Analyzing Acute Myeloid Leukemia by RNA-sequencing

Johannes Walter Bagnoli

Dissertation der Fakultät für Biologie der

Ludwig-Maximilians-Universität München

München

14.05.2020

1. Gutachter: Herr Prof. Wolfgang Enard

2. Gutachter: Herr Prof. Heinrich Leonhardt

Tag der Abgabe: 14.05.2020

Tag der mündlichen Prüfung: 20.11.2020

Table of Contents

Summary

Bulk and single cell RNA sequencing have revolutionized biomedical research and empower researchers to quantify the global gene expression of populations and single cells to further understand the development, manifestation and the treatment of diseases like cancer. Acute myeloid leukemia (AML), a cancer of the myeloid line of blood cells, could benefit from these technologies as relapse and mortality rates remain high despite the extensive research conducted over several decades. This is partly because AML is a heterogeneous disease, differing substantially between patients and hence requiring more fine-grained classifications and specialised treatment strategies, for example by incorporating expression profiles. In addition, single cell RNA sequencing (scRNA-seq) can resolve genetic and epigenetic subclonal structures within a patient to improve understanding and treatment of AML. However, improving and adapting RNA-seq technologies is still often necessary to efficiently and reliably obtain expression profiles, especially from small or suboptimally processed samples. To this end, we developed a bulk RNA-seq protocol, which copes with the major challenges of limited sample quantities, different sample types, throughput and costs and subsequently applied this method to further understand the subclonal structures in AML.

We were able to characterize a plastic cell state of AML cells that is defined by increased stemness and dormancy and could influence treatment outcome and relapse. For this, we isolated non-dividing AML cells based on a proliferation-sensitive dye from patient derived xenograft (PDX) models of two AML patients. We found that these cells have low levels of cell cycle genes confirming dormancy, and additionally had similar expression patterns to previously described dormant minimal residual disease (MRD) cells in lymphoblastic leukemia (ALL). This included high expression levels of cell adhesion molecules, potentially

reflecting the persistence of dormant AML and ALL cells in the hematopoietic niche. Lastly, we could show that resting and cycling AML cells can transition between these two states, indicating that dormancy might be a general property of AML cells and not depend on particular genetic subclones.

In a second project, we optimized a single cell RNA-seq technology. We used a systematic approach to evaluate experimental conditions of SCRB-seq, a powerful and efficient scRNA-seq method. Focussing on reverse transcription, arguably the most important and inefficient reaction, , we used a standardized human RNA (UHRR) and systematically tested nine different RT enzymes, several reaction enhancers and primer compositions to increase sensitivity. We found that Maxima H- showed the highest sensitivity and that molecular crowding using poylethylene glycol (PEG) could increase the efficiency of the reaction significantly. Together with several smaller changes in the workflow, primer design and PCR conditions, we developed mcSCRB-seq (molecular crowding SCRB-seq). We verified the 2.5x increase in sensitivity using mES cells in a side by side test between SCRB-seq and mcSCRB-seq, and further found mcSCRB-seq to be amongst the most sensitive methods using artificial RNA spike in molecules (ERCCS).

Lastly, since method comparisons between studies suffer from missing accuracy due to batch effects and external factors, we participated in a complex scRNA-seq benchmark study aiming to provide a fair comparison between methods concerning sensitivity, accuracy and applicability for building expression atlases. In contrast to before, we found that in this particular setting, mcSCRB-seq did not perform well and ídentified fields for further improvement.

In conclusion, my work described in this thesis not only contributes towards a deeper understanding of the emergence and progression of AML but also towards the development

of experimental bulk and single-cell RNA sequencing methods, improving their widespread application to biomedical problems such as leukemia.

Introduction

Gene expression levels define cellular phenotypes

The function of each cell is realized by a multilevel system with biopolymers as the key players. This hierarchically organized system is the central dogma of molecular biology (Figure 1). DNA is the essential molecule that contains all genetic information (Avery et al., 1944), which can subsequently be transcribed into transient messenger RNA (mRNA) and further translated into proteins, which are responsible for the molecular phenotype and hence the function of the cell (Crick, 1958). Although each cell of a multicellular organism contains the same genetic information, the variety of functional and phenotypic specifications require a defined system to regulate the cells' identity. Several mechanisms are responsible to manage the information transfer between the biomolecules (Edfors et al., 2016; Vogel and Marcotte, 2012). On the DNA level, DNA methylation (Jones, 2012) and chromatin modifications (Voss and Hager, 2014) determine epigenetically, i.e. across cell divisions, the packing of chromatin. This in turn regulates whether transcription factors can access and bind DNA sequences and in turn modify the level of gene expression (Ong and Corces, 2011; Vaquerizas et al., 2009). The regulatory effect of these systems arises from either silencing, and therefore switching off transcription, or via the activation of a gene by enhancing the responsible transcription machinery. Although expression level modification is mainly achieved at the level of DNA to RNA transcription, diverse mechanisms also regulate RNA and protein levels after transcription and translation, respectively. For example, on the RNA level, microRNAs (miRNAs) can inhibit the translation to proteins (Rana, 2007). Furthermore, post-transcriptional and post-translational modifications are responsible for fine

tuning the system (Knorre et al., 2009; Zhao et al., 2017). Overall these mechanisms provide an agile and complex system to regulate the quantitative relationship between genetic information and cellular function (Vogel and Marcotte, 2012). Hence, understanding these biological processes is of high significance, especially in order to understand malfunctions leading to disease developments such as cancer (Vaquerizas et al., 2009).

