
Clinical data of patients donating diagnostic ALL cells for xenotransplantation and sample characteristics.

sample	type of ALL	disease stage*	age* [years]	sex	cytogenetics	passaging time§ [days]
ALL-199	BCP-ALL pediatric	2nd relapse	8	F	somatic trisomy 21; leukemic homozygous 9p deletion	42
ALL-233	BCP-ALL pediatric	initial diagnosis	<1	M	t(2;15)(p13;q15)	76
ALL-265	BCP-ALL pediatric	1st relapse	5	F	hyperploidy with additional 6, 13, 14, 17, 18, 21, X chromosome	43
ALL-435	BCP-ALL pediatric	initial diagnosis	<1	M	MLL-ENL, t(11;19)	40
ALL-50	BCP-ALL pediatric	initial diagnosis	7	F	BCR/ABL positive	45
ALL-177	BCP-ALL pediatric	initial diagnosis	8	F	TEL/AML1 positive	130
ALL-230	T-ALL pediatric	initial diagnosis	4	M	t(11;14)(p32;q11); rearrangement of TAL1-gene with the T-cell receptor locus	35
ALL-256	BCP-ALL adult	initial diagnosis	41	F	trisomy 8; BCR/ABL positive t(9;22)(q34;q11)	75
ALL-363	BCP-ALL adult	initial diagnosis	65	M	BCR-ABL positive t(9;22)(q34,q11)	60

BCP=B-cell precursor; *when the primary ALL sample was obtained; §time of passaging through mice refers to the time from injection of the sample until mice had to be sacrificed due to end stage leukemia

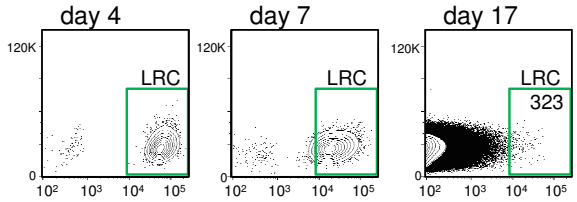
Table S2, related to Figure 1.
Two step procedure allows enrichment of minute numbers of PDX cells from mouse bone marrow.

mixed*		recovered		
mouse bone marrow cells	**PDX cells**	**number of cells**	**% recovery**	**enrichment Factor§**
$1x10^8$	1,250	1,234	99	81,000
$1x10^8$	12,500	10,262	82	9,700
$1x10^8$	37,500	34,679	92	2,666

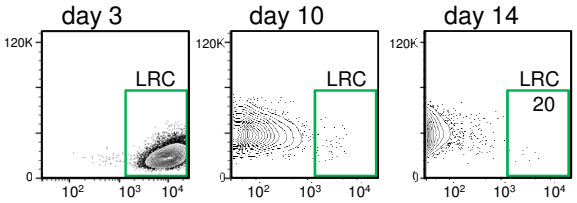
*$1x10^8$ mouse bone marrow cells were mixed with different numbers of ALL-265 PDX cells expressing NGFR and mKate; MACS-based enrichment targeting NGFR-expressing cells was followed by flow cytometry-based enrichment targeting mKate-expressing cells; §enrichment factor was calculated as ratio from “number of mouse bone marrow cells” and “recovered number of cells”

A

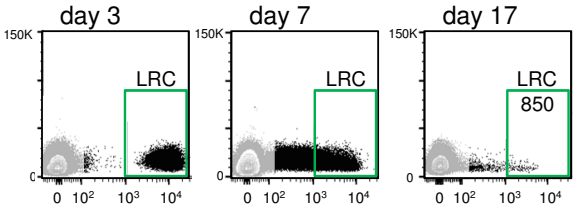
ALL-199
BCP-ALL, pediatric
2nd relapse



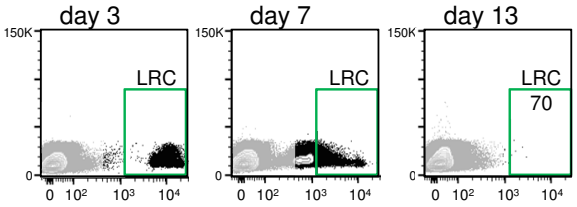
ALL-233
BCP-ALL, pediatric
initial diagnosis



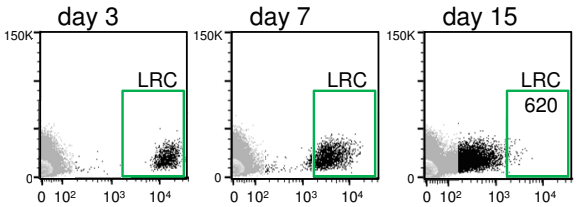
ALL-435
BCP-ALL, pediatric
initial diagnosis



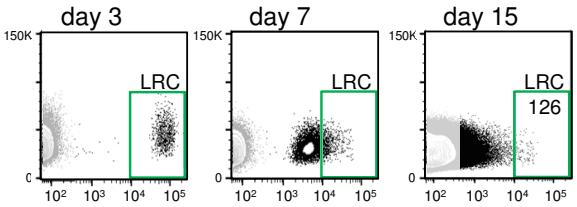
ALL-50
BCP-ALL, pediatric
initial diagnosis



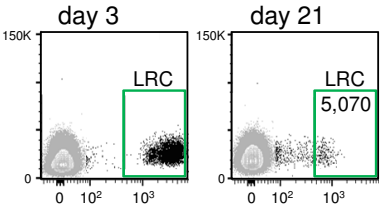
ALL-177
BCP-ALL, pediatric
initial diagnosis



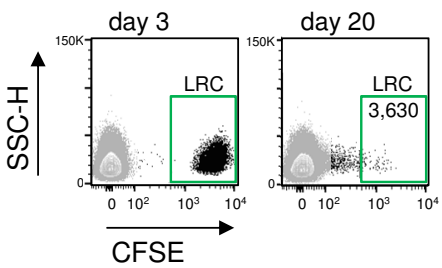
ALL-230
T-ALL, pediatric
initial diagnosis



ALL-256
BCP-ALL, adult
initial diagnosis

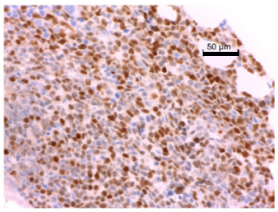


ALL-363
BCP-ALL, adult
initial diagnosis

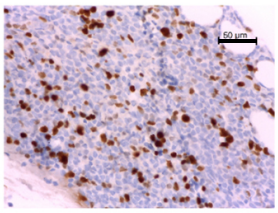


B

TdT (brown)
nucleus (blue)



Ki-67 (brown)
nucleus (blue)



TdT (red)
Ki-67 (brown)

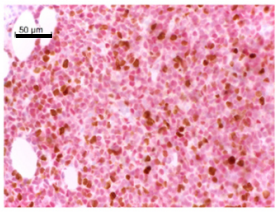


Figure S2, related to Figure 2.
A rare, long-term dormant subpopulation exists in ALL PDX cells growing in mice.

(A) 10^7 CFSE labeled ALL-199 cells/mouse were injected into 3 mice and PDX cells were enriched from bone marrow of 1 mouse at each time point using MACS sorting targeting NGFR and FACS sorting targeting mKate; LRC and non-LRC were quantified by flow cytometry. One representative out of at least 10 independent experiments is shown. All further PDX samples did not express transgenes. Here, 10% of the entire bone marrow isolate was analyzed without a prior MACS enrichment step. Unstained cells represent mouse bone marrow cells and non-LRC. Day = number of days after injection of CFSE-labeled cells.

(B) Immunohistochemistry was performed using TdT to visualize all ALL blasts and Ki-67 to visualize proliferating cells in the diagnostic BM biopsy from one 69 years old female patient with BCR/ABL positive normal karyotype ALL; double staining (lowest panel) indicates frequent dormant ALL blasts as TdT positive, Ki-67 negative cells. Hemalum staining was used for nuclei; scale bar represents 50 µm.

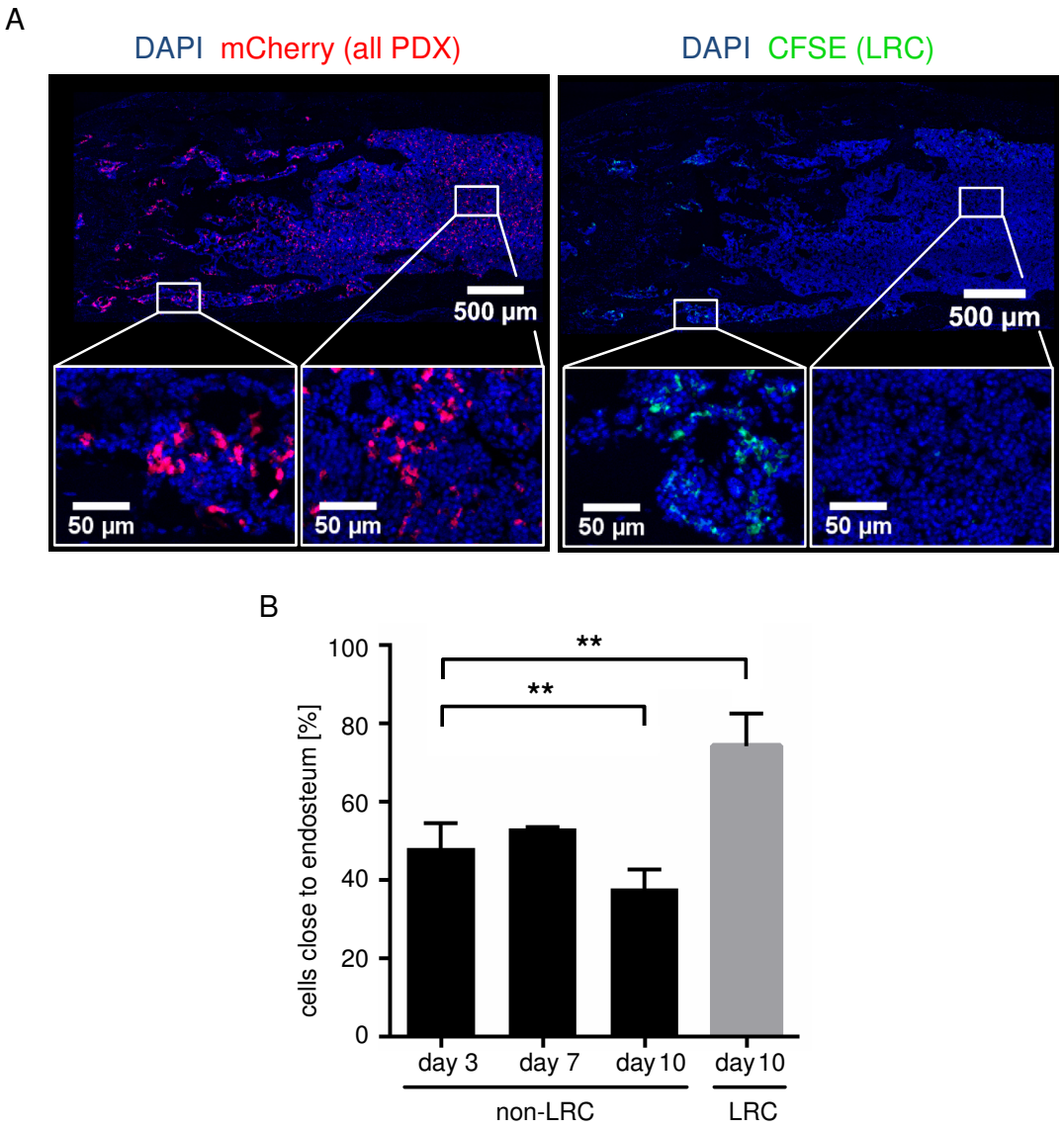


Figure S3, related to Figure 3.
LRC localize to the endosteum in ALL-199.

(A) Immunohistochemistry of consecutive murine bone marrow femur sections 10 days after injection of CFSE-stained PDX ALL-199 cells; mCherry (red) indicates all PDX cells, CFSE (green) indicates LRC.
(B) Kinetic for ALL-199; mean +/- standard error; * $p<0.05$; ** $p<0.01$ by two-tailed unpaired t-test.

Table S3, related to Figure 3.
LRC and non-LRC harbor similar numbers of leukemia initiating cells (LIC).

sample	number of cells injected per mouse*		time [days after injection] 20	28	41	48	62	75
ALL-265	LRC	333	0/2	0/2	0/2	0/2	2/2	2/2
		100	0/5	0/5	0/5	0/5	2/5	3/5
		10						8/19
	LIC frequency		1/40 (CI = 95%; lower = 1/84, upper = 1/19)					
	non-LRC	3333	0/3	1/3	3/3	3/3	3/3	3/3
		1000	0/5	0/5	2/5	4/5	5/5	5/5
		333	0/5	0/5	0/5	0/5	4/5	4/5
		100	0/5	0/5	0/5	0/5	2/5	2/5
		10						8/20
	LIC frequency		1/85 (CI = 95%; lower = 1/179, upper = 1/40)					
	number of cells injected per mouse		time [days after injection] 35	42	49	56	69	77
ALL-199	LRC	333	0/3	0/3	0/3	0/3	2/3	3/3
		100	n.d.	0/4	0/4	0/4	1/4	3/4
	LIC frequency		1/69 (CI = 95%; lower = 1/209, upper = 1/23)					
	non-LRC	1000	1/5	3/5	5/5	5/5	5/5	
		333	0/3	2/3	2/3	3/3	3/3	3/3
		100	0/4	1/4	1/4	3/4	4/4	4/4
	LIC frequency		1/100 or higher					

*LRC and non-LRC obtained 14 days after injection of CFSE labeled ALL-265 or ALL-199 cells were transplanted into secondary recipient mice in limiting dilutions at numbers indicated; bioluminescence in vivo imaging was performed repetitively at the indicated time points to determine engraftment; LIC frequency was calculated using the ELDA software; CI = confidence interval

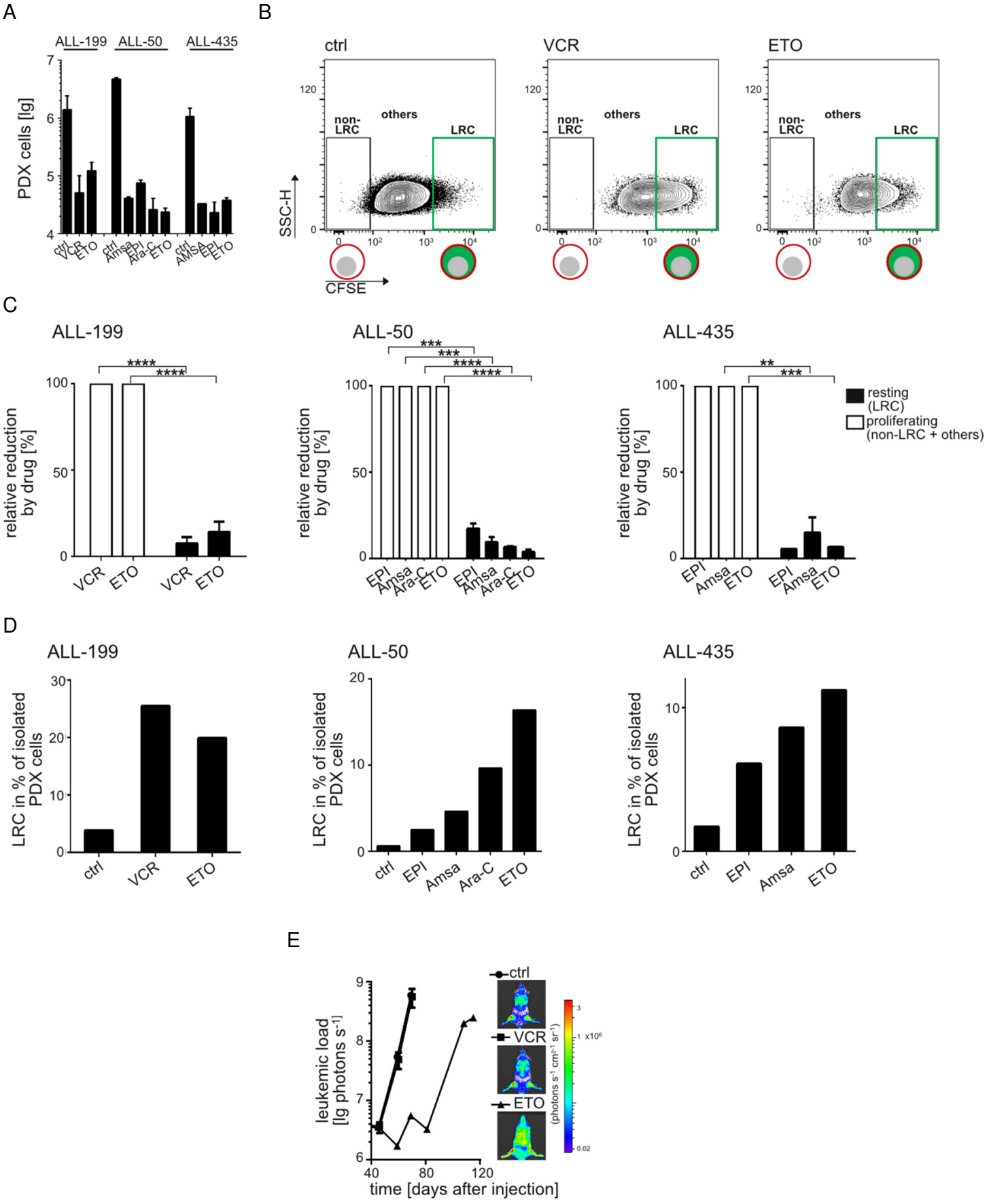

A
ALL-199
ALL-50
ALL-435
PDX cells [lg]
ctrl
VCR
ETO
ctrl
Amsa
EPI
Ara-C
ETO
ctrl
AMSA
EPI
ETO
B
ctrl
VCR
ETO
non-LRC
others
LRC
SSC-H
CFSE
C
ALL-199
ALL-50
ALL-435
relative reduction by drug [%]
resting (LRC)
proliferating (non-LRC + others)
VCR
ETO
EPI
Amsa
Ara-C
D
ALL-199
ALL-50
ALL-435
LRC in % of isolated PDX cells
ctrl
VCR
ETO
EPI
Amsa
Ara-C
E
leukemic load [lg photons s-1]
time [days after injection]
ctrl
VCR
ETO
(photons s-1 cm-2 sr-1)
x10^6

Figure S4, related to Figure 4.
LRC survive systemic drug treatment in vivo.

Mice were injected with 10^7 CFSE-labeled ALL PDX cells/mouse, were treated on day 7 and sacrificed on day 10; LRC and non-LRC were analyzed and re-transplanted into 1-2 secondary recipient mice at 2,000-5,000 LRC per mouse.

(A) Numbers of PDX cells isolated from mice with and without prior systemic drug treatment; mean of each group (n=8-11) +/- standard error.
(B) For ALL-199, a second relapse, 11 mice were treated with a single application of vincristine (VCR, 1.5 mg/kg i.v.), 8 mice were treated with a single application of etoposide (ETO, 75 mg/kg i.p.) and 8 control mice received buffer; shown are original data of representative mice.
(C) Quantification in all mice per group depicted as mean of relative drug effects on LRC compared to non-LRC (100%) +/- standard error. For ALL-50, a sample obtained at initial diagnosis, drugs were applied daily over 3 days and 2 mice were treated with cytarabine (AraC, 150 mg/kg i.p.), 2 mice with ETO (33 mg/kg i.p.), 2 mice with amsacrine (Amsa, 25 mg/kg i.p.) and 2 mice with epirubicine (EPI, 25 mg/kg i.p., single application). For ALL-435, another sample obtained at initial diagnosis, drugs were applied daily over 3 days and 2 mice were treated with ETO (33 mg/kg, i.p.), 2 mice with Amsa (25 mg/kg i.p.) and one mouse with EPI (25 mg/kg i.p., single application). ** $p< 0.01$, *** $p<0.001$, **** $p<0.0001$ by two-tailed unpaired t-test.
(D) Mean relative proportion of LRC in total PDX cells with and without treatment.
(E) To study their stem cell potential, LRC of ALL-199 LRC were isolated after treatment, re-transplanted and growth monitored by in vivo imaging mean of each group (n=1-2) +/- standard error. Imaging pictures from dpi 60 (ctrl, VCR) and dpi 108 (ETO).

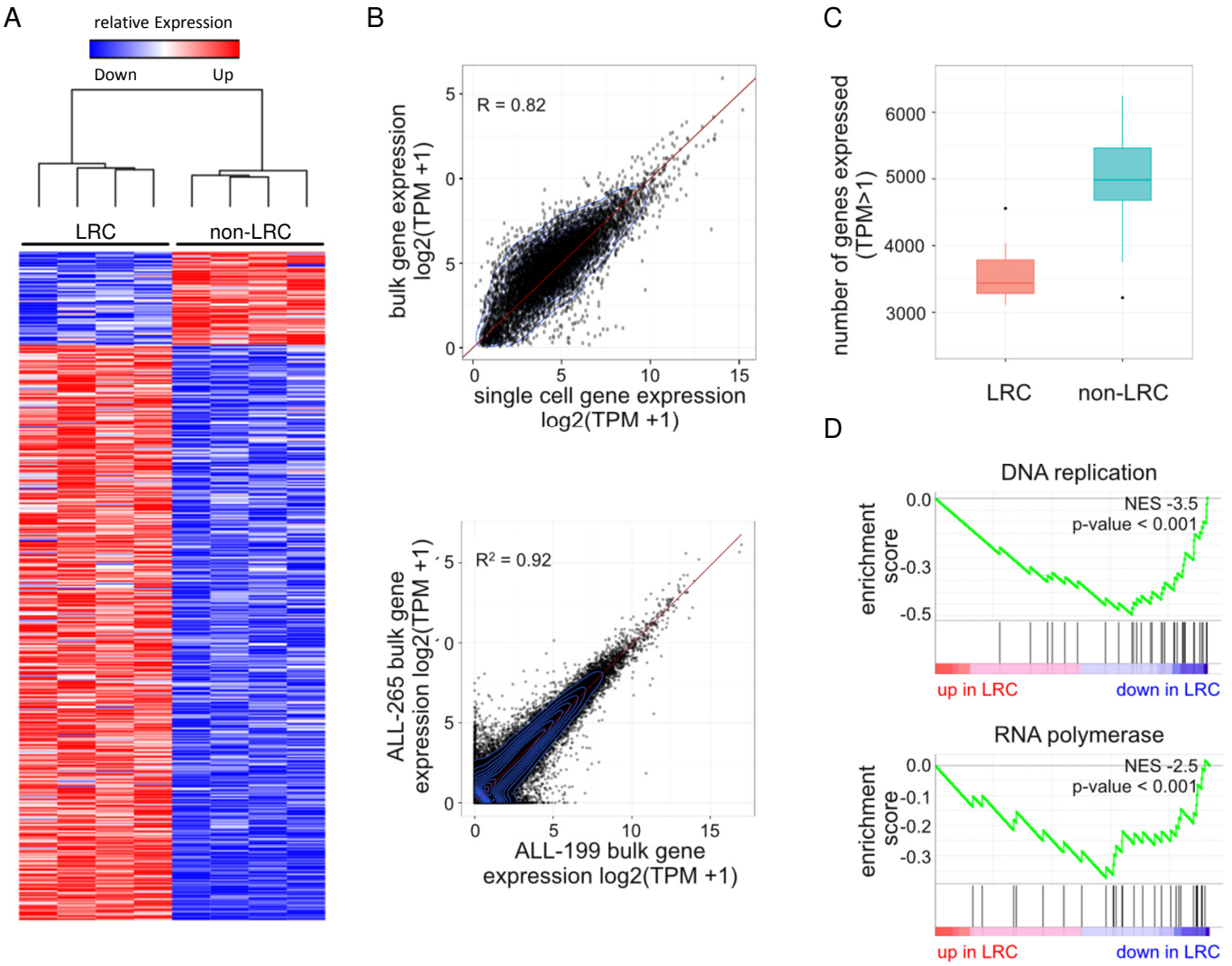


Figure S5, related to Figure 5.
Expression profile of LRC shows distinct changes to non-LRC.

15 days after transplantation of CFSE labeled PDX cells, LRC and non-LRC were subjected to RNA sequencing. For ALL-265, high quality single cell mRNA seq profiles were obtained from 15 LRC and 35 non-LRC cells. To combine single-cell and bulk RNA-seq data, median count data of single-cell experiments were summarized as a single expression profile for each LRC and non-LRC.

(A) Hierarchical clustering and gene expression heatmap across the 500 most differentially expressed genes comparing LRC and non-LRC in ALL-199 ($p < 0.01$).
(B) Comparison of Transcript Per Million (TPM) expression values between bulk versus single-cell ALL-265 (upper) and ALL-265 versus ALL-199 (lower).
(C) Quantification of expressed genes per cell (TPM > 1) in LRC versus non-LRC according to single-cell RNA-seq of ALL-265; shown is the median with upper/lower quartile and maximum/minimum, outliers are shown as dots.
(D) Gene set enrichment analysis for indicated KEGG pathways and the genes differentially regulated in LRC versus non-LRC.

Table S4, related to Figure 5.
List of 500 most differentially expressed genes between LRC and non-LRC in single cell RNA sequencing of ALL-265

Provided as an Excel file.

Table S5, related to Figure 5. Integrated LRC signature.

Provided as an Excel file.

Table S6, related to Figure 5.
KEGG pathways enriched with LRC versus non-LRC differentially expressed genes in combined analysis

Provided as an Excel file.

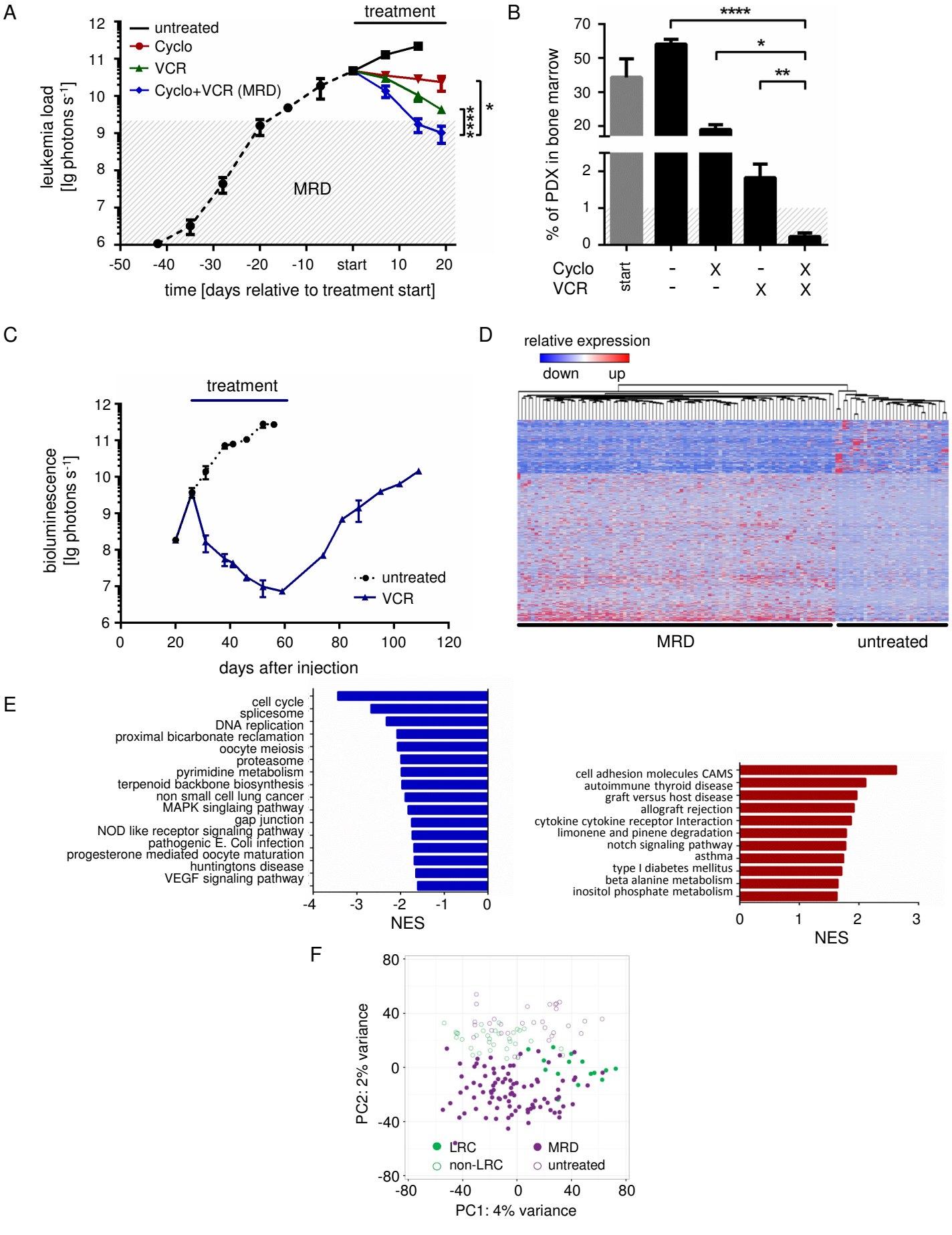

A
treatment
untreated
Cyclo
VCR
Cyclo+VCR (MRD)
leukemia load
[lg photons s-1]
MRD
time [days relative to treatment start]
B
****
*
**
% of PDX in bone marrow
Cyclo
VCR
start
C
treatment
bioluminescence
[lg photons s-1]
untreated
VCR
days after injection
D
relative expression
down
up
MRD
untreated
E
cell cycle
splicesome
DNA replication
proximal bicarbonate reclamation
oocyte meiosis
proteasome
pyrimidine metabolism
terpenoid backbone biosynthesis
non small cell lung cancer
MAPK singlaing pathway
gap junction
NOD like receptor signaling pathway
pathogenic E. Coli infection
progesterone mediated oocyte maturation
huntingtons disease
VEGF signaling pathway
NES
cell adhesion molecules CAMS
autoimmune thyroid disease
graft versus host disease
allograft rejection
cytokine cytokine receptor Interaction
limonene and pinene degradation
notch signaling pathway
asthma
type I diabetes mellitus
beta alanine metabolism
inositol phosphate metabolism
NES
F
PC2: 2% variance
LRC
non-LRC
MRD
untreated
PC1: 4% variance

Figure S6, related to Figure 6.
Characterization of cells at minimal residual disease.

(A-B) 10^7 ALL-265 cells were injected into 28 mice; when 40 % of bone marrow cells were human, therapy was started using vincristine (VCR, 0.25 mg/kg; n=4) or cyclophosphamide (Cyclo, 100 mg/kg; n=4) or a combination thereof (VCR+Cyclo; n=12), weekly for 3 weeks; VCR+Cyclo combination treatment had reduced tumor burden to minimal residual disease (MRD; < 1% human cells in bone marrow).
(A) Mean of each group +/- standard error; * $p<0.05$, **** $p<0.0001$ by two-tailed unpaired t-test; mice receiving buffer had to be sacrificed after two weeks of treatment due to end stage leukemia.
(B) Percentage of PDX ALL cells in mouse bone marrow as determined by flow cytometry post mortem as mean +/- standard error; * $p<0.05$, ** $p<0.01$, **** $p<0.0001$ by two-tailed unpaired t-test.
(C) To study their behavior after release of treatment pressure, ALL-199 cells were injected into 4 mice per group which were repetitively monitored by in vivo imaging; at substantial tumor burden, mice were treated with Vincristine (VCR) 0.4 mg/kg and left untreated thereafter; mean of each group +/- standard error.
(D-F) ALL-199 cells were injected into 19 mice; when 30 % of bone marrow cells were human, 5 untreated samples were harvested and one mouse were subjected to single cell sequencing; remaining mice received either buffer or vincristine (VCR, 0.25 mg/kg; n=5) or cyclophosphamide (Cyclo, 100 mg/kg; n=3) or a combination thereof (VCR+Cyclo; n=6) weekly for 2 weeks; when VCR+Cyclo combination treatment had reduced tumor burden to minimal residual disease (MRD; < 1% human cells in bone marrow), cells from the 6 VCR+Cyclo mice were isolated and one mouse were subjected to single cell sequencing.
(D) Hierachical clustering and gene expression heatmap across the 500 most differentially expressed genes between MRD cells and untreated cells in ALL-199 single cell RNA sequencing (MRD cells n=90; untreated cells n=32; $p < 0.01$; for gene annotation see Table S7).
(E) Significantly enriched KEGG pathways ($p < 0.05$) in MRD cells versus untreated cells as determined by geneset enrichment analysis.
(F) Principle component analysis of transcriptomes of 32 untreated control ALL-199 single cells and 90 MRD cells together with single cell data from LRC and non-LRC as in Figure 5C.

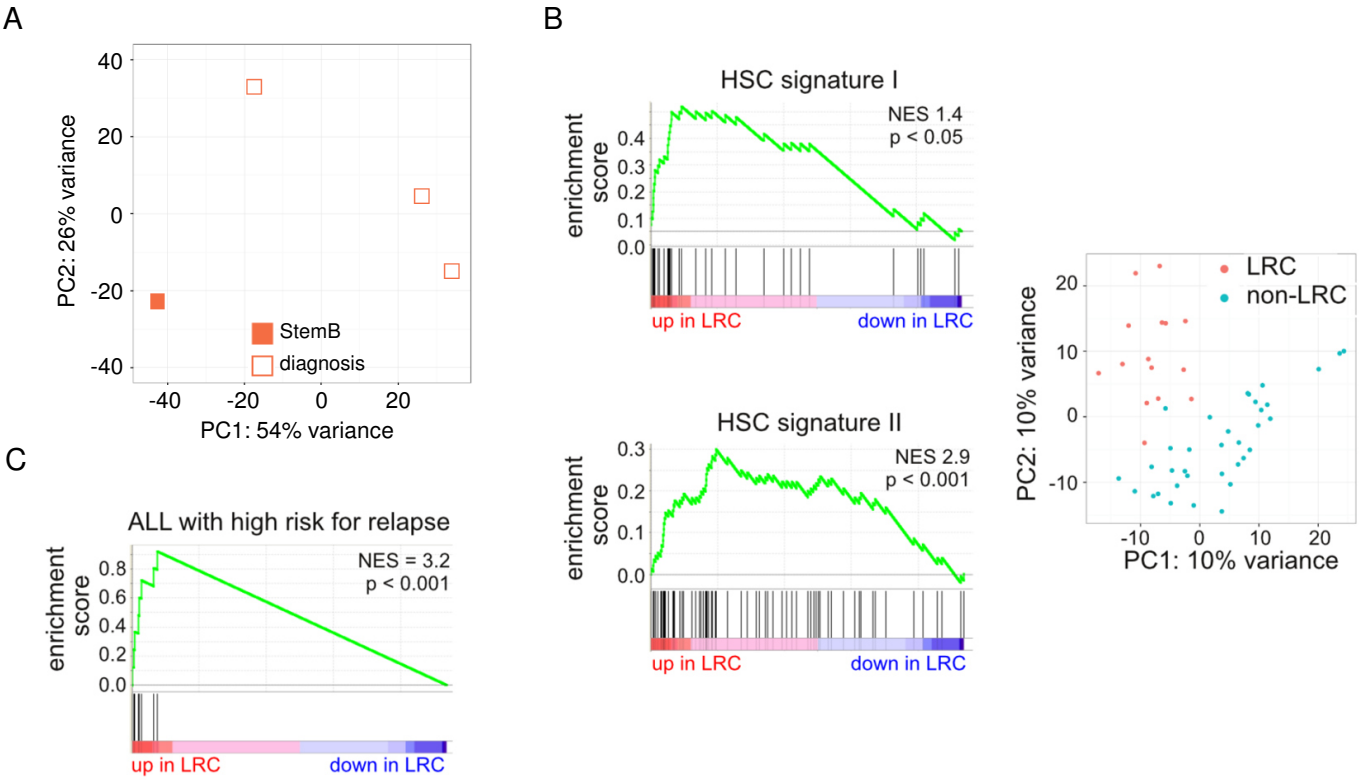


Figure S7, related to Figure 7.
LRC resemble primary MRD cells from patients.

(A) Principal component analysis of the bulk adult StemB sample compared to 3 bulk diagnosis samples of adult patients with BCR-ABL positive ALL.

(B) Gene set enrichment analysis of genes differentially expressed in LRC versus non-LRC and published signatures; HSC signature I = Georgantas et al., 2004; HSC signature II = Eppert et al., 2011 (left panel). Principal component analysis (PCA) of ALL-265 single cells on the basis HSC marker genes (Eppert et al., 2011) (right panel).

(C) Geneset enrichment analysis for a published gene signature prognostic for ALL with high risk of relapse (Kang et al., 2010).

Table S7, related to Figure 7.
Clinical data from BCP ALL patients of transcriptomes at diagnosis and/or MRD.