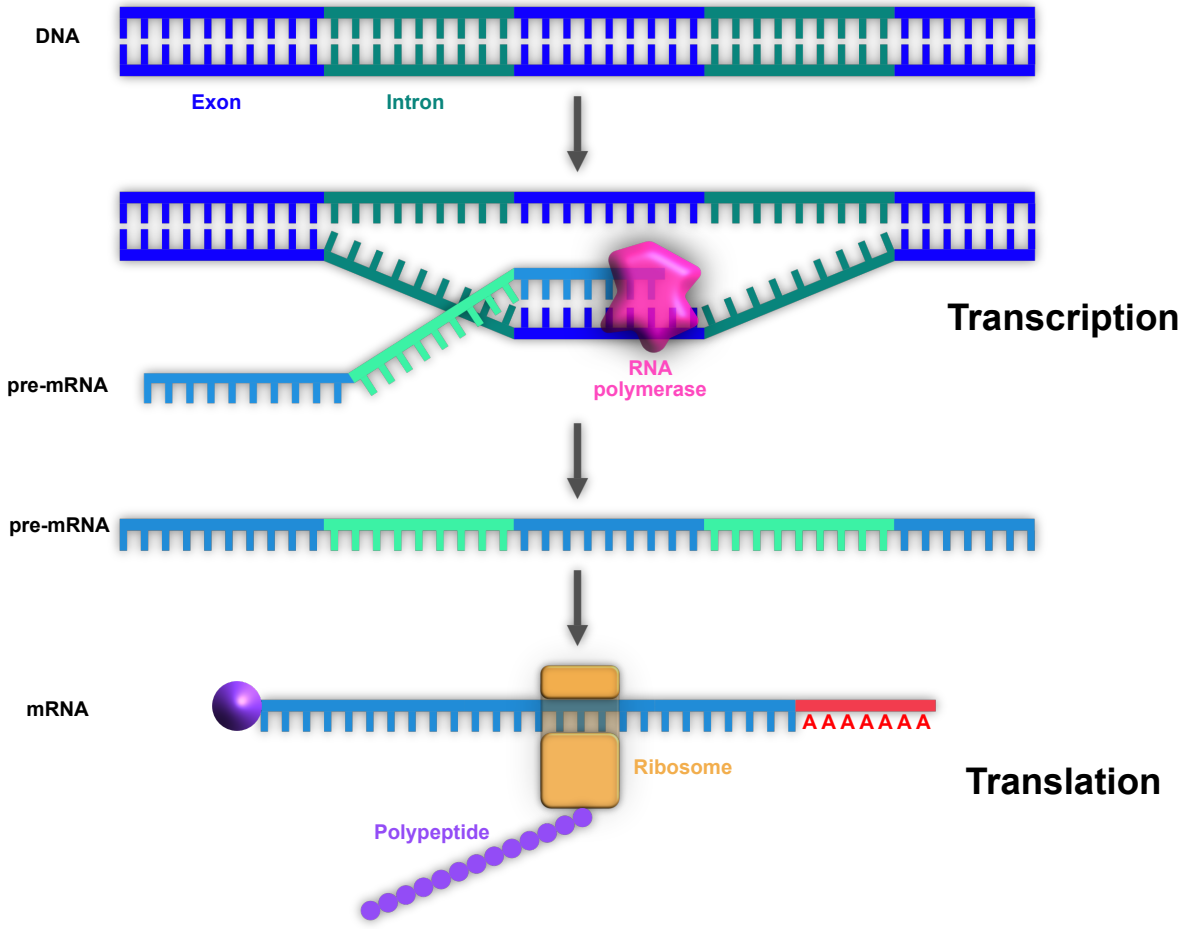


Figure 1: The central dogma of molecular biology
DNA is the essential molecule containing all genetic information. Transcription into transient pre-messenger RNA (pre-mRNA), and further maturation to mature mRNA via splicing, polyadenylation and 5'capping, is necessary to provide the molecular basis for translation into proteins, which subsequently are responsible for the cells function.

Quantification of gene expression

The high power and potential of gene expression analysis to understand important biological processes led to the development of suitable quantification methods in the 1970s. However, quantifying gene expression levels was initially limited to individual transcripts. For example, RNA molecules could be separated by gel electrophoresis, transferred to a paper membrane and detected via radioactively labelled probes in so-called Northern Blots (Alwine et al., 1977). Other studies already used DNA synthesis from RNA molecules (cDNA) coupled with a semiquantitative dot hybridization to estimate relative expression abundance (Sim et al., 1979). Further improvements in the 1980s included in-situ fluorescence probes ("fluorescence in-situ hybridization", FISH) (Pachmann, 1987) and RT-qPCR (Becker-André and Hahlbrock, 1989; Weis et al., 1992). The first methods which were able to obtain expression profiles from many or even all transcripts emerged in the 1990s. Among these, "serial analysis of gene expression" (SAGE) (Velculescu et al., 1995) and microarrays (Schena et al., 1995) provided quantitative global expression data. While SAGE relies on enzymatic digestion of cDNA and Sanger sequencing, microarrays depend on the hybridization of fluorescently labeled cDNA to custom immobilized oligonucleotide probes, complementary to cDNA sequences. Due to their low costs, good quality and simple usage, microarrays became the most popular method for global expression quantification within the 2000s and still remain in use today (Lowe et al., 2017). Although microarrays underwent several advances over the years, including oligo synthesis (Miller and Tang, 2009) and fluorescence detection (Pozhitkov et al., 2007) the *a priori* knowledge of cDNA sequences necessary to design the oligonucleotide probes remained a major drawback and limits its use for *de novo* applications and species with poorly resolved transcriptomes. In addition, a high

background noise due to mishybridization of transcripts to probes is often observed in microarrays (Okoniewski and Miller, 2006).

Next generation sequencing (NGS)

Since the emergence of Next Generation Sequencing (NGS) in 2005, sequencing costs and throughput have improved drastically compared to the traditional Sanger sequencing approach (Kircher and Kelso, 2010; Sanger et al., 1977). Although several methods were introduced to the market over the years (Clarke et al., 2009; Drmanac et al., 2010; Margulies et al., 2005; Valouev et al., 2008), Illumina's variant of sequencing by synthesis outperformed all others and currently constitutes the majority of the sequencing market worldwide (Utterback, 2020). In short, DNA libraries for Illumina sequencing are tagged with immobilization adapters, which enable the fragments to bind to a flowcell that features complimentary immobilized oligonucleotides on its surface (Fedurco et al., 2006). Then, the fragments are amplified in a "bridge amplification" reaction to form clusters consisting of thousands of copied molecules in close proximity. The sequence information of each of these clusters is then read via a sequencing by synthesis reaction in which fluorescently labelled nucleotides are integrated. These labelled nucleotides contain an additional cleavable 3` chain terminator ensuring that only one base is incorporated at a time. After incorporation, the fluorophores are illuminated via lasers and the signals are imaged. Finally, the 3` terminator and fluorescence labels are chemically cleaved and the next incorporation cycle is performed. In general, the libraries can be sequenced on both the forward and reverse strand with a read-length up to 600 bases (300 bases paired-end) (Kircher and Kelso, 2010; Kircher et al., 2009) .

RNA-sequencing revolutionized transcriptomics

With the introduction of NGS a new set of methods called RNA-sequencing (RNA-seq) overcame the limitations of previous expression quantification assays. First used by several groups in 2006 and 2007 and further refined and named in 2008, RNA-sequencing relies on the combination of high throughput sequencing of cDNA libraries and corresponding computational tools to analyze and quantify gene expression of an RNA sample both, in a quantitative and qualitative manner (Bainbridge et al., 2006; Cheung et al., 2006; Emrich et al., 2007; Marioni et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Weber et al., 2007; Wilhelm et al., 2008). Although comprising its own challenges, RNA-seq provides several key advantages over previous methods like Microarrays or RT-qPCRs. First, it is not confined by necessary *a priori* knowledge of sequences and can therefore be used to also study non-model organisms and find *de novo* transcripts (Vera et al., 2008). In addition, RNA-seq is not only capable of quantifying the expression of genes but also revealing isoform and allelic information of transcripts from the same gene as well as finding mutations such as single nucleotide polymorphism/variants (SNP/SNV), indels and even fusion genes due to its single base resolution. Furthermore, its accuracy, theoretical unlimited quantification range and reproducibility over technical and biological replicates outperforms other methods with an additional decrease in costs and necessary input amounts (Cloonan et al., 2008; Marioni et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008). With constant improvements to both the construction of libraries from RNA as well as high throughput DNA sequencing, RNA-seq became the dominant transcriptomic method by 2015 (Lowe et al., 2017). Although the rapid evolution in the methodology generated numerous different RNA-seq protocols over the years a general workflow can be attributed to most of them (Levin et al., 2010) (Figure 2).

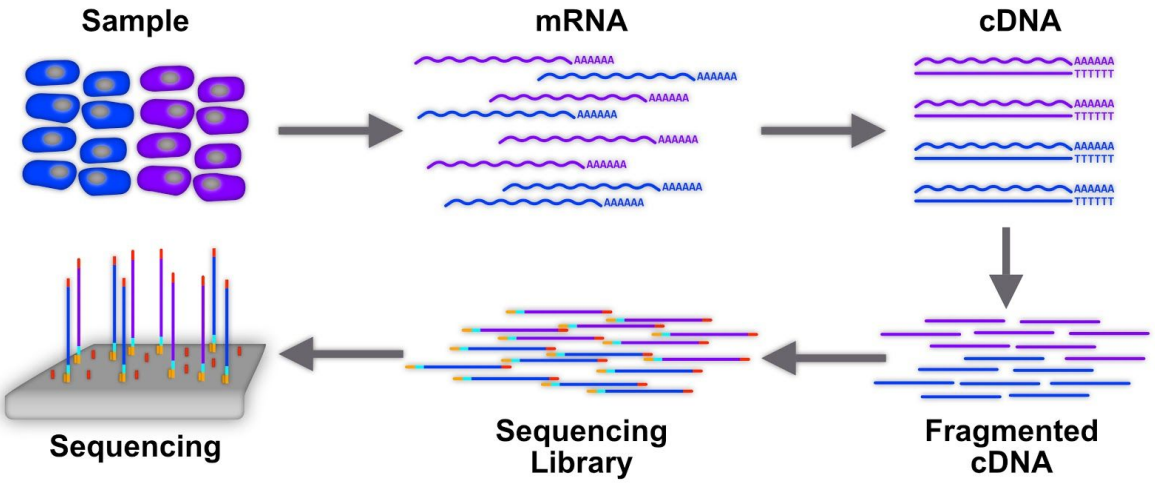


Figure 2: General experimental workflow of RNA-seq
RNA sequencing firstly requires the extraction of RNA of any biological samples. After reverse transcription, cDNA is fragmented and barcoded with multiplexing sequences and sequencing adapters. Finally, high-throughput sequencing is used to obtain the encoded information of the library.