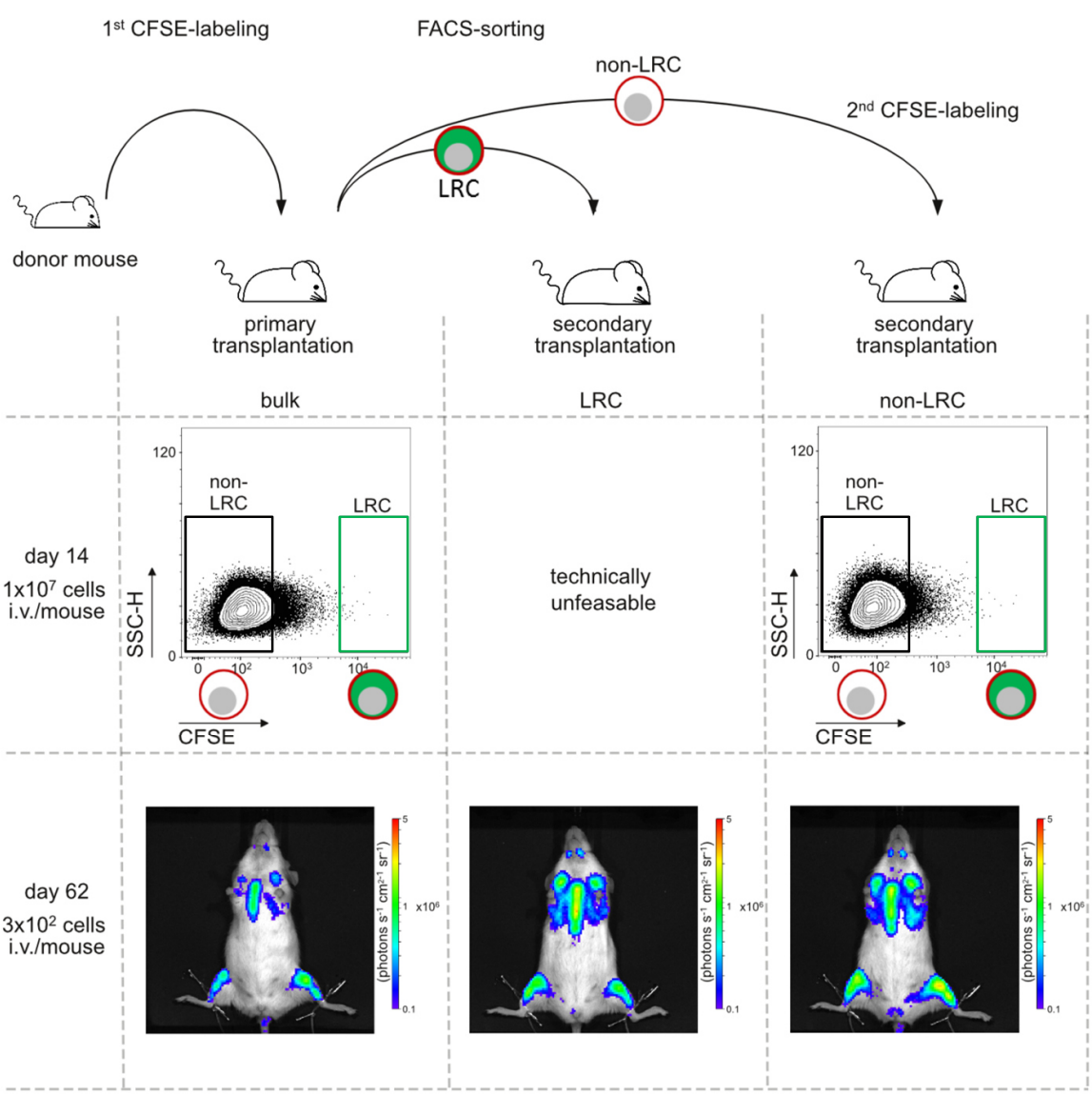
sample	age#	sex	multi-center study	genetic subtype	flow RG	proto-col RG&	stage after induc-tion II	day of MRD measure-ment §	BM blasts at MRD (%)$	sort
1	38	F	GMALL 0703	BCR-ABL	na	VHR	CR	71	0.24	StemB*
2	39	M	GMALL 0703	BCR-ABL	na	VHR	CR	71	0.32	StemB*
1	4	F	BFM 2009	ETV6/ RUNX1	MR	MR	na	na	na	$CD19^{+}$, $CD10^{+++}$, $CD20^{-}$
2	3	F	BFM 2009	ETV6/ RUNX1	MR	SR	na	na	na	$CD19^{+}$, $CD99^{bright}$, $CD10^{+++}$
3	5	M	BFM 2009	HD	MR	HR	na	33	0.69	$CD19^{+}$, $CD10^{+}$, $CD123^{+}$
4	18	M	BFM 2009	B OTHER	MR	HR	na	33	1.10	$CD19^{+}$, $CD10^{++}$, $CD45^{-/dim}$
5	3	F	BFM 2009	HD	MR	MR	na	33	0.13	$CD19^{+}$, $CD10^{++}$, $CD20^{dim}$

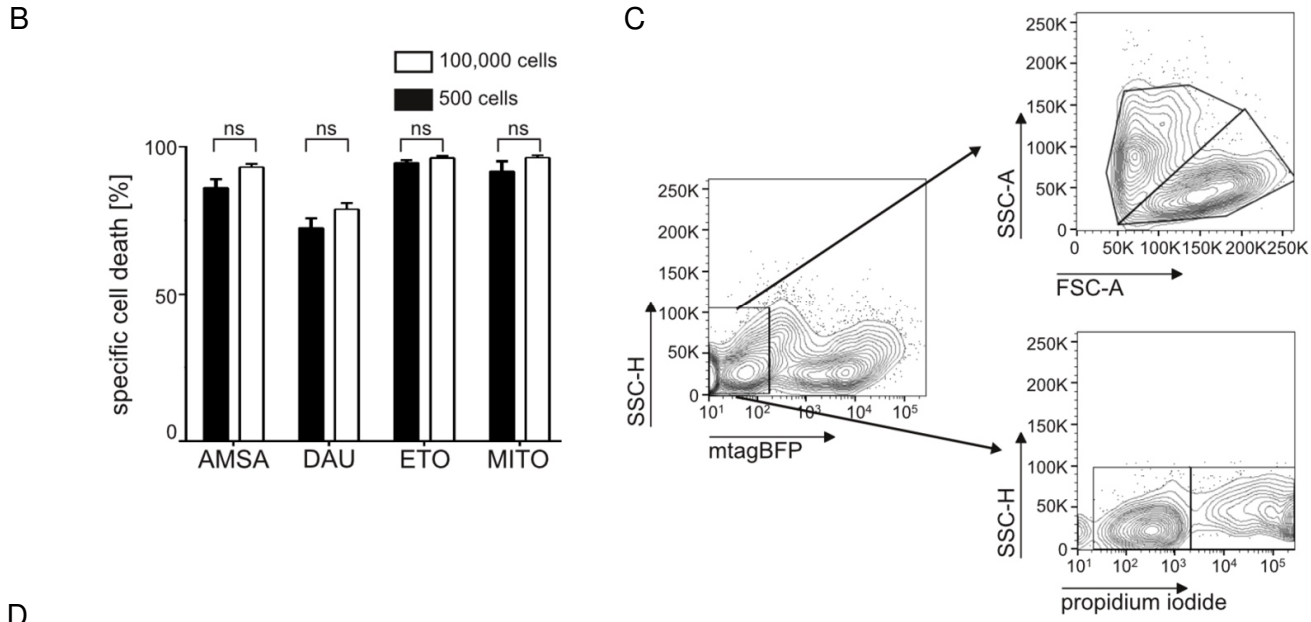
#age at diagnosis in years; F=female; M=male; GMALL=German Multicenter ALL Study Group; BFM=Berlin-Frankfurt-Münster; HD=high hyperdiploid karyotype; RG=risk group; na=not applicable; MR=medium risk; VHR=very high risk; SR=standard risk; HR=high risk; &therapy risk group (RG) assignment; §days after onset of treatment; BM=bone marrow; $in BCR-ABL positive samples, MRD was quantified by PCR using the BCR-ABL/ABL ratio; *StemB cells are $CD19^{+}$, $CD34^{+}$, $CD38^{-/low}$ according to Lutz et al., 2013; Hong et al., 2008; Castor et al., 2005

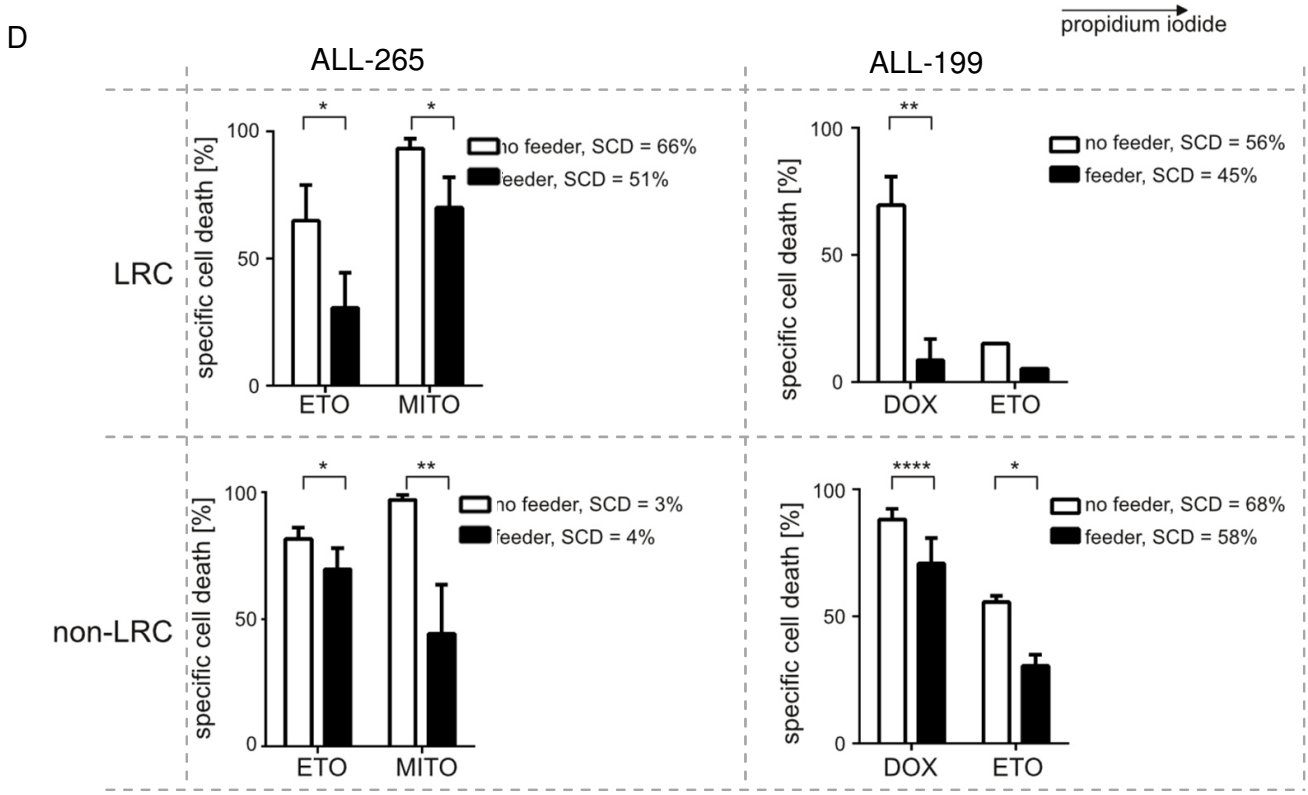
Table S8, related to Figure 7.
List of most significantly differentially expressed genes between primary samples from 5 primary ALL diagnosis and 3 MRD samples after 33 days of treatment.

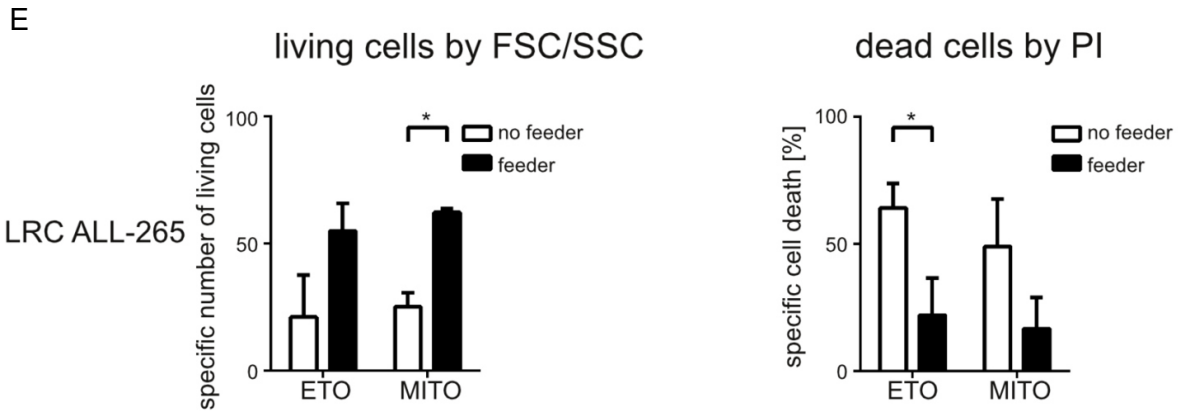
Provided as an Excel file.

A

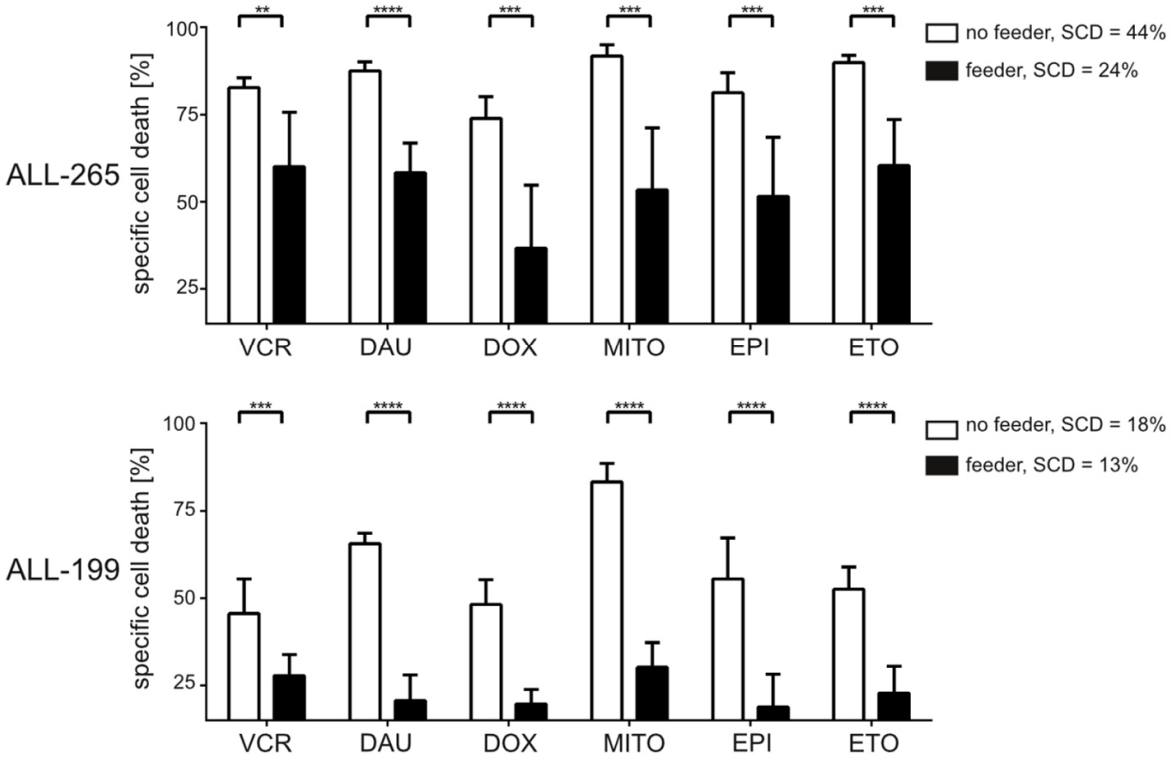

1st CFSE-labeling
FACS-sorting
non-LRC
2nd CFSE-labeling
LRC
donor mouse
primary transplantation
secondary transplantation
secondary transplantation
bulk
LRC
non-LRC
day 14
1x10^7 cells i.v./mouse
non-LRC
LRC
SSC-H
CFSE
technically unfeasable
day 62
3x10^2 cells i.v./mouse
(photons s-1 cm-2 sr-1)
x10^6


B
100,000 cells
500 cells
ns
specific cell death [%]
AMSA
DAU
ETO
MITO
C
SSC-A
FSC-A
SSC-H
mtagBFP
propidium iodide


D
ALL-265
ALL-199
LRC
non-LRC
specific cell death [%]
no feeder, SCD = 66%
feeder, SCD = 51%
no feeder, SCD = 56%
feeder, SCD = 45%
no feeder, SCD = 3%
feeder, SCD = 4%
no feeder, SCD = 68%
feeder, SCD = 58%
ETO
MITO
DOX


E
living cells by FSC/SSC
dead cells by PI
LRC ALL-265
specific number of living cells
specific cell death [%]
no feeder
feeder
ETO
MITO

F


ALL-265
specific cell death [%]
100
75
50
25
**
****
***
***
***
***
VCR
DAU
DOX
MITO
EPI
ETO
no feeder, SCD = 44%
feeder, SCD = 24%
ALL-199
specific cell death [%]
100
75
50
25
***
****
****
****
****
****
VCR
DAU
DOX
MITO
EPI
ETO
no feeder, SCD = 18%
feeder, SCD = 13%

G

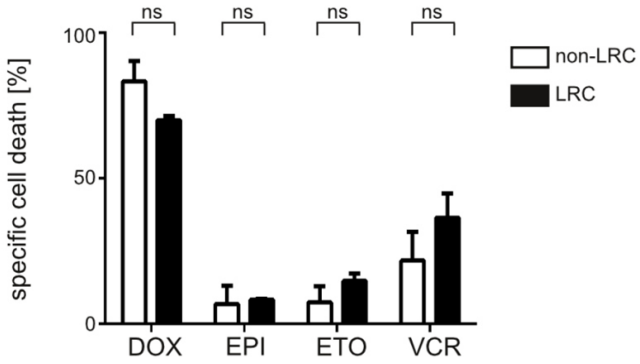

specific cell death [%]
100
50
0
ns
ns
ns
ns
DOX
EPI
ETO
VCR
non-LRC
LRC

Figure S8, related to Figure 7.
Identical growth behavior upon re-transplantation and identical ex vivo drug sensitivity in LRC versus non-LRC.