The first step requires the isolation of RNA from samples, which can be done in several ways (Kałużna et al., 2016; Shu et al., 2014). However, as mRNA only makes up about 5% of the total RNA (Warner, 1999) , all protocols either actively deplete rRNAs, which contribute to 80%, or enrich for mRNA by selecting polyadenylated RNAs (Choy et al., 2015). In the next step, RNA must be converted to cDNA via a reverse transcription reaction. Afterwards, the resulting cDNA has to be fragmented into smaller molecules in order to be able to generate clusters on a flowcell. This step is crucial as most mRNA transcripts from eukaryotic organisms exceed the maximum recommended fragment length of ~1 kb for Illumina sequencing (Lowe et al., 2017; Wang et al., 2009). It should be noted that this fragmentation process can also be performed prior reverse transcription on the RNA level mostly via RNA hydrolysis (Mortazavi et al., 2008). On the cDNA level, fragmentation is mostly achieved via sonication (Head et al., 2014) or enzymatic reactions (Adey et al., 2010; Picelli et al., 2014). In the final step, fragmented cDNA libraries are constructed via a PCR, ensuring correct

sequencing adapter addition, sufficient library concentration and possible sample barcode addition (van Dijk et al., 2014; Kircher et al., 2012; Meyer and Kircher, 2010).

Due to the large amount of sequencing data derived from typical RNA-seq experiments, interpreting the results of these experiments requires a combination of bioinformatic and statistical tools and pipelines (Lowe et al., 2017). In general, sequencers convert the image data to more suitable file formats, which include base callings and adjusted quality scores (Kircher et al., 2009). Since most sequencing runs are performed with several samples per run, a demultiplexing step, separating the different samples into independent files via their corresponding barcode sequences, is necessary (Renaud et al., 2015). Due to the relatively high error rate of NGS it is important to perform a QC step on the raw sequencing files to ensure that only reads with high base call quality are used for further analysis (Andrews, 2010) . A major computational challenge is the following alignment of the cDNA reads to the reference genome. Although there are several tools available for mapping short sequencing reads, RNA-seq reads require special tools which can deal with splice junctions and mapping of intron skipping reads towards the genome (Baruzzo et al., 2017; Dobin et al., 2013; Engström et al., 2013; Hayer et al., 2015). Similar to the first quality control, the mapping quality should be analyzed to ensure only well mapped reads are used (Wang et al., 2012). After this primary data processing, higher level analysis requires statistical models for normalization (Hrdlickova et al., 2017), dimension reduction clustering (Kobak and Berens, 2019; Yeung and Ruzzo, 2001), differential gene expression (Love et al., 2014; Ritchie et al., 2015) and gene set enrichment analysis (Tarca et al., 2013). However, these examples show only a small subset of the numerous computational analyses that are possible with the complex information provided by RNA-seq (Wang et al., 2009).

Single cell RNA-sequencing develops rapidly

Despite the increased efficiency of RNA-seq compared to other methods, it still requires relatively large amounts of input material, typically the RNA of thousands of cells. This limitation becomes increasingly significant when identifying and characterizing new subpopulations and rare cell types, understanding cell to cell heterogeneity or uncovering developmental processes (Kolodziejczyk et al., 2015; Tang et al., 2009; Wagner et al., 2016; Ziegenhain et al., 2018). However, over the last decade, continuous improvements in whole transcriptome amplification have made it possible to perform RNA-seq at the level of single cells (Kurimoto, 2006; Kurimoto et al., 2007; Tang et al., 2009). This resulted in numerous different single cell RNA-sequencing (scRNA-seq) protocols (Kolodziejczyk et al., 2015; Ziegenhain et al., 2017) and transformed our understanding of biology. Similar to the numerous conventional RNA-seq methods, almost all scRNA-seq protocols follow a similar experimental workflow overcoming the two major challenges not previously present, the isolation of single cells and amplification of the very small amounts of RNA. Most common scRNA-seq protocols either use fluorescence-activated cell sorting (FACS) (Bagnoli et al., 2018; Jaitin et al., 2014; Picelli et al., 2013; Soumillon et al., 2014), capturing cells in microfluidic chips (Hashimshony et al., 2016; Wu et al., 2014) or encapsulation of single cells in microdroplets (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017) (Figure 3). While the later two approaches are capable of capturing hundreds of thousands of cells and therefore are more suited to perform large scale experiments, FACS isolation into microwell plates offers more flexibility in study design and sample type. Therefore the corresponding cell isolation technique can already make a protocol more or less suited for a specific research question (Ziegenhain et al., 2018). In addition, all approaches require a

solution of single cells, which depending on the sample, can already be challenging and introduce biases (van den Brink et al., 2017).

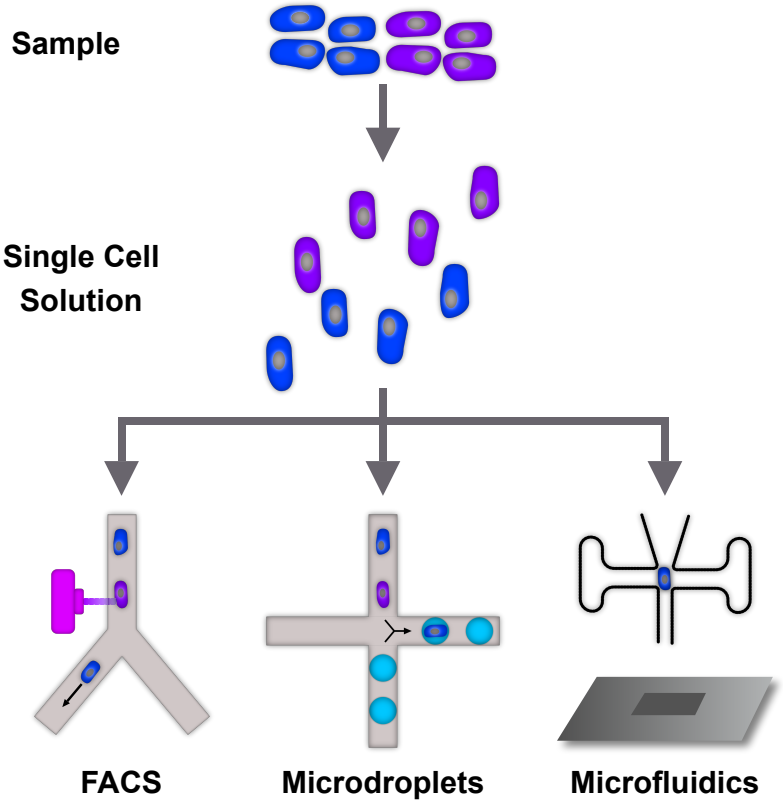


Figure 3: Isolation of single cells Illustration of typical single-cell isolation techniques. scRNA-seq requires the capture of non-damaged, living cells. Hence, each method comes with its own advantages and challenges regarding the incorporation of stainings, throughput and influence on the cells healthiness and arguably the transcriptome.

After obtaining single cells, scRNA-seq approaches follow a similar workflow as bulk RNA-seq methods (Figure 4). First, cDNA is generated via reverse transcription and is afterwards amplified either via PCR or *in vitro* transcription. Since both reactions are highly impacted by the small starting amounts it is of crucial importance that both reactions are as sensitive and unbiased as possible. For example, conversion efficiencies of mRNA to cDNA are estimated to be between 10%-49% (Bagnoli et al., 2018; Grün et al., 2014; Islam et al., 2014). This subsequently impacts the following amplification steps as more cycles are required to obtain the necessary cDNA yields introducing noise and biases (Parekh et al., 2016; Ziegenhain et al., 2017). Although, being commonly used reactions in molecular

biology, the sensitivity, efficacy and accuracy of both reactions depend on a complex combination of several factors including but not limited to, enzyme properties, buffer composition, primer sequences and reaction volume (Bagnoli et al., 2018; Hashimshony et al., 2016; Kalle et al., 2014; Picelli et al., 2013). Hence, optimizing these reactions requires substantial work, time and cost efforts (Bagnoli et al., 2018; Hagemann-Jensen et al., 2020). Therefore, some methods replace this typical pattern of reverse transcription and PCR amplification with linear amplifications using *in vitro* transcription. Although being supposedly less biased than PCR, this setup requires a second reverse transcription step (Hashimshony et al., 2016; Jaitin et al., 2014). In addition, PCR introduced biases and noise can be efficiently removed by integrating a molecular barcode or unique molecular identifier (UMI) already at the stage of cDNA conversion, enabling the removal of PCR duplicates computationally (Kivioja et al., 2012; Parekh et al., 2016). However, UMI integration is either performed at the 5` or 3` end of fragments, which precludes obtaining reads over the full gene body as one of the ends is enriched. Although all published scRNA-seq protocols require at least one reverse transcription step, several approaches are also possible in this reaction. Most methods use mRNA targeting oligo-dt primers in combination with reverse transcriptases derived from the moloney murine leukemia virus (MMLV), which are capable of performing template switching at the end of the RNA molecule and thereby introducing a second universal PCR handle using a template switching oligo (Zajac et al., 2013). Interestingly, the exact mechanism of this function was poorly understood until very recently (Wulf et al., 2019).

Finally, amplified cDNA needs to be converted to sequenceable libraries containing flowcell binding adaptors and sample barcodes. While most protocols use Illumina`s transposon based

tagmentation approach, improvements in enzymatic fragmentation followed by adaptor ligation are becoming more popular (Zheng et al., 2017).

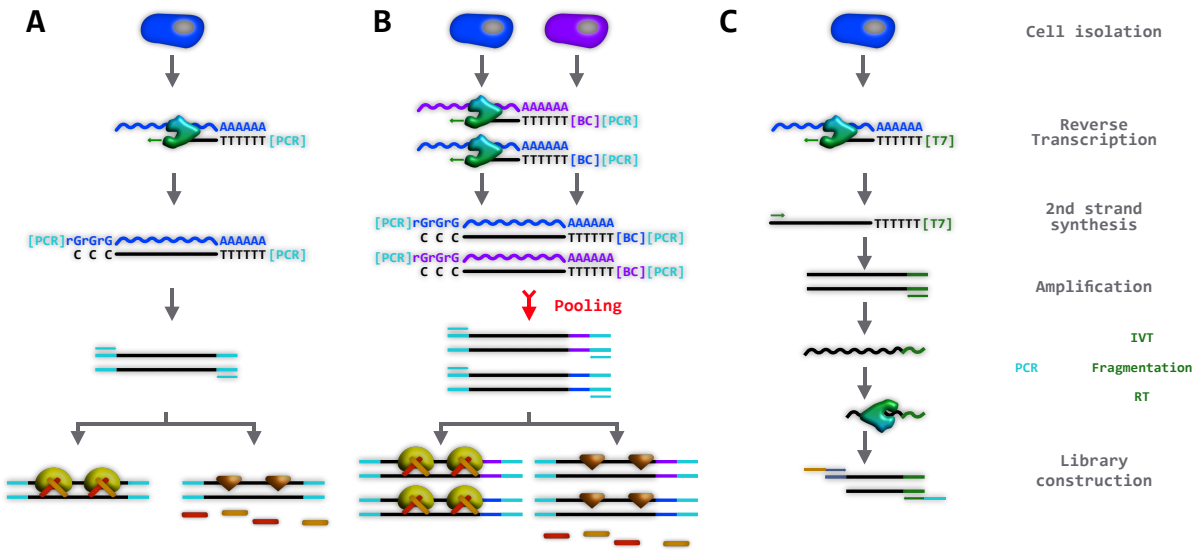


Figure 4: Common scRNA-seq workflows
After successful cell isolation, mRNA is reverse transcribed and 2nd strand synthesis is performed. The resulting cDNA is amplified separately (A) or pooled (B) either via PCR (A,B) or *in vitro* transcription (C). Final library construction involves fragmentation/tagmentation and adapter integration.

With over 50 different protocols published, scRNA-seq remains a fast evolving method. The continuous evolution in both molecular methods as well as the corresponding tools for analysis, tackle not only the major challenges described above but also the computational, statistical and biological limitations recent methods still need to overcome. This includes dealing with stochastic dropout events in gene detection (zero inflation), big data handling, sample preparation and isolation and costs (Ziegenhain et al., 2018). It becomes even more challenging with single cell RNA-seq being more and more applied ubiquitously across numerous fields of biology, facing new specific challenges and applications.

Acute Myeloid Leukemia

Pathophysiology of AML

Acute myeloid leukemia (AML) is the most common type of acute leukemias, accounting for one third of cases within adult patients and roughly one percent of all new cancer incidences. In general, AML is caused by an abnormal proliferation and differentiation of a clonal population of hematopoietic stem cells. Hence, the healthy maturation process of myeloid cells is disturbed, leading to a decrease in erythrocytes (Anemia) and platelets (Thrombocytopenia) and an increase in leukocytes (Leukocytosis), whereas the latter are non-functional myeloblasts. Without any treatment, patients typically die within several months, mostly due to secondary infections or bleeding caused by the impaired immune system (Albrecht, 2014; De Kouchkovsky and Abdul-Hay, 2016). Despite the extensive advances in cancer diagnostics and treatment over the last decades, prognosis for AML patients remains poor with 50-80% of patients dying after an initial successful treatment (Kantarjian, 2016). Especially elderly patients, which represent the majority of cases, show a very low survival rate, which is partially caused by a higher risk of treatment related mortality (TRM) (Meyers et al., 2013; Shah et al., 2013).

The underlying cause of the emergence of such transformations of a healthy hematopoietic stem cell relies on the acquisition of several mutations. These mutations permit a clone to overgrow other healthy HSCs and subsequently to fulfill other hallmarks of cancer. (Hanahan and Weinberg, 2000, 2011). In contrast to other cancers however, AML displays a very heterogeneous mutational pattern across patients and in most cases also lacks major chromosomal rearrangements and aberrations (Cancer Genome Atlas Research Network et al., 2013).

The heterogeneity of AML benefits from subclassifications

Due to the heterogeneity across patients, specific mutation patterns can only be observed in a small subset of patients. Therefore, linking them to their individual effect on prognosis and treatment success is highly restricted due to the low statistical power. (De Kouchkovsky and Abdul-Hay, 2016). However, a framework for classifying AML related mutations has been established in animal models. This two hit model of leukemogenesis relies on the simultaneous presence of mutations which lead to the activation of pro-proliferative pathways (class I, e.g. FLT3, K/NRAS, TP53) and an impairment of the normal hematopoietic differentiation (class II, e.g. NPM1 or CEBPA) (Gilliland and Griffin, 2002). In addition, a third class of mutations in epigenetic regulators, such as DNMT3A, TET2 and IDH-1/2 could be identified affecting cellular differentiation and proliferation. (Cancer Genome Atlas Research Network et al., 2013).