(A) Upper panel shows experimental procedure; ALL-265 cells were amplified in donor mice, CFSE labeled, re-transplanted into primary recipients and re-isolated after 14 days (left lane). Cells were separated into LRC (middle lane) and non-LRC (right lane) and re-transplanted into secondary recipients which were imaged after 62 days (lower row). non-LRC were additionally re-labeled with CFSE and re-transplanted at high numbers which was unfeasible for LRC due to their minor abundance (middle panel).
(B) 500 or 100,000 freshly isolated non-LRC (ALL-265) were stimulated ex vivo for 48 hours with the following cytotoxic drugs: amsacrine (AMSA; 18 µM), daunorubicine (DAU; 250 nM); etoposide (ETO; 30 µM) and mitoxantrone (MITO; 675 nM); shown is one experiment in triplicates +/- standard error; ns = not significant by two-tailed unpaired t-test. Specific cell death was determined by DAPI staining and specific cell death calculated thereof.
(C-F) Freshly isolated PDX cells were seeded in triplicates in the presence or absence of irradiated MS-5 cells expressing the blue fluorochrome mtagBFP. Cells were stimulated for 48-72 hours and all cells per well were removed by trypsin digestion and analyzed by flow cytometry. * $p<0.05$, ** $p<0.01$, *** $p<0.001$ and **** $p<0.0001$ by two-tailed unpaired t-test. SCD = spontaneous cell death in the absence of cytotoxic drugs.
(C) Feeder cells were excluded by gating on non-blue/mtagBFP-expressing cells; living cells were quantified in absolute and relative amounts using either forward/side scatter analysis or propidium iodide staining with similar results.
(D, E) 700-1,900 fresh LRC or non-LRC were stimulated with the following drugs: Etoposide (ETO; 3 µM) or mitoxantrone (MITO; 0.45 µM) for ALL-265 and etoposide (ETO; 15 µM) or doxorubicine (DOX; 0.5 µM) for ALL-199. E shows results obtained by forward sideward scatter analysis, F shows results obtained by propidium iodide (PI) staining as well as absolute number of surviving cells as estimated in forward/side scatter analysis. Shown are mean of up to 3 independent experiments; +/- standard error.
(F) 100,000 unsorted PDX ALL-265 were stimulated with vincristine (VCR; 0.3 µM), daunorubicine (DAU; 0.25 µM), doxorubicine (DOX; 0.5 µM), mitoxantrone (MITO; 0.45 µM), epirubicine (EPI; 0.4 µM) and etoposide (ETO; 3 µM) for 48 hours; 100,000 unsorted PDX ALL-199 were stimulated with vincristine (VCR; 0.03 µM), daunorubicine (DAU; 0.25 µM), doxorubicine (DOX; 0.5 µM), mitoxantrone (MITO; 0.25 µM), epirubicine (EPI; 0.4 µM) and etoposide (ETO; 3 µM) for 72 hours; mean of 9 data points from 3 independent experiments in triplicates is shown; +/- standard error; Welch's correction war required in two-tailed unpaired t-test for ALL-265 and DAU stimulation in ALL-199.
(G) 14 days after transplantation, LRC or non-LRC were isolated and 500-800 cells stimulated ex vivo for 48 hours with the following cytotoxic drugs: doxorubicine (DOX; 500 nM), epirubicine (EPI; 500 nM); etoposide (ETO; 30 µM) and vincristine (VCR; 300 nM). Specific cell death was determined after 48h by forward-side scatter and by DAPI staining and specific cell death calculated thereof; mean of 6 data points from 2 independent experiments in triplicates is shown +/- standard error; ns = not significant by two-tailed unpaired t-test.

Supplemental Experimental Procedures

The NSG mouse model of individual ALL

ALL blasts were obtained from children and adults treated within clinical multicenter studies. NSG mice (NOD/scid, IL2 receptor gamma chain knockout mice) were obtained from The Jackson Laboratory (Lund, Sweden). The animal model was performed as described (Liem et al., 2004). Briefly, fresh primary ALL cells were isolated by Ficoll gradient centrifugation from peripheral blood or bone marrow aspirates that had been obtained from leftovers of clinical routine sampling before onset of therapy. 10 million ALL cells were injected into 6-12 weeks old NSG mice via the tail vein. Engraftment was monitored by 2-weekly flow cytometry measurement of human cells in peripheral blood starting at week 6. ALL-265 was first engrafted by Jean Pierre Bourquin and Beat Bornhäuser in Zurich. Mice were sacrificed at first clinical signs of disease, as measured by quantification of human cells in peripheral blood or by in vivo imaging. From engrafted mice, PDX ALL cells were harvested from enlarged spleens and either directly re-injected or frozen at -190 °C and re-injected after thawing. Accuracy of sample identity was verified by repetitive finger printing using PCR of mitochondrial DNA (Hutter et al., 2004).

Cloning

The construct encoding for all 3 transgenes (Figure S1A) was generated by cloning a synthesized DNA-fragment (Eurofins Scientific, Luxembourg) encoding for mKate and a truncated form of the human nerve growth factor receptor lacking any intracellular signaling domain (NGFR; construct -mKateT2A-NGFR) into the pCDH-EF1α-extGLucT2A-copGFP Vector (Terziyska et al., 2012), leaving membrane anchored Gaussia luciferase and replacing copGFP gene using BamHI and SalI; T2A or P2A self-cleaving peptides enabled equimolar expression of the transgenes. For immunohistochemistry, cells were additionally transduced with a construct expressing mCherry which was obtained by amplifying mCherry from the pSicoR-U6-EF1a-mCherry Vector (addgene, Cambridge, MA, USA) and cloning it into the pCDH-EF1α-extGluc-T2A-copGFP Vector replacing the copGFP gene using BamHI and SalI.

Lentiviral transduction of ALL PDX cells and enrichment of transgenic cells

ALL-199 and ALL-265 were transduced using pCDH-EF1α-extGlucT2A-mKate-NGFR. Third generation packaging plasmids pMDLg/pRRE, pRSV-Rev and pMD2-G (Dull et al., 1998) were kindly provided by T. Schroeder. High-titer vesicular stomatitis virus (VSV) G protein-pseudotyped lentivector was prepared by transient 4-plasmid transfection of 293T cells using TurboFect Transfection Reagent (Thermo Scientific, Waltham, MA, USA) and supernatant concentration as described (Klier et al., 2008; Terziyska et al., 2012). The functional titer of virus was determined by transduction of NALM-6 B-ALL cell line cells with serial dilutions of the vector stock, followed by analysis of transgene positive cells using flow cytometry.

Generation of transgenic PDX cells was performed as previously described (Terziyska et al., 2012). In brief, PDX cells were transduced over night with lentivirus at MOI > 10 in the presence of 8 µg/ml polybrene. The next day, cells were washed thoroughly and injected into mice. After passaging, cells expressing the transgenes were enriched in two consecutive rounds by flow cytometry using FACSAriaIII (BD Biosciences) and gating on the red fluorochrome before cell re-amplification in mice. Although lentiviral transduction could in principle alter cells due to the transduction process or genomic integration, we could not detect adverse effects so far in comprehensive quality controls (Terziyska et al., 2012).

Bioluminescence in vivo imaging

For bioluminescence in vivo imaging mice were anesthetized with isoflurane and D-Luciferin (BIOMOL GmbH, Hamburg, Germany) dissolved in sterile PBS was used as substrate. Immediately after intravenous tail vein injection of 150 mg/kg of native D-Luciferin per mouse, mice were imaged for 30 seconds or up to 2 minutes using a field of view of 12.5 cm with binning 8, f/stop 1 and open filter setting using the IVIS Lumina II Imaging System (Perkin Elmer, MA, USA). The Living Image software 4.x (Perkin Elmer, MA, USA) was used for data acquisition and quantification of light emission using a scale with a minimum of 1.8×10^4 photons per second per cm^2 per solid angle of 1 steradian (sr) (Terziyska et al., 2012). Mice were considered positive for engraftment, if light emission by the entire mouse exceeded $5x10^5$ photons s^{-1} and positive signals were detected at typical sites at the lower extremities.

Reagents

For flow cytometry, analysis of NGFR, mKate, mCherry, BrdU, Annexin V, DAPI and PI was performed by flow cytometry, using BD LSRFortessa and BD FACSAriaIII (BD Biosciences, Heidelberg, Germany). The following antibodies were used: NGFR-PerCP-Cy5.5 (Biolegend, CA, USA), BrdU-APC, Annexin V-FITC detection kit (both

from BD Biosciences, Heidelberg, Germany). Mouse CD45-APC-Cy7 (Biolegend, San Diego, CA, USA) was used to exclude mouse cells.
BrdU incorporation was detected using the BrdU Flow Kit (BD Biosciences, Heidelberg, Germany). For analysis of cell viability, DAPI and/or PI were added to the cells at a concentration of 1µg/ml. All antibodies and reagents were used according to the manufacturer's instructions.
For chemotherapy treatments in vivo and ex vivo vincristine (VCR; Merck, Darmstadt, Germany), etoposide (ETO; Sigma Aldrich, St. Louis, USA), cyclophosphamide (Cyclo; Baxter, USA), epirubicine (EPI; Sigma Aldrich, St. Louis, USA), amsacrine (Amsa, Sigma Aldrich, St. Louis, USA), cytarabine (Ara-C; cell pharm GmbH, Bad Vilbel, Germany), daunorubicin (DAU; Sigma Aldrich, St. Louis, USA), mitoxantrone (MITO; Sigma Aldrich, St. Louis, USA) or doxorubicin (DOX, Sigma Aldrich, St. Louis, USA) were used.

Labeling of PDX cells with BrdU and CFSE
To label PDX cells with BrdU, donor mice were fed with BrdU (VWR, Radnor, PA, USA) during the 7 last days before cell isolation, at approximately 0.8 mg/kg/d BrdU using BrdU-containing drinking water. Freshly isolated PDX cells were labeled with CFDASE (Life Technologies, Carlsbad, CA, USA) according to manufacturer's instructions. Cells were washed and directly injected into recipient mice. The procedures resulted in both BrdU and CFSE positivity of well above 98% of PDX cells, as validated by flow cytometry. As PDX ALL cells are heterogeneous in size, loss of CFSE appears as continuum in flow cytometry devoid of the distinct peaks known from normal lymphocytes.

Enriching human PDX ALL cells from murine bone marrow
The aim was to isolate and enrich minute numbers of human PDX ALL cells out of a major excess of murine bone marrow cells. The procedure was designed according to published protocols for isolating normal mouse hematopoietic stem cells from murine bone marrow (Takizawa et al., 2011). Our studies concentrated on the first 3 weeks of ALL growth in mice, when low tumor burden is mainly restricted to bone marrow without major involvement of further organs (data not shown).
Isolation of bone marrow cells from mice
To collect as many bone marrow cells as possible from each mouse, the hip, femura, tibiae, spine and sternum were isolated and crushed in a porcelain mortar. The suspension was washed with cold PBS, filtered through a 70 µm cell strainer, washed again with PBS and re-suspended in cold PBS at $1x10^7$ cells/ml.

Step 1: Enriching NGFR expressing PDX cells from the bone marrow suspension
A first enrichment step consisted in magnetic cell separation (MACS) of NGFR-expressing PDX ALL cells from the entire mouse bone marrow isolated. 20 µl per $1x10^7$ cells of anti-human NGFR MicroBeads (Miltenyi Biotech, Bergisch Gladbach, Germany) were added to the isolated mouse bone marrow cell suspension and incubated 10 minutes at 4°C. A maximum of $2x10^8$ cells were loaded onto a LS column (Miltenyi Biotech, Bergisch Gladbach, Germany), prepared according to manufacturer's instructions. Cells were recovered from the column according to manufacturer's instructions and washed with PBS.

Step 2: Enriching and quantifying fluorochrome expressing PDX cells from NGFR-expressing cells
The second consecutive enrichment step consisted in flow cytometry enrichment of red fluorochrome expressing cells out of the cell suspension obtained after MACS enrichment. Cells obtained after MACS enrichment were stained with DAPI to exclude dead cells and with anti-muCD45-APC-Cy7 (anti-mouse CD45) to exclude murine hematopoietic cells. Cells were quantified and sorted using a BD FACSAriaIII (BD Biosciences, Heidelberg, Germany), gating (i) on the lymphocyte gate in forward/side scatter, (ii) the negative gate for both mouse CD45 and DAPI and ultimately (iii) the positive gate for the red fluorochrome.
Alternatively and to quality control for the MACS enrichment step, 10% of the entire population of bone marrow cells was directly analyzed by flow cytometry without prior MACS enrichment and with the identical staining procedure (Figure S1D). The disadvantage of this procedure lies in the prolonged periods of time required for flow cytometric cell enrichment disabling measuring more than 10% of all cells.

Enriching dormant cells (LRC) from human PDX ALL cells
Step 3: Separating PDX ALL cells into LRC and non-LRC
Separating PDX ALL cells into LRC and non-LRC was performed within the flow cytometry enrichment step described above (Step 2) by addition of a 4^{th} gating strategy. Additionally to gating on (i) the lymphocyte gate in

forward/side scatter, (ii) the negative gate for both mouse CD45 and DAPI and (iii) the positive gate for the red fluorochrome, gating (iv) on CFSE was used to discriminate LRC and non-LRC as shown in Figure 1D. To set gate 4, CFSE intensity was measured at day 3 after injection when major bleaching had stopped; maximum CFSE MFI was used to define start of any cell proliferation ("0 divisions"). Maximum CFSE MFI was divided by factor 2 to calculate CFSE bisections mimicking cell divisions. 7 CFSE MFI bisections or more were defined as entire loss of the CFSE signal characterizing non-LRC. The LRC gate was set to include all cells harboring high CFSE signal of below 3 bisections of the maximum CFSE MFI (Schillert et al., 2013) (Figure 1D). All further analyses were done and analyzed with the same instrument settings and gates as determined using the sample on day 3 sample of the experiment.

Ex vivo culture of PDX cells

PDX cells were cultured in RPMI medium supplemented with 20% FSC, 1% pen/strep, 1% gentamycin, 6 mg/l insulin, 3 mg/l transferrin, 4 µg/l selenium (ITS-G, Gibco, San Diego, CA, USA), 2 mM glutamine, 1 mM sodium pyruvate, 50 µM α-thioglycerol (Sigma-Aldrich, St. Louis, MO, USA).

Limiting dilution transplantation assay (LDTA)

For LDTAs, NSG mice were injected intravenously with different amounts of PDX cells from ALL-265 or ALL-199. Development of leukemia was monitored by bioluminescence in vivo imaging every 7 to 14 days after cell injection. LIC frequencies were determined according to Poisson statistics, using the ELDA software application (http://bioinf.wehi.edu.au/software/elda/) (Hu and Smyth, 2009).

Drug stimulation ex vivo

500 LRC and 500 or 100,000 non-LRC were cultured in 100 µl medium in 96-well plates, in cell concentrations of 5,000 cells/ml or 10^6 cells/ml. Cytotoxic drugs were added in triplicates at the clinically relevant concentrations described in each Figure legend. Cell death was measured after 48h by forward-side scatter and DAPI or propidium iodide staining in a flow cytometer. Specific cell death induced by each drug was calculated as follows: specific cell death = [(cell death(stimulated) – cell death(control) / (100 – cell death(control))] * 100.

For co-cultures, MS-5 cells stably expressing mtagBFP as blue fluorochrome were irradiated in suspension with 60 Gy and seeded at 10^4 per well in a 96 well plate; 700 - 1,900 freshly isolated PDX cells were incubated with and without feeder cells in 100 µl medium for 24-48h stimulated with conventional cytotoxic drugs at clinically relevant concentrations; entirely all cells of each well were removed using trypsin digestion and stained with propidium iodide; feeder cells were excluded by gating on non-blue-expressing cells independently from CFSE or propidium iodide staining; absolute numbers of living PDX cells were measured using forwardside scatter analysis and cell death was additionally measured by propidium iodide staining in flow cytometry.

In vivo treatment of mice

For treatment of LRC, NSG mice were injected i.v. with $1x10^7$ PDX cells. 7 days after cell injection, control animals received physiological salt solution i.p., while treatment group mice were injected with chemotherapeutic drugs as indicated in Figure legends. Mice were taken down 3 days later, bone marrow was collected, and PDX cells were isolated and analyzed for CFSE label retention. For calculation of relative drug effect on LRC compared to non-LRC (Figure 4D), first absolute number of control LRC or non-LRC were divided by the absolute number of treated LRC or non-LRC, respectively. In a second step, relative cell reduction in non-LRC was set to 100% and cell reduction in LRC was calculated relative to non-LRC. A maximum of 4 animals could be included into the same experiment as a maximum of 4 animals could be analyzed for CFSE distribution at the same day.

To obtain cells at minimal residual disease, $1x10^6$ ALL-199 or ALL-265 were injected into 19 NSG mice and leukemic growth was followed by weekly in vivo imaging. Treatment was started at an average of $1x10^{11}$ photons s^{-1}, when untreated cells were recovered from 5 mice. Mice were divided into different treatment groups which were treated as indicated in Figures legends.

Immunostaining of bone marrow sections

Mouse femurs were fixed in zinc formalin fixative for 1 day at 4°C. Bones were washed with PBS and decalcified with Osteosoft (Merck) for 3 days at 4°C, infiltrated with 30% sucrose for 1 day at 4°C, embedded in O.C.T. compound (Sakura) and stored at -80°C. Cryosections of decalcified bones were obtained by using the CryoJane tape transfer system (Leica). For immunostaining, sections were permeabilized and blocked with 5% goat serum and 0.1% Tween-20 serum in PBS for 45 min at room temperature. Primary antibodies were applied for 1 day at 4°C and

followed by secondary antibody incubation for 45 min at room temperature. Sections were finally stained with 10 mg/ml DAPI for 15 min and the slides were mounted with prolong gold (Invitrogen). Washing in between each staining steps was performed. Primary antibodies were rabbit-anti-FITC (ThermoFisher; 1:100) and rabbit-anti-mCherry (Abcam; 1:100) and goat-anti-rabbit with Alexa 594 (Invitrogen) was used as secondary antibody. Images were acquired on a Leica SP5 confocal microscope and analyzed with ImageJ. CFSE signal intensity was adapted to the mCherry signal by adjusting the 8 bit threshold for quantification of the LRC population based on FACS data. The endosteal region was defined as less than 100 µm from bone matrix (Nombela-Arrieta et al., 2013). Cells of interests were counted semi-automatically by the program ImageJ. Relative endosteal cells were calculated as absolute cell numbers in the endosteal region divided by absolute cell numbers in entire bone marrow section. Mean and standard error were calculated from at least 3 sections of each femur from 2 independent mice.
For immunhistology of primary bone marrow biopsies, bone marrow biopsies were fixed and stained using the avidin-biotin-peroxidase complex (ABC) method (Hsu et al., 1981) and anti-TdT antibody (Leica, Germany) and anti-Ki-67 antibody (Dako, Germany).

Flow cytometric cell enrichment of StemB cells from BCR-ABL positive ALL
Thawed mononuclear bone marrow cells were handled on ice and stained with CD3-FITC, CD19-PE, CD34-APC, CD38-PECy7 (all Becton Dickinson) and DAPI 0.1 µg/ml; StemB cells expressing $CD3^-$ $CD34^+$ $CD38^{-/low}$ $CD19^+$ cells were enriched using the FACSAria™ (Becton Dickinson) according to (Castor et al., 2005; Hong et al., 2008; Lutz et al., 2013).

Flow cytometric cell enrichment of diagnostic and MRD pediatric BCP-ALL cells
Thawed mononuclear bone marrow cells were handled on ice and stained using antibodies appropriate for minimal residual disease (MRD) detection against CD10, CD19, CD20, CD34, CD38, CD45, CD58, CD99, and CD123. Leukemic blasts were enriched to >95% purity using a FACSAria™ Fusion cell sorter equipped with an automatic cell deposition unit (ACDU; Becton Dickinson); data analysis was performed using the FACSDiva™ software (Becton Dickinson).

Bulk RNA sequencing library construction
PDX LRC and non-LRC cell populations were sorted into lysis buffer composed of 0.2 % Triton X-100 (Sigma) and 2 U/µl of RNAse Inhibitor (Life Technologies). ERCC spike-in controls (Life Technologies) were added to the cell lysis mix at 1:5,000 dilution. RNA was cleaned-up from the crude lysate with Agencourt RNAclean XP SPRI beads (Beckman-Coulter). cDNA was synthesized and pre-amplified from 5 µl of lysate according to the Smart-seq2 protocol (Picelli et al., 2013).
For each pediatric ALL MRD and PDX MRD sample, 2000 cells were sorted into TCL buffer (Qiagen). RNA was cleaned up using Agencourt RNAclean XP SPRI beads from half of the lysate and used to generate UMI-seq libraries as previously described (Parekh et al., 2016).
For all libraries, 1 ng of pre-amplified cDNA was used as input for tagmentation by the Nextera XT Sample Preparation Kit (Illumina), where a second amplification round was performed for 12 cycles.

RNA sequencing library construction of primary StemB single cells
Single adult StemB cells were deposited in 96-well plates containing 5 µl lysis buffer composed of a 1:500 dilution of Phusion HF buffer (NEB). Single-cell RNA-seq libraries were constructed using the SCRB-seq method according to (Soumillon et al., 2014).

RNA-seq analysis
All sequencing reads were demultiplexed from the Nextera (i5 and i7) indices.
For Smart-seq libraries, demultiplexed reads were aligned to the human genome (hg19) and ERCC reference using NextGenMap (Sedlazeck et al., 2013). Count data was generated from mapped reads using featureCounts (Liao et al., 2014) on ENSEMBL gene models (GRCh38.74).
For UMI-seq and SCRB-seq libraries, read pairs were processed by tagging the cDNA read with barcode and UMI sequences using the Drop-seq tools pipeline (Macosko et al., 2015). Tagged reads were aligned to the human genome (hg19) using STAR (Dobin et al., 2013) and sample-wise count tables generated using Drop-seq tools.

To remove noise from lowly expressed genes, count data sets were subjected to data-driven gene filtering using the HTSFilter R package (Rau et al., 2013). For PDX single cell sequencing libraries, only those cell data sets were used

which came from viable cells, obtained at least 1 million reads and detected at least 3000 genes (TPM > 1). For combined bulk (1x ALL-265; 4x ALL-265) and single cell (1x ALL-265) analysis (Figure 5), filtered single cell datasets were included summarized by gene-wise median read count as one LRC and non-LRC replicate. Differential expression (DE) analysis was done in the DESeq2 R package (Love et al., 2014) using the Wald test.
A combined LRC signature (ALL-265 & ALL-199; 250 genes; FC > 1; padj < 0.05) was obtained from this data.
Overrepresentation of significantly differentially expressed genes in KEGG pathways was tested by a fixed network enrichment analysis (FNEA) implemented in the neaGUI R package (Alexeyenko et al., 2012).
We applied hierarchical clustering gene-wise and sample-wise with complete linkage based on Euclidian distances of variance stabilized counts of DE genes (500 genes with lowest padj, FDR adjustment (Benjamini and Hochberg, 1995)) and plotted as heatmap. The reference expression value is the expression average of non-LRC cells.
Principal Component Analysis (PCA) of LRC PDX cells was performed on variance stabilized counts of the 500 most variable genes to display the main variance of the samples.
To analyse combined data from all obtained single-cells, count data was normalized accounting for batch effects using SCONE (Risso et al., 2014). PCA and k-means clustering of combined single-cell data was performed on all shared detected genes.
Gene set enrichment analysis was performed using GSEA Desktop Application. For ranking all genes, a metric score was calculated by multiplying their log fold changes with the –log10(p_adj) values and submitted to the Pre-Ranked GSEA tool. The statistical significance was determined by 1000 gene set per mutations (Subramanian et al., 2005).

Dynamical modelling
The growth behavior of ALL cells in bone marrow has been analyzed using mechanistic ordinary differential equation models describing the population growth and the CFSE dilution. To gain insights into the in vivo growth behavior of ALL cells, we compared three alternative models. The first model assumed exponential growth, the second model assumed logistic growth caused by a decreased rate of cell division at higher cell densities, and the third model assumed logistic growth caused by an increased rate of cell death at higher cell densities.
The state variables of all three models are the cell number $n(t)$ and the mean fluorescence intensity $m(t)$. The governing equations for $n(t)$ and $m(t)$,

$$dn/dt = (\alpha(n)-\beta(n))\, n, \qquad N(0) = n_0,$$
$$dm/dt = -(\alpha(n)+k)\, m, \qquad m(0) = m_0,$$

have been deduced from existing partial differential equation models (Hasenauer et al., 2012). In this governing equations $\alpha(n)$ denotes the rate of cell division, $\beta(n)$ denotes the rate of cell death, and k denotes the rate of CFSE degradation.
The three model alternatives only differed in the parameterization of rates of cell division and cell death, $\alpha(n)$ and $\beta(n)$. For the exponential growth model all rates were constant, $\alpha(n) = \alpha_0$ and $\beta(n) = \beta_0$. For the logistic growth model with decreasing cell division at higher cell densities $\alpha(n) = \alpha_0 (1-n/n_\alpha)$ and constant $\beta(n) = \beta_0$ were used. For the logistic growth model with increasing cell death at higher cell densities constant $\alpha(n) = \alpha_0$ and $\beta(n) = \beta_0 (1+n/n_\beta)$ were used. As the measurement of the mean intensity induced by CFSE is corrupted by the cell's autofluorescence, we measure $m'(t) = m + m_a$, in which m_a denotes the average autofluorescence.
The parameters of the three models were determined from measurement of $n(t)$ and $m'(t)$ using maximum likelihood estimation, assuming normally distributed measurement noise. For model comparison the Akaike information criterion (AIC) was used.

Statistics
All statistical analyses were calculated using GraphPad Prism 6 software. Two-tailed unpaired t-test was applied to evaluate differences after drug treatment. F-test was applied to compare standard deviations; in cases, when standard deviations differed significantly, Welch's correction was applied. LIC frequencies were calculated according to Poisson statistics using the ELDA software application (http://bioinf.wehi.edu.au/software/elda) (Hu and Smyth, 2009).

Supplemental References

Alexeyenko, A., Lee, W., Pernemalm, M., Guegan, J., Dessen, P., Lazar, V., Lehtio, J., and Pawitan, Y. (2012). Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinformatics *13*, 226.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B, Methodological *57*, 289-300.

Castor, A., Nilsson, L., Astrand-Grundstrom, I., Buitenhuis, M., Ramirez, C., Anderson, K., Strombeck, B., Garwicz, S., Bekassy, A. N., Schmiegelow, K.*, et al.* (2005). Distinct patterns of hematopoietic stem cell involvement in acute lymphoblastic leukemia. Nat Med *11*, 630-637.

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.

Dull, T., Zufferey, R., Kelly, M., Mandel, R. J., Nguyen, M., Trono, D., and Naldini, L. (1998). A Third-Generation Lentivirus Vector with a Conditional Packaging System. Journal of Virology *72*, 8463-8471.

Hasenauer, J., Schittler, D., and Allgöwer, F. (2012). Analysis and Simulation of Division- and Label-Structured Population Models. Bull Math Biol *74*, 2692-2732.

Hsu, S. M., Raine, L., and Fanger, H. (1981). Use of avidin-biotin-peroxidase complex (ABC) in immunoperoxidase techniques: a comparison between ABC and unlabeled antibody (PAP) procedures. J Histochem Cytochem *29*, 577-580.

Hu, Y., and Smyth, G. K. (2009). ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. J Immunol Methods *347*, 70-78.

Hutter, G., Nickenig, C., Garritsen, H., Hellenkamp, F., Hoerning, A., Hiddemann, W., and Dreyling, M. (2004). Use of polymorphisms in the noncoding region of the human mitochondrial genome to identify potential contamination of human leukemia-lymphoma cell lines. Hematol J *5*, 61-68.

Klier, M., Anastasov, N., Hermann, A., Meindl, T., Angermeier, D., Raffeld, M., Fend, F., and Quintanilla-Martinez, L. (2008). Specific lentiviral shRNA-mediated knockdown of cyclin D1 in mantle cell lymphoma has minimal effects on cell survival and reveals a regulatory circuit with cyclin D2. Leukemia *22*, 2097-2105.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

Liem, N. L. M., Papa, R. A., Milross, C. G., Schmid, M. A., Tajbakhsh, M., Choi, S., Ramirez, C. D., Rice, A. M., Haber, M., Norris, M. D.*, et al.* (2004). Characterization of childhood acute lymphoblastic leukemia xenograft models for the preclinical evaluation of new therapies. Blood *103*, 3905-3914.
Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

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.

Nombela-Arrieta, C., Pivarnik, G., Winkel, B., Canty, K. J., Harley, B., Mahoney, J. E., Park, S. Y., Lu, J., Protopopov, A., and Silberstein, L. E. (2013). Quantitative imaging of haematopoietic stem and progenitor cell localization and hypoxic status in the bone marrow microenvironment. Nat Cell Biol *15*, 533-543.

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

Picelli, S., Bjorklund, A. 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.

Rau, A., Gallopin, M., Celeux, G., and Jaffrezic, F. (2013). Data-based filtering for replicated high-throughput transcriptome sequencing experiments. Bioinformatics *29*, 2146-2152.

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.

Sedlazeck, F. J., Rescheneder, P., and von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics *29*, 2790-2791.

Soumillon, M., Cacchiarelli, D., Semrau, S., Oudenaarden, A. v., and Mikkelsen, T. S. (2014). Characterization of directed differentiation by high-throughput single-cell RNA-Seq. bioRxiv.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Discussion

Whole-transcriptome amplification introduces noise

Single-cell RNA sequencing is a novel tool with a multitude of applications in biology and biomedical sciences (Wagner et al. 2016). Determining the global state of gene expression in a given cell gives a rich and powerful view of a cell's identity. Furthermore, scRNA-seq reveals implications of molecular circuitry within cells for development, fate decisions or disease processes. Being an emerging technology, there are a large number of newly developed experimental and computational strategies. Importantly for further work, power, biases and limitations of scRNA-seq approaches need to be better understood.

In order to access the small quantities of mRNA contained in single cells, whole transcriptome amplification prior to library construction and sequencing is a necessity. Investigating how amplification affects data quality, we analysed bulk and single-cell datasets from three protocols: Smart-seq, TruSeq and UMI-seq (Parekh et al. 2016).

First, we analysed whether duplicates stemming from amplification can be identified computationally. Commonly, duplicates are identified by their mapping position, e.g. prior to SNP-calling (DePristo et al. 2011). Comparing rates of duplication for single-end and paired-end data, we found that single-end duplicates are much more common but can be easily explained by a sampling model. Depending on the expression level of a gene, there is a certain probability for the same mapping position being sampled from different RNA-molecules. Furthermore, we identified that fragmentation is biased and fragmentation patterns are highly reproducible within library preparation methods. Thus, considering mapping position for the removal of duplicates disqualifies because not every position of the transcript has an equal chance to be a read start position. Furthermore, methods like Smart-seq or SCRB-seq perform fragmentation only after pre-amplification of full-length cDNA. Thus, only the duplicates arising from the subsequent library PCR could have been found by looking at read start

positions. Confirming this notion, power simulations using the PROPER framework (Wu et al. 2015) showed a reduction in power for simulated RNA sequencing data when removing duplicates.

The only instance where removing duplicates increased power was the UMI-seq dataset. Importantly, here we removed duplicates on the basis of the UMI instead of considering read mapping positions. Interestingly, the UMI-seq dataset showed ~65% duplicates, although it was constructed from a bulk of 10 ng total RNA. It can be expected that duplication fractions would be even higher from the small starting RNA amount of a single-cell.

Indeed, our paper comparing various single-cell RNA sequencing methods (Ziegenhain et al. 2017) showed that, depending on the protocol used, as much as 98% of reads per cell are amplification duplicates when analysing UMI libraries sequenced to one million reads.

In order to provide descriptive statistics on amplification noise, we leveraged the fact that the drawing of read counts can be described as a poisson process. Thus, under the poisson distribution, the variation of counts can be predicted from the mean. Any variation above the poisson expectation, such as variation stemming from amplification noise can be expressed as Extra-Poisson variability. We show that methods containing UMIs show lower Extra-Poisson variability than methods without UMIs, or when disregarding UMI information. This illustrates how UMIs are indeed capable of removing amplification noise. When performing power simulations on single-cell data containing UMIs, the increase in power by removing duplicates is much more pronounced than for the UMI-seq bulk dataset, indicating that larger amounts of amplifications lead to more amplification noise. Similarly, within the comparison of single-cell RNA sequencing methods, less sensitive methods generally showed less power than more sensitive methods, when not considering UMI information. Comparing several strategies for whole-transcriptome amplification, IVT (linear amplification) and PCR (exponential amplification), one can clearly see that amplification noise is more pronounced in PCR-based methods, as biases can propagate exponentially (Ziegenhain et al. 2017).

In conclusion, it is recommended to eliminate amplification noise by the inclusion of UMIs when performing whole-transcriptome amplification, especially in single-cell RNA sequencing applications, to avoid a loss in power.

Technical performance of scRNA-seq methods varies widely and can be improved

High sensitivity is a necessity to obtain sequencing data from the limited quantities of RNA of a single cell. Because of this, a large focus of scRNA-seq method development has been on improving sensitivity. Still, single-cell RNA sequencing methods are significantly less sensitive than single molecule fluorescence in-situ hybridisation (smFISH) (Raj et al. 2008; Torre et al. 2017). Reverse transcription is regarded as the a major limiting step for sensitivity (Picelli et al. 2013) and the absolute efficiency of scRNA-seq methods to detect a given RNA molecule in a cell is estimated to be between 12% (determined by comparison to smFISH) and 48% (determined by ERCC spike-in UMI counts) (Grün et al. 2014; Islam et al. 2014). In this work, we could for the first time directly compare relative sensitivities of six major single-cell RNA sequencing protocols because we generated data from the same cell type and culture condition. Using the number of detected genes per cell as a proxy, we show that Smart-seq2 is the most sensitive protocol, outperforming the microfluidics-based Smart-seq/C1 method. Among UMI-based methods, SCRB-seq and CEL-seq2/C1 are significantly more sensitive than MARS-seq and Drop-seq. In our comparative analysis, all cells were sequenced and downsampled to exactly one million reads to exclude influence of varying sequencing depths. In many practical applications, however, it will not be economical to sequence cells to saturation. Thus, library complexity at low sequencing coverage and how quickly saturation of libraries is reached are important parameters. We show that the slope of gene detection (sensitivity) relative to the number of sequenced reads is variable between protocols.

Importantly, this slope is largely dependent on amplification bias where methods with in-vitro amplification feature a steeper slope and thus contain more of their total information at lower sequencing coverages. Moreover, there are also clear differences among PCR-based methods that correlate with difference in amplification bias measured as Extra-Poisson Variability. Hence, although amplification bias can be removed by unique-molecular identifiers, low bias is crucial for large information content at low sequencing coverage and hence crucial for cost-efficiency.

Based on insights gained from this comparison, we set out to systematically optimize the sensitivity and amplification bias of the already efficient SCRB-seq method (Soumillon et al. 2014). Counterintuitively, our data indicates that the systematic optimizations applied to the sensitive Smart-seq2 protocol (Picelli et al. 2013) can not be directly transferred to SCRB-seq, although both methods employ PCR amplification of full-length cDNA after reverse-transcriptase template switching. Thus, there may be complex interactions of enzymes, buffers and additives that can lead to method-specific increases in sensitivity. In our case we find that adding polyethylene glycol greatly increases cDNA yield and sensitivity, probably due to molecular crowding leading to higher chances of interactions between reverse transcriptase and mRNA molecules in the crowded environment. Another important improvement was the use of a different PCR polymerase that generates less amplification bias and hence increases the efficiency. Together with several more minor modifications, we established “molecular crowding SCRB-seq” (mcSCRB-seq) (Bagnoli et al., 2017). In order to show the relative increase in sensitivity to the original SCRB-seq protocol, we again generated comparable scRNA-seq data from the same cells in the same batch. Thus, we can quantify that the mcSCRB-seq protocol detects 2.5x more unique RNA molecules than SCRB-seq, which represent a large fraction of the cellular transcriptome. In other studies lacking comparative data, ERCC spike-ins are often used to experimentally determine sensitivity (Islam et al. 2014; Liu et al. 2016; Genshaft et al. 2016). ERCCs consist of 92 synthetic poly-adenylated mRNA

transcripts divided into four groups of 23 transcripts each, represented at known concentrations spanning a 10^6-fold range (Baker et al. 2005; Jiang et al. 2011). To not only compare the mcSCRB-seq protocol to the SCRB-seq protocol, we modeled detection of ERCC transcripts in relation to their concentration using a binomial logistic regression, as proposed by others (Marinov et al. 2014; Svensson et al. 2017). This enables us to integrate the results of another recently published study comparing technical performance of various scRNA-seq protocols based on the analysis of ERCC spike-ins (Svensson et al. 2017). *Svensson et al.* focussed only on spike-ins to try to integrate large amounts of public data from 15 protocols generated from different cell types in many labs and varying sequencing depth. Although of broad relevance, this large collection of datasets may also contain pitfalls. For instance, the data shown for Smart-seq2 varied in three discrete populations over a 10^3 detection limit range, showing that in this case other underlying factors have to drive the variance observed, not the protocol itself.