Nevertheless, the classification of AML related mutations into these three classes cannot explain the substantial variance in treatment success or relapse rates between patients. Another major entity contributing to this is the fraction of cells that acquired specific mutations within each tumour population, measured by variant allele frequencies (VAF). Mutations with a high variant allele frequency probably emerged very early on, possibly in the founder cell of the tumour, whereas mutations which can only be found in a subset of the tumour are more likely to have developed later on. Although a low variant allele frequency of a specific mutation might be initially interpreted as less important for the tumour development it can play an essential role in relapse formation and hence for prognosis (Döhner et al., 2017).

For example, it was shown that class II mutations are often not retained in the relapse and are therefore assumed to be unstable. On the other hand, DNMT3A and NPM1 mutations showed

high stabilities in general (Cocciardi et al., 2019). The loss or gain of mutations between the initial cancer population and the corresponding relapse can be explained by two different models. Either the primary clone is not completely eradicated by therapy and remains in small numbers in the patient, or an already resistant subclone is selected for during treatment due to its fitness advantage. During the second outgrowth, further mutations can lead to a treatment resistant relapse or complement a resistance clone with further aggressiveness (Ding et al., 2012) (Figure 5).

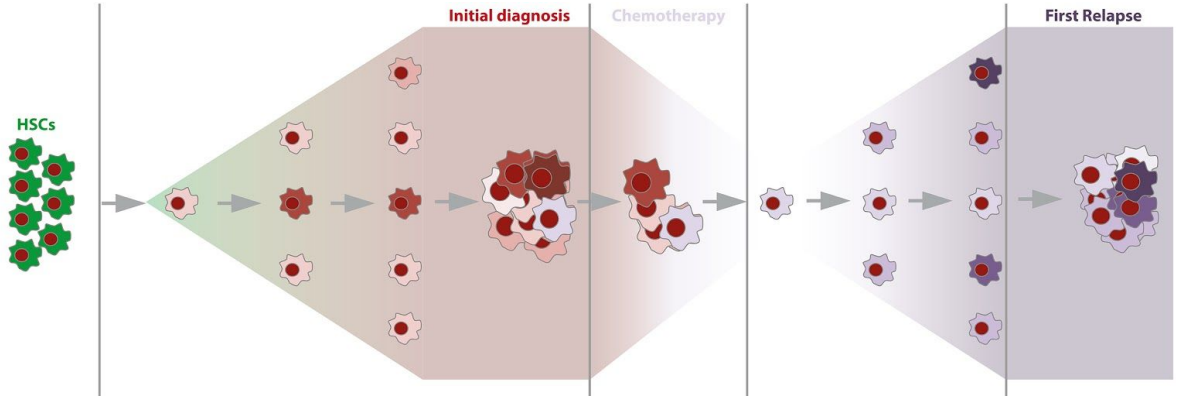


Figure 5: Schematic presentation of tumorigenesis and relapse formation in AML.
Colouring of cells represent cell types: green, healthy hematopoietic stem cell; red/purple, leukemic cells. Shadings represent various subclones of transformed leukemic cells.

Hence, in order to provide an optimal outcome for each patient, the combined forces of French, American and British haematologists proposed the FAB (French American British) System in 1976, classifying patients into different categories and treatment strategies. Based on morphologic and cytochemical characteristics of peripheral blood and bone-marrow films it describes eight AML subtypes (M0 through M7) (Bennett et al., 1976). This system remained the fundamental classification scheme until 2001, when a better understanding of molecular and disease biology caused the World Health Organization (WHO) to incorporate genetic, immunophenotypic, biological as well as clinical features to a new classification

scheme (World Health Organization, 2001)(Walter et al., 2013), (World Health Organization, 2001). Revised versions of this classification in 2008 and 2016 led to the definition of six different major disease entities, which are now generally adapted by clinicians (Arber et al., 2016; Vardiman et al., 2009).

To further guide the treatment of patients, AML cases can be arranged into 3 major risk groups with favourable, intermediate and adverse prognosis based on cytogenetic aberrations and gene mutations as described in the revised version of the diagnosis and management recommendations of the European AML network in 2017 (ELN2017) (Döhner et al., 2017). Not surprisingly, the effect of a mutation in a specific gene must always be considered in its mutational environment (De Kouchkovsky and Abdul-Hay, 2016)(Gale et al., 2008)(De Kouchkovsky and Abdul-Hay, 2016). Nevertheless, a mutation in TP53 (class I) remains the single worst genetic prognostic factor (Grossmann et al., 2012).

As heterogeneous as the disease itself, the possible treatments are numerous (Figure 6). Most patients undergo at least one round of induction therapy comprising 7 days of continuous infusion of cytarabine, a chemotherapy medication mainly used to treat leukemias, followed by 3 days of anthracycline, a common chemotherapeutic regimen (Wiernik et al., 1992). Consolidation therapy after achieving remission mostly involves either continuing chemotherapy with lower doses of cytarabine or allogeneic hematopoietic stem cell transplantation (allo-HSCT). Due to the high general toxicity and high risk of TRM of this treatment, especially with elderly patients, newly designed targeted therapies such as FLT3 inhibitors hold great promise for higher specificity and therefore lower toxicity in AML patients (Bcop et al., 2020; Estey, 2014; Ravandi et al., 2013; Röllig et al., 2015; Swaminathan et al., 2017; Wang et al., 2016). However, these specific therapies can only be effective for a subset of patients and do not present a general treatment scheme.