Indeed, ERCC spike-ins have recently been criticized for various reasons as they may be subject to different technical effects stemming from library preparation than endogenous genes in certain conditions (Risso et al. 2014). Furthermore, physical properties, such as length (range 250 to 2000 nucleotides), GC content (5–51%), short poly-A tails and lack of 5' cap structure do not model endogenous transcripts perfectly (Grün et al. 2014; Stegle et al. 2015; Svensson et al. 2017). Still, *Svensson et al.* generally confirm our results in the data generated from their own laboratory, where Smart-seq2 was the most sensitive method, while a high-throughput droplet method, in their case 10x Genomics (Zheng et al. 2017), was the least sensitive. When analyzing ERCC spike-in data generated using the mcSCRB-seq protocol, we can show that it is now the most sensitive protocol available.

In conclusion, high sensitivity relative to sequencing depth is a decisive factor for single-cell RNA sequencing in order to extract as much information per cell as possible, but varies widely among methods. Currently, the extensively optimized Smart-seq2 protocol is the most

sensitive full-length method available, while mcSCRB-seq is the most sensitive method featuring unique molecular identifiers.

Power simulations inform scRNA-seq studies

In practice, the power to detect differential gene expression is at the heart of many analyses of RNA sequencing data, but it is not intuitive how the empirical parameters discussed so far, eg. sensitivity and amplification noise, effect this power. In order to be able to judge power, we developed a simulation framework for single-cell RNA sequencing data. Previously reported power simulation tools for RNA sequencing (Poplawski & Binder 2017) suffer from limitations regarding the integrated analysis tools (e.g. PROPER (Wu et al. 2015)), consideration of sequencing depth (e.g. RNASeqPower Calculator (Ching et al. 2014)) or the input of user-specified pilot data (e.g. RSPS). Moreover, none of the previously published frameworks is able to work with the special characteristics of single-cell RNA sequencing data. Especially for sparse data (e.g. droplet-based methods with low sensitivity), we found that it is necessary to consider a dropout rate (Kharchenko et al. 2014). The dropout rate (p_0) describes the chance of missing a gene expression value for a certain gene in any given cell due to technical limitations of scRNA-seq protocols (Bacher & Kendziorski 2016). Thus, our power simulations integrate the modelling of mean-variance and mean-dropout relationship to reliably recapitulate characteristics of scRNA-seq count data (Ziegenhain et al. 2017; Vieth et al. 2017). We furthermore overcome previous limitations by implementing twelve tools for differential gene expression analysis, consideration of sequencing depth (library size) factors and easy user-defined input of empirical pilot data and fold-change distributions.

Power analyses have high importance to inform researchers in several stages of a given study. First, in order to obtain optimal experimental design, assessment of expected power is helpful. Here, statistical power should be considered together with sample material limitations and financial constraints, given the scRNA-seq technology of choice. By this, the practically feasible

number of replicates and sequencing depth can be determined along with an expectation of how much of the differences between analysed populations may be found.

Even after experiments have been conducted, powsimR can help in *a posteriori* power evaluation. Thus, researchers can get an estimate on how many of the differences between populations have been found and quantify whether major effects may have been missed.

Furthermore, power simulations are valuable for computational and experimental method development, which is dynamically ongoing in the field of single-cell transcriptomics.

Indeed, powsimR has already been used by others to conduct a comparison of differential testing algorithms (Soneson & Robinson 2017). In line with our findings, the authors find that, after pre-filtering of lowly/stochastically detected genes (Lun et al. 2016), limma/voom (Ritchie et al. 2015) and MAST (Finak et al. 2015) have high sensitivity to call truly differentially expressed genes while controlling FDR appropriately. Furthermore, *Soneson et al.* characterise biases and computational time requirements for the various algorithms, which has implications for the application to high-throughput scRNA-seq.

In addition to analysis of differential gene expression testing algorithms, we have applied an early version of *powsimR* to our comparative data generated from six important single-cell RNA sequencing methods (Ziegenhain et al. 2017). Assuming that all protocols were powerful enough to detect big effects already with rather small sample sizes, we chose to draw empirical fold-changes from the more moderate differences of two microglial subpopulations described by *Zeisel et al.* (Zeisel et al. 2015). Indeed, at small sample sizes (16 cells per group) power to detect differential gene expression was generally low but increased with larger sample sizes. Of all scRNA-seq protocols present in the comparison, SCRB-seq needed the smallest sample size (64 cells per group) to reach 80% power. Likely, this high power is owed to the favorable combination of fairly high sensitivity and low noise due to the use of UMIs. Additionally, *powsimR* allowed us to investigate power in relation to sequencing depth. Overall, lowering sequencing to 500,000 or 250,000 reads per cell reduced power modestly. However, we could

observe that methods featuring in-vitro amplification (ie. CEL-seq2 and MARS-seq) were less affected by downsampling, because the lack of PCR bias leads to better sampling of the transcriptome at lower coverages. Clearly, IVT-based methods are at an advantage for lower coverage sequencing, but there are also significant differences in the amount of bias between PCR-based methods, as discussed above. Thus, optimizing protocols towards uniform amplification is an important factor for maximizing the information obtained per number of reads (Sasagawa et al. 2017, Bagnoli et al., 2017).

Determining the optimal balance between replication and sequencing depth for a given experiment will thus depend on availability of samples or whether lowly expressed genes are of interest. It should be highlighted that a thorough assessment of power for high throughput transcriptomic profiling of single-cells (eg > 100,000 cells) (Zheng et al. 2017) with more and more sparse sequencing depths (as low as 10,000 - 20,000) reads per cell has not been done yet. This would constitute a particularly important and timely contribution to the field, as the sequencing costs are substantial at this scale. For instance, doubling the sequencing depth of 100,000 cells from 50,000 to 100,000 reads, which is still fairly low by conventional scRNA-seq standards, would cost approximately 18,000 € using the popular NextSeq550 sequencer.

Another important area of development in the field of scRNA-seq concerns the requirement for fresh live cells. Single cell RNA sequencing is currently being used in many applications from diverse sources, but fresh cells are not always available due to practical, logistical or ethical constraints - especially in studies involving human tissue.

Overcoming this limitation, cryopreservation strategies can be applied (Guillaumet-Adkins et al. 2017). In this study, the authors showed that cryopreservation of cell lines and tissues yielded good single-cell RNA-seq data and did not alter transcriptional profiles of cells using both full-length and a 3'-counting scRNA-seq methods. Important technical performance parameters, such as number of detected genes (i.e. sensitivity) and genebody coverage (i.e.

transcript degradation) were indistinguishable from scRNA-seq data of fresh cells. Furthermore, cells always clustered according to cell type instead of by freezing/thawing condition, confirming that transcriptomes were largely unchanged.

For more short-term preservation of primary tissues, whole tissues can be stored at 4 °C for several days in organ transplant solution (Wang et al. 2017). This enables to process preserved tissues by dissociation and single-cell isolation for RNA-seq in the same way as fresh tissue. Similar to *Guillaumet-Adkins et al.*, the authors characterized technical parameters such as successful cDNA synthesis, sensitivity and transcript degradation at several timepoints of storage. Storage of tissues for up to three days was possible with only minor decrease in technical performance. However, this study was only performed with one single-cell RNA-seq protocol, limiting the broad applicability of the findings.

Both of the discussed preservation techniques rely on the conservation of live, intact single-cells that can be processed similar to freshly obtained cells. Additionally, several other approaches use cell fixation to decouple sampling and library preparation. For instance, formaldehyde fixation has been used in conjunction with downstream reverse-crosslinking and library preparation in a plate-based method (Thomsen et al. 2016). Furthermore, methanol fixation has been demonstrated to be compatible with droplet-based single-cell RNA sequencing (Alles et al. 2017). Another innovative method, SPLiT-seq, relies on formaldehyde fixation and leverages the cell itself as the vessel for in-situ reverse transcription (Rosenberg et al. 2017). Together with combinatorial barcoding, this method is able to process large numbers of fixed cells. Thus, SPLiT-seq can be used to access precious fixed material, for instance from human brain biobanks.

In summary, single-cell RNA sequencing technology is rapidly evolving, becoming more and more widely applicable. Additionally, there are considerable further steps to be made. The presented comparative data and the *powsimR* simulation framework constitute an excellent resource for benchmarking ongoing developments. It will become increasingly important in the

future to judge the power of the quickly growing number of scRNA-seq protocols available to chose the most appropriate methods.

scRNA-seq enables characterization of rare leukemia cells

Acute lymphoblastic leukemia (ALL) is of high relevance, as it is the most common pediatric malignancy. Although modern chemotherapy treatments are able to cure most patients, further research needs to be done to prevent relapse with adverse outcome. Research on leukemia cells is challenging because it is not possible to culture patient cells *in vitro*, likely due to a lack of microenvironment-dependent factors (Vick et al. 2015). Although short-term culture models with various cytokine factors and immortalized cell lines exist, these are often disadvantageous because of alterations in important functional characteristics occur frequently (Pan et al. 2009). Furthermore, primary patient material is rare and can obviously not be studied in the absence of treatments. To overcome these limitations, our study leverages patient-derived xenograft models (Kamel-Reid et al. 1989), i.e. primary patient cells transplanted in immunocompromised mice. This allows us to obtain and perturb leukemia cells over time. Crucially, we can track cancer cells by the integration of transgenes or stainings. In patients, relapse with dismal outcome is one of the biggest reason for mortality (van Dongen et al. 2015). So far, conflicting evidence on the mechanism of relapses exists. On the one hand, some studies suggest that genetically heterogeneous and therapy-resistant subclones may exist at onset of disease that populate relapse (Irving et al. 2014), yet, there is currently no evidence that a majority of ALL incidences are genetically complex (Pal et al. 2016). On the other hand, microenvironment-mediated resistance is one of the discussed mechanisms for persistence of ALL cells (Meads et al. 2009). Here, it is thought that the close communication between ALL cells and bone marrow cells leads to expression of genes conferring increased resistance to treatment (Polak et al. 2015).

In order to better understand the biology of rare persistent and relapse-inducing cells, we generated the first single-cell RNA sequencing data from acute lymphoblastic leukemia (ALL) cells. Leveraging our patient-derived xenograft model, we could identify and sequence RNA from a rare, persistent cell population at minimal residual disease. We can show that these cells feature greatly downregulated cell-cycle genes and are thus dormant. Furthermore, persistent cells highly express a number of cell adhesion factors, indicating they must be located in the hematopoietic niche. Importantly, functional characterization showed that these cells were indeed treatment-resistant. In conclusion, our findings confirm the notion that blasts can modulate the microenvironment to form a protective niche (Duan et al. 2014). Additionally, we show that this niche-association is strongly linked to dormancy. Thus, persistence and treatment-resistance seem to consist of two joint mechanisms: niche-adherence and dormancy. Furthermore, we present evidence that there is plasticity between the cycling blast phenotype and the resting persistent phenotype, which dispels the notion that a fixed genetically distinct population of cells is responsible for driving relapse.

In conclusion, leveraging the power of single-cell RNA sequencing has allowed us to access and characterize this rare and clinically important cell population, furthering the understanding of how ALL cells persist to cause relapse.

Conclusion and Outlook

In this work focussing on single-cell RNA sequencing, I investigated the current state of technology, developed method improvements and applied it to a relevant biomedical question. Because of the dynamic nature of the field and the growing interest in the technique, scRNA-seq has matured in few years from an expensive niche method to a key approach in diverse fields of biology. Still, the technology has converged neither for wet lab protocols nor for computational analysis tools. Here, we made contributions to both aspects of method development by introducing our own highly sensitive molecule-counting protocol *mcSCRB-seq*, our fast and flexible data processing pipeline *zUMIs* and our statistical power analysis framework *powsimR*. We aim to provide easily implemented tools that make it possible for more molecular biology labs to start using the power of single-cell transcriptomics and enable its application to more and more research questions as we did to elucidate persistence and dormancy in ALL.

Today, single-cell analyses have generated enough buzz to spark the Human Cell Atlas initiative, an international consortium with the ambitious goal of mapping all cells of the human body (Regev et al. 2017). From this initiative that is driven by the development of novel wet-lab technologies and computational analysis techniques, we will learn more about what defines cell types, states and their transitions. Furthermore, we will be able to better distinguish diseased cells from healthy ones, determine molecular signatures for diagnosis and even discover targets for therapeutic intervention.

Such a comprehensive single-cell molecular map of the human body is unprecedented and will be a truly revolutionary resource for researchers of virtually all areas in biology and medicine.

In conclusion, single-cell RNA sequencing will continue to revolutionize the way we see and understand the life's building blocks: the cell.

References

Adams, M.D. et al., 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013), pp.1651–1656. Available at: https://www.ncbi.nlm.nih.gov/pubmed/2047873.

Adey, A. et al., 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology*, 11(12), p.R119. Available at: http://dx.doi.org/10.1186/gb-2010-11-12-r119.

Aird, D. et al., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology*, 12(2), p.R18. Available at: http://dx.doi.org/10.1186/gb-2011-12-2-r18.

Alles, J. et al., 2017. Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC biology*, 15(1), p.44. Available at: http://dx.doi.org/10.1186/s12915-017-0383-5.

Alwine, J.C., Kemp, D.J. & Stark, G.R., 1977. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5350–5354. Available at: https://www.ncbi.nlm.nih.gov/pubmed/414220.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Auer, P.L. & Doerge, R.W., 2010. Statistical design and analysis of RNA sequencing data. *Genetics*, 185(2), pp.405–416. Available at: http://dx.doi.org/10.1534/genetics.110.114983.

Avery, O.T., Macleod, C.M. & McCarty, M., 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *The Journal of experimental medicine*, 79(2), pp.137–158. Available at: https://www.ncbi.nlm.nih.gov/pubmed/19871359.

Bacher, R. & Kendziorski, C., 2016. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome biology*, 17(1), p.63. Available at: http://dx.doi.org/10.1186/s13059-016-0927-y.

Bagnoli, J.W. et al., 2017. mcSCRB-seq: sensitive and powerful single-cell RNA sequencing. Unsubmitted Manuscript.

Baker, S.C. et al., 2005. The External RNA Controls Consortium: a progress report. *Nature methods*, 2(10), pp.731–734. Available at: http://dx.doi.org/10.1038/nmeth1005-731.

Ballouz, S., Verleyen, W. & Gillis, J., 2015. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* , 31(13), pp.2123–2130. Available at: http://dx.doi.org/10.1093/bioinformatics/btv118.

Baron, M. et al., 2016. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell systems*, 3(4), pp.346–360.e4.

Available at: http://dx.doi.org/10.1016/j.cels.2016.08.011.

Baruzzo, G. et al., 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature methods*, 14(2), pp.135–139. Available at: http://dx.doi.org/10.1038/nmeth.4106.

Bayerlová, M. et al., 2015. Comparative study on gene set and pathway topology-based enrichment methods. *BMC bioinformatics*, 16, p.334. Available at: http://dx.doi.org/10.1186/s12859-015-0751-5.

Becker-André, M. & Hahlbrock, K., 1989. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). *Nucleic acids research*, 17(22), pp.9437–9446. Available at: https://www.ncbi.nlm.nih.gov/pubmed/2479917.

Bentley, D.R. et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–59. Available at: http://dx.doi.org/10.1038/nature07517.

Bhargava, V. et al., 2014. Technical variations in low-input RNA-seq methodologies. *Scientific reports*, 4, p.3678. Available at: http://dx.doi.org/10.1038/srep03678.

Biase, F.H., Cao, X. & Zhong, S., 2014. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome research*, 24(11), pp.1787–1796. Available at: http://dx.doi.org/10.1101/gr.177725.114.

Blanco, L. et al., 1989. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *The Journal of biological chemistry*, 264(15), pp.8935–8940. Available at: https://www.ncbi.nlm.nih.gov/pubmed/2498321.

Brennecke, P. et al., 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods*, 10(11), pp.1093–1095. Available at: http://dx.doi.org/10.1038/nmeth.2645.

van den Brink, S.C. et al., 2017. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature methods*, 14(10), pp.935–936. Available at: http://dx.doi.org/10.1038/nmeth.4437.

Butler, A. & Satija, R., 2017. Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. *bioRxiv*, p.164889. Available at: https://www.biorxiv.org/content/early/2017/07/18/164889.abstract [Accessed October 3, 2017].

Castro Alves, C. et al., 2012. Leukemia-initiating cells of patient-derived acute lymphoblastic leukemia xenografts are sensitive toward TRAIL. *Blood*, 119(18), pp.4224–4227. Available at: http://dx.doi.org/10.1182/blood-2011-08-370114.

Ching, T., Huang, S. & Garmire, L.X., 2014. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* , 20(11), pp.1684–1696. Available at: http://dx.doi.org/10.1261/rna.046011.114.

Choy, J.Y.H. et al., 2015. A resource of ribosomal RNA-depleted RNA-Seq data from different normal adult and fetal human tissues. *Scientific data*, 2, p.150063. Available at: http://dx.doi.org/10.1038/sdata.2015.63.

Clevers, H., 2011. The cancer stem cell: premises, promises and challenges. *Nature medicine*, 17(3), pp.313–319. Available at: http://dx.doi.org/10.1038/nm.2304.

Conesa, A. et al., 2016. A survey of best practices for RNA-seq data analysis. *Genome biology*, 17, p.13. Available at: http://dx.doi.org/10.1186/s13059-016-0881-8.

Crick, F.H., 1958. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12, pp.138–163. Available at: https://www.ncbi.nlm.nih.gov/pubmed/13580867.

van Dam, S. et al., 2017. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*. Available at: https://academic.oup.com/bib/article/doi/10.1093/bib/bbw139/2888441/Gene-co-expression-analysis-for-functional [Accessed June 5, 2017].

Datlinger, P. et al., 2017. Pooled CRISPR screening with single-cell transcriptome readout. *Nature methods*. Available at: http://dx.doi.org/10.1038/nmeth.4177 [Accessed January 18, 2017].

Deamer, D., Akeson, M. & Branton, D., 2016. Three decades of nanopore sequencing. *Nature biotechnology*, 34(5), pp.518–524. Available at: http://dx.doi.org/10.1038/nbt.3423.

Deng, Q. et al., 2014. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167), pp.193–196. Available at: http://dx.doi.org/10.1126/science.1245316.

DePristo, M.A. et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5), pp.491–498. Available at: http://dx.doi.org/10.1038/ng.806.

van Dijk, E.L., Jaszczyszyn, Y. & Thermes, C., 2014. Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental cell research*, 322(1), pp.12–20. Available at: http://dx.doi.org/10.1016/j.yexcr.2014.01.008.

Dixit, A. et al., 2016. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7), pp.1853–1866.e17. Available at: http://www.sciencedirect.com/science/article/pii/S0092867416316105.

van Dongen, J.J.M. et al., 2015. Minimal residual disease diagnostics in acute lymphoblastic leukemia: need for sensitive, fast, and standardized technologies. *Blood*, 125(26), pp.3996–4009. Available at: http://dx.doi.org/10.1182/blood-2015-03-580027.

Duan, C.-W. et al., 2014. Leukemia propagating cells rebuild an evolving niche in response to therapy. *Cancer cell*, 25(6), pp.778–793. Available at: http://dx.doi.org/10.1016/j.ccr.2014.04.015.

Edfors, F. et al., 2016. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Molecular systems biology*, 12(10), p.883. Available at: http://dx.doi.org/10.15252/msb.20167144.

Eid, J. et al., 2009. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910), pp.133–138. Available at: http://dx.doi.org/10.1126/science.1162986.

Engström, P.G. et al., 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods*, 10(12), pp.1185–1191. Available at:

http://dx.doi.org/10.1038/nmeth.2722.

Fan, H.C., Fu, G.K. & Fodor, S.P.A., 2015. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222), p.1258367. Available at: http://dx.doi.org/10.1126/science.1258367.

Fan, X. et al., 2015. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome biology*, 16, p.148. Available at: http://dx.doi.org/10.1186/s13059-015-0706-1.

Fedurco, M. et al., 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research*, 34(3), p.e22. Available at: http://dx.doi.org/10.1093/nar/gnj023.

Finak, G. et al., 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1), p.278. Available at: http://genomebiology.com/2015/16/1/278.

Fuzik, J. et al., 2016. Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nature biotechnology*, 34(2), pp.175–183. Available at: http://dx.doi.org/10.1038/nbt.3443.

Garalde, D.R. et al., 2016. Highly parallel direct RNA sequencing on an array of nanopores. *bioRxiv*, p.068809. Available at: http://biorxiv.org/content/early/2016/08/12/068809 [Accessed January 30, 2017].

Genohub, 2017. Next Generation Sequencing Instrument Guide. Available at: https://genohub.com/ngs-instrument-guide/ [Accessed June 3, 2017].

Genshaft, A.S. et al., 2016. Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome biology*, 17(1), p.188. Available at: http://dx.doi.org/10.1186/s13059-016-1045-6.

Gierahn, T.M. et al., 2017. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature methods*, 14(4), pp.395–398. Available at: http://dx.doi.org/10.1038/nmeth.4179.

Gökbuget, N. et al., 2012. Adult patients with acute lymphoblastic leukemia and molecular failure display a poor prognosis and are candidates for stem cell transplantation and targeted therapies. *Blood*, 120(9), pp.1868–1876. Available at: http://dx.doi.org/10.1182/blood-2011-09-377713.

Goodwin, S., McPherson, J.D. & McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews. Genetics*, 17(6), pp.333–351. Available at: http://dx.doi.org/10.1038/nrg.2016.49.

Grün, D. et al., 2015. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568), pp.251–255. Available at: http://dx.doi.org/10.1038/nature14966.

Grün, D., Kester, L. & van Oudenaarden, A., 2014. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6), pp.637–640. Available at: http://dx.doi.org/10.1038/nmeth.2930.

Guillaumet-Adkins, A. et al., 2017. Single-cell transcriptome conservation in cryopreserved

cells and tissues. *Genome biology*, 18(1), p.45. Available at: http://dx.doi.org/10.1186/s13059-017-1171-9.

Habib, N. et al., 2017. DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq. *bioRxiv*, p.115196. Available at: http://biorxiv.org/content/early/2017/03/09/115196 [Accessed March 28, 2017].

Halpern, K.B. et al., 2017. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*. Available at: http://dx.doi.org/10.1038/nature21065 [Accessed February 6, 2017].

Hashimshony, T. et al., 2016. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome biology*, 17(1), p.77. Available at: http://dx.doi.org/10.1186/s13059-016-0938-8.

Hashimshony, T. et al., 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports*, 2(3), pp.666–673. Available at: http://dx.doi.org/10.1016/j.celrep.2012.08.003.

Hayashi, T. et al., 2010. Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its "index sorting" function for stem cell research. *Development, growth & differentiation*, 52(1), pp.131–144. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1440-169X.2009.01157.x/full.

Hayden, E.C., 2014. Is the $1,000 genome for real? *Nature News*. Available at: http://www.nature.com/news/is-the-1-000-genome-for-real-1.14530 [Accessed June 3, 2017].

Hayer, K.E. et al., 2015. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics* , 31(24), pp.3938–3945. Available at: http://dx.doi.org/10.1093/bioinformatics/btv488.

Head, S.R. et al., 2014. Library construction for next-generation sequencing: overviews and challenges. *BioTechniques*, 56(2), pp.61–4, 66, 68, passim. Available at: http://dx.doi.org/10.2144/000114133.

Hicks, S.C., Teng, M. & Irizarry, R.A., 2015. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv*, p.025528. Available at: http://biorxiv.org/content/early/2015/12/27/025528 [Accessed November 13, 2016].

Hochgerner, H. et al., 2017. STRT-seq-2i: dual-index 5′ single cell and nucleus RNA-seq on an addressable microwell array. *bioRxiv*, p.126268. Available at: http://biorxiv.org/content/early/2017/04/20/126268 [Accessed May 8, 2017].

Huang, H.-L. et al., 2010. Trypsin-induced proteome alteration during cell subculture in mammalian cells. *Journal of biomedical science*, 17, p.36. Available at: http://dx.doi.org/10.1186/1423-0127-17-36.

Hunger, S.P. et al., 2012. Improved survival for children and adolescents with acute lymphoblastic leukemia between 1990 and 2005: a report from the children's oncology group. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 30(14), pp.1663–1669. Available at: http://dx.doi.org/10.1200/JCO.2011.37.8018.

Illumina, 2015. Patterned Flow Cell Technology. *Illumina Technology Whitepaper*. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/techno

tes/patterned-flow-cell-technology-technical-note-770-2015-010.pdf [Accessed June 6, 2017].

Inaba, H., Greaves, M. & Mullighan, C.G., 2013. Acute lymphoblastic leukaemia. *The Lancet*, 381(9881), pp.1943–1955. Available at: http://dx.doi.org/10.1016/S0140-6736(12)62187-4.

Irving, J. et al., 2014. Ras pathway mutations are prevalent in relapsed childhood acute lymphoblastic leukemia and confer sensitivity to MEK inhibition. *Blood*, 124(23), pp.3420–3430. Available at: http://dx.doi.org/10.1182/blood-2014-04-531871.

Islam, S. et al., 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome research*, 21(7), pp.1160–1167. Available at: http://dx.doi.org/10.1101/gr.110882.110.

Islam, S. et al., 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods*, 11(2), pp.163–166. Available at: http://dx.doi.org/10.1038/nmeth.2772.

Jain, M. et al., 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17(1), p.239. Available at: http://dx.doi.org/10.1186/s13059-016-1103-0.

Jaitin, D.A. et al., 2016. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*, 167(7), pp.1883–1896.e15. Available at: http://www.cell.com/cell/fulltext/S0092-8674(16)31611-7 [Accessed December 15, 2016].

Jaitin, D.A. et al., 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172), pp.776–779. Available at: http://dx.doi.org/10.1126/science.1247651.

Jiang, L. et al., 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome research*, 21(9), pp.1543–1551. Available at: http://dx.doi.org/10.1101/gr.121095.111.

Jones, P.A., 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13(7), pp.484–492. Available at: http://dx.doi.org/10.1038/nrg3230.

Kamel-Reid, S. et al., 1989. A model of human acute lymphoblastic leukemia in immune-deficient SCID mice. *Science*, 246(4937), pp.1597–1600. Available at: https://www.ncbi.nlm.nih.gov/pubmed/2595371.

Kharchenko, P.V., Silberstein, L. & Scadden, D.T., 2014. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7), pp.740–742. Available at: http://dx.doi.org/10.1038/nmeth.2967.

Kim, J.K. et al., 2015. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature communications*, 6, p.8687. Available at: http://dx.doi.org/10.1038/ncomms9687.

Kircher, M. & Kelso, J., 2010. High-throughput DNA sequencing--concepts and limitations. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 32(6), pp.524–536. Available at: http://dx.doi.org/10.1002/bies.200900181.

Kircher, M., Sawyer, S. & Meyer, M., 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research*, 40(1), p.e3. Available at: http://dx.doi.org/10.1093/nar/gkr771.

Kircher, M., Stenzel, U. & Kelso, J., 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome biology*, 10(8), p.R83. Available at: http://dx.doi.org/10.1186/gb-2009-10-8-r83.

Klein, A.M. et al., 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5), pp.1187–1201. Available at: http://dx.doi.org/10.1016/j.cell.2015.04.044.

Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C.H., et al., 2015. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell stem cell*, 17(4), pp.471–485. Available at: http://dx.doi.org/10.1016/j.stem.2015.09.011.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., et al., 2015. The technology and biology of single-cell RNA sequencing. *Molecular cell*, 58(4), pp.610–620. Available at: http://dx.doi.org/10.1016/j.molcel.2015.04.005.

Lake, B.B. et al., 2016. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, 352(6293), pp.1586–1590. Available at: http://dx.doi.org/10.1126/science.aaf1204.

La Manno, G. et al., 2016. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell*, 167(2), pp.566–580.e19. Available at: http://dx.doi.org/10.1016/j.cell.2016.09.027.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921. Available at: http://dx.doi.org/10.1038/35057062.

Langfelder, P. & Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9, p.559. Available at: http://dx.doi.org/10.1186/1471-2105-9-559.

Levin, J.Z. et al., 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods*, 7(9), pp.709–715. Available at: http://dx.doi.org/10.1038/nmeth.1491.

Liu, S.J. et al., 2016. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome biology*, 17, p.67. Available at: http://dx.doi.org/10.1186/s13059-016-0932-1.

Lowe, R. et al., 2017. Transcriptomics technologies. *PLoS computational biology*, 13(5), p.e1005457. Available at: http://dx.doi.org/10.1371/journal.pcbi.1005457.

Lun, A.T.L., Bach, K. & Marioni, J.C., 2016. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome biology*, 17(1), p.75. Available at: http://dx.doi.org/10.1186/s13059-016-0947-7.

Lun, A.T.L. & Marioni, J.C., 2017. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics* . Available at: http://dx.doi.org/10.1093/biostatistics/kxw055.

Lutz, C. et al., 2013. Quiescent leukaemic cells account for minimal residual disease in childhood lymphoblastic leukaemia. *Leukemia*, 27(5), pp.1204–1207. Available at: http://dx.doi.org/10.1038/leu.2012.306.

Macosko, E.Z. et al., 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5), pp.1202–1214. Available at: http://dx.doi.org/10.1016/j.cell.2015.05.002.

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–380. Available at: http://dx.doi.org/10.1038/nature03959.

Marinov, G.K. et al., 2014. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome research*, 24(3), pp.496–510. Available at: http://dx.doi.org/10.1101/gr.161034.113.

Marioni, J.C. et al., 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9), pp.1509–1517. Available at: http://dx.doi.org/10.1101/gr.079558.108.

Marra, M.A., Hillier, L. & Waterston, R.H., 1998. Expressed sequence tags--ESTablishing bridges between genomes. *Trends in genetics: TIG*, 14(1), pp.4–7. Available at: http://dx.doi.org/10.1016/S0168-9525(97)01355-3.

Martinez-Jimenez, C.P. et al., 2017. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, 355(6332), pp.1433–1436. Available at: http://dx.doi.org/10.1126/science.aah4115.

Meads, M.B., Gatenby, R.A. & Dalton, W.S., 2009. Environment-mediated drug resistance: a major contributor to minimal residual disease. *Nature reviews. Cancer*, 9(9), pp.665–674. Available at: http://dx.doi.org/10.1038/nrc2714.

Meyer, M. & Kircher, M., 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols*, 2010(6), p.db.prot5448. Available at: http://dx.doi.org/10.1101/pdb.prot5448.

Miller, M.B. & Tang, Y.-W., 2009. Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical microbiology reviews*, 22(4), pp.611–633. Available at: http://dx.doi.org/10.1128/CMR.00019-09.

Milligan, J.F. et al., 1987. Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucleic acids research*, 15(21), pp.8783–8798. Available at: https://www.ncbi.nlm.nih.gov/pubmed/3684574.

Moignard, V. et al., 2015. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature biotechnology*, 33(3), pp.269–276. Available at: http://dx.doi.org/10.1038/nbt.3154.

Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7), pp.621–628. Available at: http://dx.doi.org/10.1038/nmeth.1226.

Mullis, K. et al., 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology*, 51 Pt 1, pp.263–273.

Available at: https://www.ncbi.nlm.nih.gov/pubmed/3472723.

Muraro, M.J. et al., 2016. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell systems*, 3(4), pp.385–394.e3. Available at: http://dx.doi.org/10.1016/j.cels.2016.09.002.

Nagalakshmi, U. et al., 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881), pp.1344–1349. Available at: http://dx.doi.org/10.1126/science.1158441.

Nature Methods, 2014. Method of the year 2013. *Nature methods*, 11(1), p.1. Available at: http://dx.doi.org/10.1038/nmeth.2801.

Nestorowa, S. et al., 2016. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, 128(8), pp.e20–31. Available at: http://dx.doi.org/10.1182/blood-2016-05-716480.

Nguyen, K. et al., 2008. Factors influencing survival after relapse from acute lymphoblastic leukemia: a Children's Oncology Group study. *Leukemia*, 22(12), pp.2142–2150. Available at: http://dx.doi.org/10.1038/leu.2008.251.

Okonechnikov, K., Conesa, A. & García-Alcalde, F., 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* , 32(2), pp.292–294. Available at: http://dx.doi.org/10.1093/bioinformatics/btv566.

Pachmann, K., 1987. In situ hybridization with fluorochrome-labeled cloned DNA for quantitative determination of the homologous mRNA in individual cells. *The Journal of molecular and cellular immunology: JMCI*, 3(1), pp.13–19. Available at: https://www.ncbi.nlm.nih.gov/pubmed/3151062.

Pal, D., Heidenreich, O. & Vormoor, J., 2016. Dormancy Stems the Tide of Chemotherapy. *Cancer cell*, 30(6), pp.825–826. Available at: http://dx.doi.org/10.1016/j.ccell.2016.11.014.

Pan, C. et al., 2009. Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. *Molecular & cellular proteomics: MCP*, 8(3), pp.443–450. Available at: http://dx.doi.org/10.1074/mcp.M800258-MCP200.

Parekh, S. et al., 2016. The impact of amplification on differential expression analyses by RNA-seq. *Scientific reports*, 6, p.25533. Available at: http://dx.doi.org/10.1038/srep25533.

Patel, A.P. et al., 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190), pp.1396–1401. Available at: http://dx.doi.org/10.1126/science.1254257.

Picelli, S., Faridani, O.R., et al., 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols*, 9(1), pp.171–181. Available at: http://dx.doi.org/10.1038/nprot.2014.006.

Picelli, S. et al., 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11), pp.1096–1098. Available at: http://dx.doi.org/10.1038/nmeth.2639.

Picelli, S., Björklund, A.K., et al., 2014. Tn5 transposase and tagmentation procedures for massively-scaled sequencing projects. *Genome research*. Available at: http://dx.doi.org/10.1101/gr.177881.114.

Polak, R. et al., 2015. B-cell precursor acute lymphoblastic leukemia cells use tunneling nanotubes to orchestrate their microenvironment. *Blood*, 126(21), pp.2404–2414. Available at: http://dx.doi.org/10.1182/blood-2015-03-634238.

Pollen, A.A. et al., 2014. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology*. Available at: http://dx.doi.org/10.1038/nbt.2967.

Poplawski, A. & Binder, H., 2017. Feasibility of sample size calculation for RNA-seq studies. *Briefings in bioinformatics*. Available at: http://dx.doi.org/10.1093/bib/bbw144.

Poulin, J.-F. et al., 2016. Disentangling neural cell diversity using single-cell transcriptomics. *Nature neuroscience*, 19(9), pp.1131–1141. Available at: http://dx.doi.org/10.1038/nn.4366.

Pozhitkov, A.E., Tautz, D. & Noble, P.A., 2007. Oligonucleotide microarrays: widely applied--poorly understood. *Briefings in functional genomics & proteomics*, 6(2), pp.141–148. Available at: http://dx.doi.org/10.1093/bfgp/elm014.

Raj, A. et al., 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature methods*, 5(10), pp.877–879. Available at: http://dx.doi.org/10.1038/nmeth.1253.

Ramsköld, D. et al., 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology*, 30(8), pp.777–782. Available at: http://dx.doi.org/10.1038/nbt.2282.

Rapaport, F. et al., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology*, 14(9), p.R95. Available at: http://dx.doi.org/10.1186/gb-2013-14-9-r95.

Regev, A. et al., 2017. The Human Cell Atlas. *bioRxiv*, p.121202. Available at: http://biorxiv.org/content/early/2017/05/08/121202 [Accessed June 5, 2017].

Renaud, G. et al., 2015. deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics* , 31(5), pp.770–772. Available at: http://bioinformatics.oxfordjournals.org/content/31/5/770.abstract.

Risso, D. et al., 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9), pp.896–902. Available at: http://dx.doi.org/10.1038/nbt.2931.

Ritchie, M.E. et al., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7), p.e47. Available at: http://dx.doi.org/10.1093/nar/gkv007.

Robles, J.A. et al., 2012. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC genomics*, 13, p.484. Available at: http://dx.doi.org/10.1186/1471-2164-13-484.

Rosenberg, A.B. et al., 2017. Scaling single cell transcriptomics through split pool barcoding. *bioRxiv*, p.105163. Available at: http://biorxiv.org/content/early/2017/02/02/105163 [Accessed April 9, 2017].

Sandberg, R., 2014. Entering the era of single-cell transcriptomics in biology and medicine.

Nature methods, 11(1), pp.22–24. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24524133.

Sanger, F. et al., 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596), pp.687–695. Available at: https://www.ncbi.nlm.nih.gov/pubmed/870828.