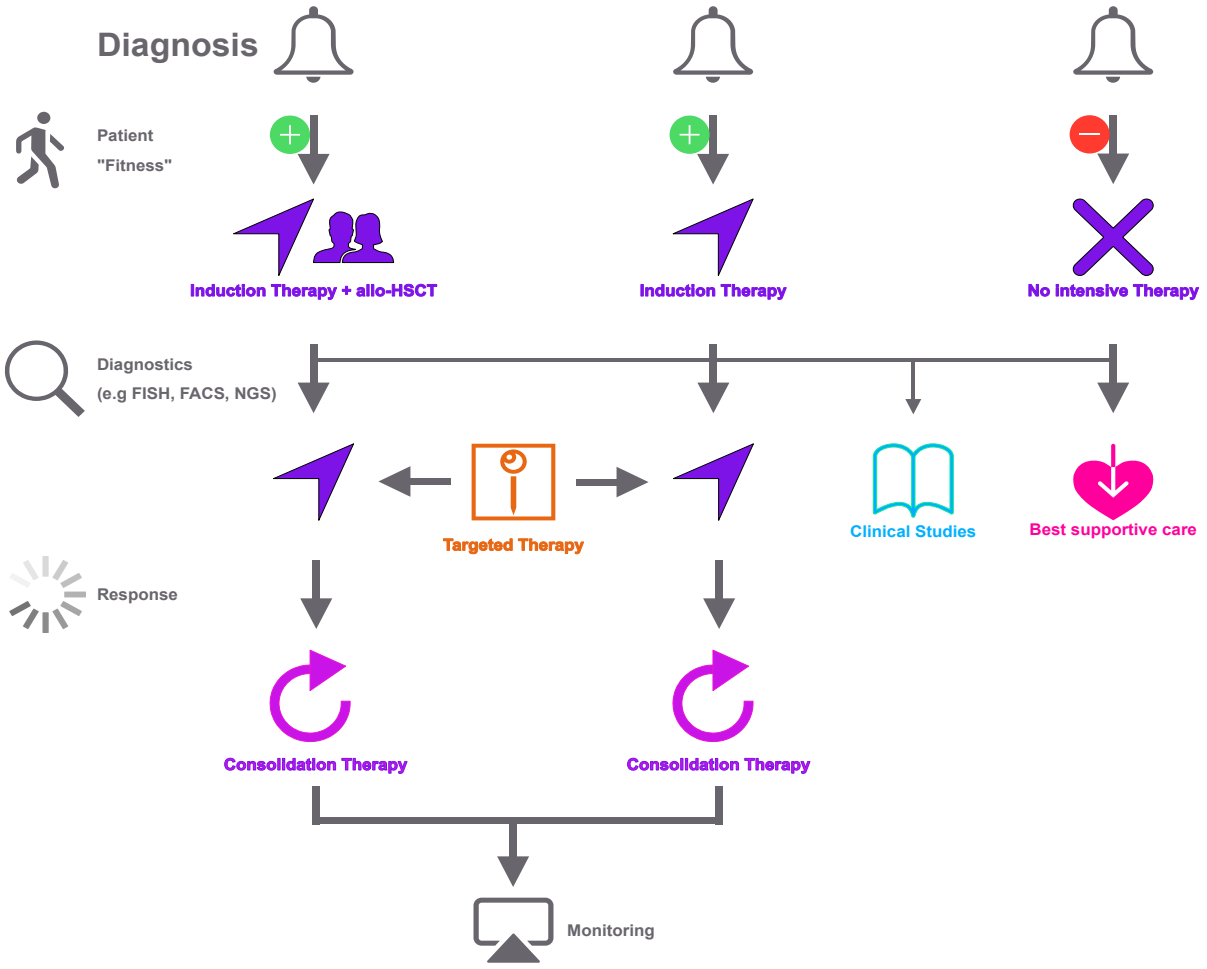


Figure 6: Schematic outline of possible treatments in AML
The treatment of AML is not standardized and requires the integration of clinical and molecular information in addition to constant monitoring and reevaluation by the physicians. The possible treatments include classical chemotherapy, allogeneic hematopoietic stem cell transplantation (allo-HSCT) and targeted therapies as well as yet not established treatments via inclusions into clinical studies.

Overall, the underlying biological and genetic heterogeneity requires future developments, not only in drug specificity and efficacy but also in a deeper understanding of the molecular processes involved in disease progression, relapse formation and treatment resistance. Refined diagnostic tools and better prognostic stratifications could lead to a step by step improvement to successfully treat AML patients (Herold et al., 2018).

Quantification of gene expression in AML

In the last 5 years multiple studies have shown the great potential of RNA expression analysis in helping to understand clinical and molecular features of AML and thereby unlock at least some of the hidden secrets.

As mentioned above, one of the major challenges of AML is the low rate of achieving complete remission (Walter et al., 2015a). Up to 50% of older and 20-30% of younger adult patients are refractory to induction therapy (Döhner et al., 2010). This treatment failure can be explained by a resistant disease (primary refractory AML). Earlier attempts to predict treatment response based on clinical but also genetic variables performed poorly (Krug et al., 2010; Walter et al., 2015a, 2015b). However, in 2018 a group of AML researchers from Munich provided a better estimator for treatment response by including RNA expression data (Herold et al., 2018).

Combining the cytogenetically defined MRC risk group assignment (Grimwade et al., 2010) as well as the expression values of 29 genes, the newly defined PS29MRC (predictive score 29 MRC) outperformed previous models, especially for patients above the age of 60. In addition, a high-risk patient group with a median survival of only eight months could be identified. This high risk group comprises about 20% of all intensively cared AML patients within the validation cohort of the study. Surprisingly, when comparing the ELN2017 classification of these patients, 14% were assigned an intermediate and 86% an adverse risk. This finding does not only illustrate potential improvements of the current risk classification in use, but also raises the question whether standard induction therapy is the right choice for this group of patients. Furthermore, the gene expression markers incorporated in the PS29MRC score itself could help to understand the poor success of therapy. For example, the two most powerful marker genes within PS29MRC, CYP2E1 and MIR155HG (hosting

miR-155), have both been already described in relation to induction treatment. CYP2E1 expression is a strong predictor of treatment response and is known to be involved in cytarabine metabolism (Iacobucci et al., 2013). On the other hand, high miR-155 expression was highly correlated with a refractory phenotype and is known to be upregulated in high risk CN-AML cases (Marcucci et al., 2013).

More recently, high throughput RNA sequencing in combination with Chromatin Immunoprecipitation-sequencing (ChiP-seq) revealed that changes in gene regulatory element (GRE) activity in relapsed AML patients can be linked to an expression signature enabling to predict relapse in these patients (Wiggers et al., 2019). However, in contrast to the PS29MRC score, it was found that prediction capabilities of different genes differ drastically between subgroups of patients. Using a weighted gene coexpression network analysis (WGCNA) 12 clusters of genes that share a similar expression across the cohorts sample could be identified. Six of these clusters were comprised of five or more relapse predictive genes. While most clusters showed no distinctive preference for FAB classified subcategories of patients, some indeed were highly predictive for only one specific FAB classification. This highlights that although RNA-expression can be a useful tool to predict relapse risk already at diagnosis, an overall pattern across all AML patients is not likely, further highlighting the need for a finer classification scheme of the disease heterogeneity (Wiggers et al., 2019).

In addition to the informative expression of genes, full length RNA-seq is capable of detecting differential isoform expression and alternative splicing (Wang et al., 2009). The effect of specific alternative splicing events is known to have a possible impact on treatment resistance (Sveen et al., 2016). For example, it has been shown that alternative splicing via skipping exon 12 leads to an inactive deoxycytidine kinase (dCK) protein which is

commonly found in patients with restinant AML (Veuger et al., 2000). Low expression or inactive splice forms enabled AML cells to cope with Cytarabine treatment, *in vitro* (Veuger et al., 2002). Recent studies have shown that aberrant splicing affects a major portion of expressed genes in AML samples and might play a key role in initial treatment resistance and relapse formation (Adamia et al., 2014; Li et al., 2014; Zhou and Chng, 2017).

A major advantage of RNA-seq is the possibility to perform expression quantification at the level of single cells. The field of single cell RNA sequencing (scRNA-seq) has been rapidly evolving over the last decade with numerous protocols published (Ziegenhain et al., 2017). Although comprising its own set of difficulties and challenges scRNA-seq holds the great promise to be able to finally understand the underlying heterogeneity of AML (Bagnoli et al., 2019; Ziegenhain et al., 2018). In addition to the interpatient heterogeneity, even within one specific AML subgroup, cancerous cells are known to show a variety of cell states. In a simplified manner, AML cell states can be seen as a mirror of the healthy hematopoietic cell hierarchy. Interestingly, it has been shown that some leukemic cells harbour a stem cell like phenotype, being quiescent and rare and are hypothesized to be able to retain the tumour (Pollyea and Jordan, 2017). Using high throughput scRNA-seq of healthy and AML bone marrow aspirations, in combination with short read (Illumina) and long read (Nanopore) sequencing of targeted RNA genotyping amplicons, it was shown that leukemic cells can be projected to cell types of healthy hematopoiesis. When comparing the ratios of these malignant cell types across 35 patients to clinical characteristics like morphology and surface phenotypes, the assignments were in good concordance. Interestingly, using a subset of genes differentiating between these leukemic cell types in the single cell data, publicly available bulk RNA-seq data of The Cancer Genome Atlas (TCGA) could be clustered into 7 distinct clusters, which showed a high correlation with their underlying genetic mutations.

Furthermore, primitive AML cells were found to up regulate genes involved in stress response, proliferation and self renewal, when compared to their healthy counterparts. However, gene signatures which could identify early stage hematopoietic cells (HSCs and GMPs) were coexpressed in primitive AML cells and patients with a high HSC like expression pattern showed a decreased survival (van Galen et al., 2019).

A similar but more efficient approach in combining single cell expression analysis with single cell RNA genotyping was recently established by using only commercially available 3` and 5` high throughput scRNA-sequencing kits, showing great promise for further research (Petti et al., 2019).