Sasagawa, Y. et al., 2017. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *bioRxiv*. Available at: http://www.biorxiv.org/content/early/2017/07/05/159384.abstract.

Sawyers, C.L., Denny, C.T. & Witte, O.N., 1991. Leukemia and the disruption of normal hematopoiesis. *Cell*, 64(2), pp.337–350. Available at: https://www.ncbi.nlm.nih.gov/pubmed/1988151.

Schena, M. et al., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235), pp.467–470. Available at: https://www.ncbi.nlm.nih.gov/pubmed/7569999.

Seyednasrollah, F., Laiho, A. & Elo, L.L., 2015. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in bioinformatics*, 16(1), pp.59–70. Available at: http://dx.doi.org/10.1093/bib/bbt086.

Shapiro, E., Biezuner, T. & Linnarsson, S., 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics*, 14(9), pp.618–630. Available at: http://dx.doi.org/10.1038/nrg3542.

Shekhar, K. et al., 2016. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*, 166(5), pp.1308–1323.e30. Available at: http://dx.doi.org/10.1016/j.cell.2016.07.054.

Sheng, K. et al., 2017. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nature methods*. Available at: http://dx.doi.org/10.1038/nmeth.4145.

Simpson, J.T. et al., 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nature methods*, 14(4), pp.407–410. Available at: http://dx.doi.org/10.1038/nmeth.4184.

Soneson, C. & Delorenzi, M., 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, 14, p.91. Available at: http://dx.doi.org/10.1186/1471-2105-14-91.

Soneson, C. & Robinson, M.D., 2017. Bias, Robustness And Scalability In Differential Expression Analysis Of Single-Cell RNA-Seq Data. *bioRxiv*, p.143289. Available at: http://biorxiv.org/content/early/2017/05/28/143289 [Accessed May 29, 2017].

Soumillon, M. et al., 2014. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*. Available at: http://dx.doi.org/10.1101/003236.

Stegle, O., Teichmann, S.A. & Marioni, J.C., 2015. Computational and analytical challenges in single-cell transcriptomics. *Nature reviews. Genetics*, 16(3), pp.133–145. Available at: http://dx.doi.org/10.1038/nrg3833.

Streets, A.M. et al., 2014. Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 111(19), pp.7048–7053. Available at: http://dx.doi.org/10.1073/pnas.1402030111.

Svensson, V. et al., 2017. Power analysis of single-cell RNA-sequencing experiments. *Nature methods*. Available at: http://dx.doi.org/10.1038/nmeth.4220.

Tang, F. et al., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5), pp.377–382. Available at: http://dx.doi.org/10.1038/nmeth.1315.

Tarca, A.L., Bhatti, G. & Romero, R., 2013. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS one*, 8(11), p.e79217. Available at: http://dx.doi.org/10.1371/journal.pone.0079217.

Tasic, B. et al., 2016. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience*. Available at: http://dx.doi.org/10.1038/nn.4216.

Thomsen, E.R. et al., 2016. Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nature methods*, 13(1), pp.87–93. Available at: http://dx.doi.org/10.1038/nmeth.3629.

Tirosh, I. et al., 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282), pp.189–196. Available at: http://science.sciencemag.org/content/352/6282/189 [Accessed April 15, 2016].

Torre, E.A. et al., 2017. A Comparison Between Single Cell RNA Sequencing And Single Molecule RNA FISH For Rare Cell Analysis. *bioRxiv*, p.138289. Available at: http://biorxiv.org/content/early/2017/05/18/138289 [Accessed June 17, 2017].

Treutlein, B. et al., 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500), pp.371–375. Available at: http://dx.doi.org/10.1038/nature13173.

Trumpp, A. & Wiestler, O.D., 2008. Mechanisms of Disease: cancer stem cells—targeting the evil twin. *Nature reviews. Clinical oncology*, 5(6), pp.337–347. Available at: http://dx.doi.org/10.1038/ncponc1110 [Accessed June 11, 2017].

Tung, P.-Y. et al., 2017. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7, p.39921. Available at: http://dx.doi.org/10.1038/srep39921.

Turcatti, G. et al., 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research*, 36(4), p.e25. Available at: http://dx.doi.org/10.1093/nar/gkn021.

Usoskin, D. et al., 2015. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature neuroscience*, 18(1), pp.145–153. Available at: http://dx.doi.org/10.1038/nn.3881.

Valouev, A. et al., 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome research*, 18(7), pp.1051–1063. Available at: http://dx.doi.org/10.1101/gr.076463.108.

Vaquerizas, J.M. et al., 2009. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4), pp.252–263. Available at: http://dx.doi.org/10.1038/nrg2538.

Velculescu, V.E. et al., 1995. Serial analysis of gene expression. *Science*, 270(5235), pp.484–487. Available at: https://www.ncbi.nlm.nih.gov/pubmed/7570003.

Venteicher, A.S. et al., 2017. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science*, 355(6332). Available at: http://dx.doi.org/10.1126/science.aai8478.

Vick, B. et al., 2015. An advanced preclinical mouse model for acute myeloid leukemia using patients' cells of various genetic subgroups and in vivo bioluminescence imaging. *PloS one*, 10(3), p.e0120925. Available at: http://dx.doi.org/10.1371/journal.pone.0120925.

Vieth, B. et al., 2017. powsimR: Power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* . Available at: https://academic.oup.com/bioinformatics/article-abstract/doi/10.1093/bioinformatics/btx435/3952669/powsimR-Power-analysis-for-bulk-and-single-cell?redirectedFrom=fulltext [Accessed July 25, 2017].

Villani, A.-C. et al., 2017. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335). Available at: http://dx.doi.org/10.1126/science.aah4573.

Vogel, C. & Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13(4), pp.227–232. Available at: http://dx.doi.org/10.1038/nrg3185.

Voss, T.C. & Hager, G.L., 2014. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature reviews. Genetics*, 15(2), pp.69–81. Available at: http://dx.doi.org/10.1038/nrg3623.

Wagner, A., Regev, A. & Yosef, N., 2016. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11), pp.1145–1160. Available at: http://dx.doi.org/10.1038/nbt.3711.

Wang, L., Wang, S. & Li, W., 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* , 28(16), pp.2184–2185. Available at: http://dx.doi.org/10.1093/bioinformatics/bts356.

Wang, W. et al., 2017. High fidelity hypothermic preservation of primary tissues in organ transplant preservative for single cell transcriptome analysis. *bioRxiv*, p.115733. Available at: http://www.biorxiv.org/content/early/2017/03/10/115733 [Accessed May 3, 2017].

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), pp.57–63. Available at: http://dx.doi.org/10.1038/nrg2484.

Weis, J.H. et al., 1992. Detection of rare mRNAs via quantitative RT-PCR. *Trends in genetics: TIG*, 8(8), pp.263–264. Available at: https://www.ncbi.nlm.nih.gov/pubmed/1509514.

Wilhelm, B.T. et al., 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199), pp.1239–1243. Available at: http://dx.doi.org/10.1038/nature07002.

Winkler, H., 1920. Verbreitung und Ursache der Parthenogenesis im Pflanzen-und Tierreiche. Available at: http://agris.fao.org/agris-search/search.do?recordID=US201300440276.

Wu, A.R. et al., 2014. Quantitative assessment of single-cell RNA-sequencing methods. *Nature methods*, 11(1), pp.41–46. Available at: http://dx.doi.org/10.1038/nmeth.2694.

Wu, H., Wang, C. & Wu, Z., 2015. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* , 31(2), pp.233–241. Available at: http://dx.doi.org/10.1093/bioinformatics/btu640.

Xie, S. et al., 2017. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular cell*, 66(2), pp.285–299.e5. Available at: http://dx.doi.org/10.1016/j.molcel.2017.03.007.

Xin, Y. et al., 2016. Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. *Proceedings of the National Academy of Sciences of the United States of America*, 113(12), pp.3293–3298. Available at: http://dx.doi.org/10.1073/pnas.1602306113.

Yamamoto, J.F. & Goodman, M.T., 2008. Patterns of leukemia incidence in the United States by subtype and demographic characteristics, 1997–2002. *Cancer causes & control: CCC*, 19(4), pp.379–390. Available at: https://link.springer.com/article/10.1007/s10552-007-9097-2 [Accessed June 8, 2017].

Yamamoto, M. et al., 2001. Use of serial analysis of gene expression (SAGE) technology. *Journal of immunological methods*, 250(1-2), pp.45–66. Available at: https://www.ncbi.nlm.nih.gov/pubmed/11251221.

Yan, L. et al., 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9), pp.1131–1139. Available at: http://dx.doi.org/10.1038/nsmb.2660.

Zajac, P. et al., 2013. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PloS one*, 8(12), p.e85270. Available at: http://dx.doi.org/10.1371/journal.pone.0085270.

Zeisel, A. et al., 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*. Available at: http://dx.doi.org/10.1126/science.aaa1934.

Zheng, G.X.Y. et al., 2017. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8, p.14049. Available at: http://dx.doi.org/10.1038/ncomms14049.

Zhu, Y.Y. et al., 2001. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques*, 30(4), pp.892–897. Available at: https://www.ncbi.nlm.nih.gov/pubmed/11314272.

Ziegenhain, C. et al., 2017. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular cell*, 65(4), pp.631–643.e4. Available at: http://dx.doi.org/10.1016/j.molcel.2017.01.023.

Zilionis, R. et al., 2017. Single-cell barcoding and sequencing using droplet microfluidics. *Nature protocols*, 12(1), pp.44–73. Available at: http://dx.doi.org/10.1038/nprot.2016.154.

Zimmerman, E., 2014. 50 Smartest Companies: Illumina. *MIT Technology Review*. Available at: https://www.technologyreview.com/s/524531/why-illumina-is-no-1/ [Accessed June 3, 2017].

List of Figures

Acknowledgements

During my PhD, I met so many great people that it may not be possible to list everyone here.
If I forgot to mention you, sorry! Here goes my best try, in no particular order:

First and foremost, I want to thank my PhD advisor and mentor Wolfgang Enard. Wolfi, you are the most amazing boss I could have wished for. Thank you so much for always being there for advice while giving me freedom to explore my interests and grow as a person. Your ideas, thinking and guidance have shaped my idea of what really makes a scientist.
I am very grateful to Ines Hellmann for teaching and advising me in computational work so I can call myself "computational biologist" now. Big thanks to my fellow monkeys Beate and Swati for sharing all the emotions of PhD student life, great friendship, helping me out with big and small (computational) problems, occasional rants, yummy breakfasts, deflecting phone calls and many many more things that are impossible to list.

Furthermore I would like to thank:
Sabrina for being a great first student. Jojo for nice music, lots of help and cleaning the lab. Lu and Jojohanna for charming rides home, cool cells and their help. Daniel, Ilse and Mari for awesome beers, food and company. Ines B, Karin, Michi and Steffi for being the backbone of our group with helpful hands and open ears. All our lovely past and present students: Khalis, Gunnar, Chris, Lukas, Zane, Aleks, ...
In short: thanks to everyone in the Enard lab for being great colleagues and even better friends.

Thanks to all our collaborators for interesting projects and the successful work.

I owe big thanks to my family and friends. Thank you Enrico for reminding me that there is a life outside of work and science. Thank you Mama and Papa for being the best parents ever, your never ending support and unconditional love.

Thank you Elisa for the coolest microbio collab, day-to-day PhD advice and pushing me to get shit done. Thanks to kleine Chrissi, große Chrissi, Basti and Tina for being awesome and sticking together after all the years.

Christoph Ziegenhain

Education

Ph.D. candidate, Ludwig-Maximilians-University Munich, Germany **2013 – present**

Advisor: Prof. Wolfgang Enard

Master of Science, Biology & Biotechnology, University of Copenhagen, Denmark **2011 – 2013**

Bachelor of Science, Biology, Ludwig-Maximilians-University Munich, Germany **2008 – 2011**

Research Experience

PhD research, Ludwig-Maximilians-University Munich, Germany **2013 – present**

Advisor: Prof. Wolfgang Enard
Application of single-cell transcriptomics in evolutionary context

Master thesis (12 months), Helmholtz Zentrum München, Germany **2012 – 2013**

Advisors: Søren Skov, University of Copenhagen & Ruth Brack-Werner, Helmholtz Zentrum München
Title: *Establishment & optimization of a latently HIV-infected human neural reporter cell line.*

Bachelor thesis (6 months), Ludwig-Maximilians-University Munich, Germany **2011**

Advisor: Ralf Heermann, Ludwig-Maximilians-University Munich
Title: *The role of PAS4-LuxR solos for the host-bacterium interaction of Photorhabdus luminescens.*

Publications

1. Schneider M, Tigges B, Meggendorfer M, Helfer M, **Ziegenhain C**, Brack-Werner R: *A new model for post-integration latency in macroglial cells to study HIV-1 reservoirs of the brain.* **AIDS 2015**

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

3. Schreck C*, Istvanffy R*, **Ziegenhain C**, Sippenauer T, Ruf F, Henkel L, Gärtner F, Vieth B, Florian MC, Mende N, Taubenberger A, Prendergast A, Wagner A, Pagel C, Grziwok S, Götze KS, Guck J, Dean DC, Massberg S, Essers M, Waskow C, Geiger H, Schiemann M, Peschel C, Enard W, Oostendorp RAJ:
Niche WNT5A regulates the actin cytoskeleton during regeneration of hematopoietic stem cells. **Journal of Experimental Medicine 2016**

4. Ebinger S*, Özdemir EZ*, **Ziegenhain C***, Tiedt S*, Alves CC*, Grunert M, Dworzak M, Lutz C, Horny HP, Sotlar K, Parekh S, Spiekermann K, Hiddemann W, Schepers A, Polzer B, Kirsch S, Hoffmann M, Knapp B, Hasenauer J, Pfeifer H, Panzer-Grümayer R, Enard W, Gires O, Jeremias I: *Characterization of rare, dormant and therapy resistant stem cells in acute lymphoblastic leukemia.* **Cancer Cell 2016**

5. **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.* **Molecular Cell 2017**

Publications (continued)

6. Witzel M, Petersheim D, Fan Y, Bahrami E, Racek T, Rohlfs M, Puchalka J, Mertes C, Gagneur J, **Ziegenhain C**, Enard W, Stray-Pederson A, Arkwright PD, Abboud MR, Pazhakh V, Lieschke GJ, Mundlos S, Krawitz PM, Dahlhoff M, Schneider MR, Wolf E, Horny HP, Schmidt H, Schäffer AA, Klein C:
Chromatin remodelling factor SMARCD2 regulates transcriptional networks controlling early and late differentiation of neutrophil granulocytes. **Nature Genetics 2017**

7. Vieth B, **Ziegenhain C**, Parekh S, Enard W, Hellmann I:
powsimR: Power analysis for bulk and single cell RNA-seq experiments. **Bioinformatics 2017**

8. Parekh S*, **Ziegenhain C***, Vieth B, Enard W, Hellmann I:
zUMIs: A fast and flexible pipeline to process RNA sequencing data with UMIs. **bioRxiv 2017**

9. Garz AK, Wolf S, Grath S, Harbinger S, Vick B, Gaidzik V, Rudelius M, Smets M, Herold S, **Ziegenhain C**, Weickert MT, Zwick A, Oostendorp RAJ, Peschel C, Bultmann S, Döhner K, Jeremias I, Thiede C, Keller U, Götze KS:
Crenolanib in combination with azacitidine abrogates stromal protection and survival of leukemia-initiating cells in FLT3-ITD+ AML. **Oncotarget 2017**

10. Krendl C*, Shaposhnikov D*, Rishko V, Ori C, **Ziegenhain C**, Sass S, Simon L, Müller NS, Straub T, Brooks KE, Chavez SL, Enard W, Theis FJ, Drukker M:
GATA2/3-TFAP2A/C transcription factor network couples human ES cell differentiation to trophectoderm with repression of pluripotency. **PNAS 2017**

11. Böttcher A, Büttner M, Tritschler S, Sterr M, Aliluev A, Burtscher I, Sass S, Irmler M, Beckers J, **Ziegenhain C**, Enard W, Schamberger AC, Verhamme FM, Eickelberg O, Theis FJ, Lickert H:
Wnt/planar cell polarity primed intestinal stem cells directly differentiate into enteroendocrine or Paneth cells. **In Revision**

12. Granato ET, **Ziegenhain C**, Kümmerli R:
Virulence evolution in an opportunistic bacterial pathogen. **Submitted**

13. Müller S*, Engleitner T*, Maresch R*, Zukowska M, Lange S, Konukiewitz B, Kaltenbacher T, Zwiebel M, Öllinger R, Strong A, Yen H, Steiger K, Banerjee R, Louzada S, Fu B, Seidler B, Götzfried J, Hassan Z, Schuck K, Schönhuber N, Veltkamp C, Friedrich M, Rad L, Barenboim M, **Ziegenhain C**, Dovey OM, Eser S, Parekh S, Constantino-Casas F, de la Rosa J, Cadiñanos J, Sierra MI, Fraga M, Klöppel G, Weichert W, Liu P, Vassiliou G, Schmid RM, Enard W, Yang F, Unger K, Schneider G, Varela I, Bradley A, Saur D, Rad R:
Evolutionary trajectories and KRAS gene dosage define pancreatic cancer phenotypes. **In Revision**

14. Gegenfurtner FA, Jahn B, Wagner H, **Ziegenhain C**, Wolfgang Enard W, Geistlinger L, Rädler JO, Vollmar AM, Zahler S:
MRTF-A and YAP share extracellular triggers but underlie different 2 kinetics and regulatory tasks in endothelial cells. **Submitted**

15. Bagnoli, JW*, **Ziegenhain C***, Janjic A*, Wange LE, Vieth B, Parekh S, Geuder J, Hellmann I, Enard W:
mcSCRB-seq: sensitive and powerful single-cell RNA sequencing. **In preparation**

Presentations

Single Cell Genomics Conference
Poster: *Single-cell expression profiling of rare subpopulations in acute leukemia.* **Utrecht, Netherlands 2015**

Cutting-Edge Technologies Event
Talk: *One by one: The age of single-cell genomics.* **Munich, Germany 2015**

ISMB
Poster: *Comparative analysis of single-cell RNA sequencing methods* **Orlando, USA 2016**

Keystone Single-Cell Omics
Poster: *Comparative analysis of single-cell RNA sequencing methods* **Stockholm, Sweden 2017**

Grants

Boehringer Ingelheim Travel Grant
1 month research stay at Broad Technology Labs, The Broad Institute, Cambridge MA. **2015**

Teaching

Master thesis advisor **2013 - 2017**
Supervision and mentoring of several students for their MSc thesis

Human Biology I **2014, 2015, 2016**
1 week lab course for bachelor students