Although these studies have shown the potential of RNA expression analysis to further understand the disease and to improve prognosis accuracy, its broader application in AML research and diagnostic routine is often unfeasible. The complex interactions in which the disease evolves requires advanced model systems and large sample sizes. However, the biological samples needed to investigate AML are rare and very limited. Patient samples are especially hard to obtain as they are used for diagnostic procedures and require painful and risky bone marrow aspiration from the patient. While there are several leukemic cell line models, which can be cultured in a relatively easy manner, they fail to model outside influences via the immune system, different treatments or the formation of possible niches that play an important role for the underlying disease pathways. A possible system proposed to overcome this limitation are Patient Derived Xenograft (PDX) models, in which human leukemic cells derived from biopsy samples are cultured within immunodeficient mice (Vick et al., 2015). These models provide a much better environment, as niche interactions within the bone marrow of the mouse can be exploited by the AML cells and the effect of possible treatments can be investigated in a living organism. However, due to the missing intact

immune system as well as the murine environment , the transfer of observations in PDX models to human patients is difficult. In addition, they are very expensive and require substantial work to be propagated and thereby limiting sample availability.

Overall, bulk and single cell RNA-seq help to further understand the transformation of healthy hematopoietic stem cells, the progression and evolution of the cancerous cell population as well as the underlying causes of treatment resistance and relapse formation. However, special adaptations of these powerful methods are required to cope with the general restrictions in AML research.

Results

Applying bulk RNA-seq in AML

Plasticity in Growth Behavior of Patients' Acute Myeloid Leukemia Stem Cells Growing in Mice.

Published Ahead of Print on February 6, 2020, as doi:10.3324/haematol.2019.226282.

haematologica
Journal of The Ferrata Storti Foundation

Plasticity in growth behavior of patients' acute myeloid leukemia stem cells growing in mice

by Sarah Ebinger, Christina Zeller, Michela Carlet, Daniela Senft, Johannes W. Bagnoli, Wen-Hsin Liu, Maja Rothenberg-Thurley, Wolfgang Enard, Klaus H. Metzeler, Tobias Herold, Karsten Spiekermann, Binje Vick, and Irmela Jeremias

Haematologica 2020 [Epub ahead of print]

Citation: Sarah Ebinger, Christina Zeller, Michela Carlet, Daniela Senft, Johannes W. Bagnoli, Wen-Hsin Liu, Maja Rothenberg-Thurley, Wolfgang Enard, Klaus H. Metzeler, Tobias Herold, Karsten Spiekermann, Binje Vick, and Irmela Jeremias.
Plasticity in growth behavior of patients' acute myeloid leukemia stem cells growing in mice.
Haematologica. 2020; 105:xxx
doi:10.3324/haematol.2019.226282

Plasticity in growth behavior of patients' acute myeloid leukemia stem cells growing in mice

Running Title: Plasticity of low-cycling AML PDX cells

Sarah Ebinger[1], Christina Zeller[1], Michela Carlet[1], Daniela Senft[1], Johannes W. Bagnoli[2], Wen-Hsin Liu[1], Maja Rothenberg-Thurley[3], Wolfgang Enard[2], Klaus H. Metzeler[3-5], Tobias Herold[1,3,4], Karsten Spiekermann[3-5], Binje Vick[1,4], Irmela Jeremias[1,4,6]

1 Research Unit Apoptosis in Hematopoietic Stem Cells, Helmholtz Zentrum München, German Research Center for Environmental Health (HMGU), Munich, Germany
2 Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Martinsried, Germany
3 Laboratory for Leukemia Diagnostics, Department of Medicine III, University Hospital, LMU Munich, Munich, Germany
4 German Cancer Consortium (DKTK), partner site Munich, Germany
5 German Cancer Research Center (DKFZ), Heidelberg, Germany
6 Department of Pediatrics, Dr. von Hauner Childrens Hospital, Ludwig Maximilian University, Munich, Germany

corresponding author
Irmela Jeremias
Helmholtz Zentrum München
Marchioninistrasse 25
81377 Munich, Germany
Phone: +49-89-3187-1424; Fax: +49-89-3187-4225;
E-mail: Irmela.Jeremias@helmholtz-muenchen.de

Main Text:

Resistance against chemotherapy remains a major obstacle in treating patients with acute myeloid leukemia (AML).(1) Novel therapeutic concepts are especially desired to target and eliminate resistant AML stem cells. Here we show that AML stem cells harbor plasticity, a changing pattern of biological behaviour, by demonstrating that AML stem cells reversibly switch from a low-cycling, chemotherapy resistant state into an actively proliferating state associated with response to standard chemotherapy.

We used patient-derived xenograft (PDX) cells from patients with high risk or relapsed AML that were lentivirally transduced for marker expression. We stained these cells with the proliferation-sensitive dye Carboxyfluorescein succinimidyl ester (CFSE), and re-transplanted them into next-recipient mice. A rare subpopulation of AML cells displayed reduced proliferation *in vivo*, associated with resistance against standard chemotherapy. The proportion of AML cells with stem cell potential was identical in both, the high and low proliferative sub-fractions. In re-transplantation experiments, proliferation behavior proved reversible, and AML stem cells were able to switch between a high and low proliferation state. Our data indicate that AML stem cells display functional plasticity *in vivo,* which might be exploited for therapeutic purposes, to prevent AML relapse and ultimately improve the prognosis of patients with AML.

AML patients are at risk to suffer disease relapse associated with dismal prognosis. The rare subpopulation of AML stem cells (or LIC for leukemia initiating cells) might be responsible for relapse by combining self-renewal capacity with dormancy and resistance against standard chemotherapy.(2) AML LIC features, i.e. growth phenotype have long been considered mainly constant and persistent (2-5); in contrast, recent data suggest unsteady features under therapeutic pressure (6, 7), while data without experimental treatment pressure remain elusive. Putative functional plasticity of AML LIC is of major clinical importance as it might enable novel therapeutic options.

We previously reported functional plasticity in acute lymphoblastic leukemia (ALL), where we showed *in vivo* that long-term dormant, treatment-resistant ALL cells were able to convert into highly proliferative, treatment-sensitive cells and vice versa (8).

Nevertheless, AML and ALL differ widely regarding stem cell biology and a defined stem cell hierarchy - characteristic for AML - was never proven in ALL. Based on diverse stem cell characteristics, we considered functional plasticity of LIC conceivable in ALL, but hypothesized its absence in AML.

To test our hypothesis, we studied cells from ten patients with high-risk or relapsed AML of different karyotypes, genotypes and clinical histories (Table S1). As a clinic-close model system, primary cells were transplanted into immunocompromised mice, and AML patient-derived xenograft (PDX) models were established.(9) PDX models were selected to allow for serial transplantation; as this ability is restricted to highly aggressive disease, our study is biased towards high risk AML. AML PDX models were genetically engineered to express luciferase for bioluminescence *in vivo* imaging and mCherry for cell enrichment by flow cytometry. Marker expression remained stable over serial re-transplantation and allowed enrichment of minute numbers of PDX AML cells from murine bone marrow (Figure 1A, detailed in supplemental methods). As controls, three samples (AML-356, AML-358 and AML-538) were studied without prior genetic engineering.
AML PDX samples showed more than three-fold differences in doubling times *in vivo*, resulting in variable time to overt disease in mice (Figure S1AB). When PDX cells were re-isolated from murine bone marrow, mCherry expression enabled unbiased enrichment of AML PDX cells, independent of other, putatively subpopulation-restricted, surface markers on AML cells (Figure 1B).(8, 10) Re-isolation of PDX cells revealed that homing was heterogeneous between samples, as 0.01 to 1% of PDX cells could be re-isolated from mice early after transplantation (Figure S1C). The frequency of LIC, as determined in limiting dilution transplantation assays, varied by a factor of 10 between samples (Figure S1D, Table S2). Thus, our AML PDX cohort of aggressive samples displayed major functional inter-sample heterogeneity *in vivo*, reflecting the known phenotypic heterogeneity of AML.(11)

To track *in vivo* proliferation of AML cells from individual samples, PDX cells were stained with CFSE, a dye that is not metabolized in eukaryotic cells, but decreases upon cell divisions, indicating proliferation.(12) CFSE records a cell's proliferative history rather than providing a snapshot of the cell´s proliferative state at a given

moment. CFSE content was measured by flow cytometry at different time points following injection into groups of mice.

In accordance with an increase in leukemic burden and numbers of re-isolated cells (Figures 1CD and S2A), most AML PDX cells entirely lost CFSE within days of *in vivo* growth, indicating high proliferative activity in the majority of cells (Figures 1EF and S2A). However, a minor subpopulation of cells retained CFSE over several weeks, indicating a low-cycling, putatively dormant phenotype (Figures 1EF and S2A). We called these cells label-retaining cells (LRC) according to literature (8). LRC were found in 9/10 samples tested (Figures 1EF and S2AB). Only a single sample originating from child with fatal AML relapse had entirely lost the LRC population between day 7 to 15 (Figure S2C), again highlighting the known heterogeneity of AML.(11) Cell cycle analysis confirmed that LRC divide less as compared to non-LRC (Figure S3A). Together, our data reveal, in the majority of cases, heterogeneity of *in vivo* growth behavior within individual AML PDX samples, including a subpopulation of low-cycling LRC. Hence, our results add an important level of phenotypic heterogeneity to AML on top of the known heterogeneity of e.g. immunophenotypes, or gene expression profiles. As a large range of AML subtypes were studied (Table S1), the novel characteristic is not limited to a specific cytogenetic or genetic subgroup.

To further characterize attributes of LRC, gene expression analysis of 24 LRC and non-LRC samples isolated from AML-393 and AML-491 was performed.(13) Among the top down regulated gene sets in LRC were cell cycle regulators, confirming the reduced proliferative state of these cells (Figure S3BC); among the top upregulated gene sets were cell adhesion molecules (Figure S3C). Notably, LRC of AML-393 were more similar to LRC of AML-491 than to their own non-LRCs (Figure 1G), despite the substantial differences in the mutational profile of AML-393 and AML-491 (Table S1). Even more striking, gene-set enrichment analysis identified a high overlap of significantly deregulated genes between AML LRC and our previously defined LRC signature in ALL (8) (Figure 1H and S3C), suggesting comparable biologic processes activated in LRC of both, AML and ALL.

Given the long-known link between dormancy and chemo-resistance(14), we compared drug response between low-cycling LRC and high-cycling non-LRC.

Groups of mice engrafted with CFSE-labeled cells were treated with a chemotherapeutic regimen mimicking “7+3” induction therapy(1), consisting of cytarabine and liposomal daunorubicin (DaunoXome) (Figure 2A). *In vivo* treatment diminished tumor burden as monitored by *in vivo* imaging (Figure 2B), resulting in a decrease of total isolated PDX cells by at least one order of magnitude (Figure 2C). Interestingly, while non-LRC were strongly reduced by treatment, even to undetectable levels in some mice (Figures 2DE), low-cycling LRC revealed decreased sensitivity towards systemic treatment in all samples tested. As net effect, the relative proportion of LRC was significantly enriched among cells surviving after treatment in 3 of 4 samples (Figures 2DF). Thus, low-cycling LRC show increased resistance against conventional chemotherapy *in vivo* compared to high-cycling non-LRC.

We next asked whether LRC and non-LRC differ in their ability to form tumors and performed re-transplantation experiments. Low numbers of sorted LRC and non-LRC were re-injected into secondary recipient mice in limiting dilutions close to sample-specific LIC frequency (Figures 3A and S4A). Interestingly, both, LRC and non-LRC gave rise to leukemia upon re-transplantation, indicating both subpopulations contained LIC (Figures 3B and S4B). As leukemia development was highly similar in mice transplanted with either LRC or non-LRC, low-cycling LRC must have converted into an actively proliferative state. Furthermore, we found similar LIC frequencies in LRC and non-LRC (Figure 3C, S4C and Table S3), and no difference in $CD34^{+}CD38^{-}$ cells between the two groups (Figure S5), strengthening previous findings.(15) Notably, $CD34^{+}CD38^{-}$ cells were barely detectable in the aggressive AML-393 sample, despite high LIC frequency (Figure S5, Table S2). These data indicate that LIC reside not only in the low-cycling LRC, but also in the high-cycling non-LRC compartment, indicating heterogeneity in proliferation dynamics within the AML LIC pool.

As low-cycling cells were able to convert to active proliferation, we asked whether the switch could also occur vice versa. To test whether LRC could be replenished from non-LRC, we re-transplanted high cell numbers of non-LRC restained with CFSE (Figure 3D and S4D). Upon secondary transplantation, non-LRC gave rise to a clear LRC fraction, comparable to the one from bulk cells at first transplantation, even at late time points (Figures 3EF and S4EF), indicating that high-cycling cells converted to a low-cycling phenotype.

These experiments revealed major functional plasticity of AML LIC phenotypes, and the ability to change their proliferation rate upon changes in external stimuli, such as re-transplantation.

Taken together, our data shows that low proliferation or dormancy characterizes a temporary, reversible cell state rather than a defined subpopulation of cells. AML contains a rare fraction of low-cycling, chemo-resistant LIC which are functionally plastic; AML LIC might temporarily adopt a low-cycling LRC phenotype or switch to a rapidly proliferating non-LRC phenotype, triggered by external stimuli such as re-transplantation. Even the highly aggressive AML samples used in this study harbor the potential to adopt a proliferative phenotype associated with response to standard chemotherapy.

Unexpectedly, we detected similar functional plasticity in AML as previously observed in ALL (8). This was accompanied by similar changes in gene expression profiles, although both diseases differ substantially regarding their stem cell biology as ALL never revealed a stem cell hierarchy as proven in AML. In contrast to ALL, AML plasticity comes as a major surprise, as we show here that high-cycling cells harbor the potential to convert into low-cycling cells, while both populations retain stem cell capacities. In our experiments, neither functionally nor immunophenotypically defined LIC were enriched in the LRC fraction, suggesting that dormancy and stemness are not consistently linked in AML, but that dormancy characterizes a temporary cell state rather than a defined subpopulation of cells. In addition to the known constant, presumably deterministic factors defining stemness, AML stem cells appear to be regulated by additional, transient and putatively stochastic factors.(16)

Our data indicates that stemness and resistance to anti-leukemic therapy is not strictly linked in AML. This opens exciting therapeutic potential to prevent relapse and strongly support the concept that recruiting AML LIC from their low-cycling phenotype into proliferation might sensitize them towards, e.g., conventional chemotherapy.(3, 4, 7) Taking advantage of the discovered heterogeneity and reversibility of the low- and high-cycling phenotypes implicates the need to identify factors responsible for AML plasticity, in addition to known microenvironment-derived regulators such as G-CSF.(2) The detected similarity in transcriptome signature between LRC of AML and

ALL might aid identifying factors that regulate these processes in both diseases. As attractive therapeutic concept, inhibition of the reversible low-cycling state might enable overcoming treatment resistance, remove AML LIC, prevent relapse, and ultimately increase patients' prognosis.

Acknowledgments

We thank Liliana Mura, Fabian Klein, Maike Fritschle, Annette Frank and Miriam Krekel for excellent technical assistance; Markus Brielmeier and team (Research Unit Comparative Medicine, Helmholtz Zentrum München) for animal care services; Andreas Beyerlein (Core Facility Statistical Consulting, Institute of Computational Biology, Helmholtz Zentrum München) for assisting with statistical analysis of treatment studies; Helmut Blum and Stefan Krebs (Laboratory for Functional Genome Analysis, Gene Center, LMU Munich) for sequencing, Claudia Baldus and Lorenz Bastian (Divison of Hematology and Oncology, Charité Universitätsmedizin Berlin, Germany) for kindly providing primary cells of AML-538, and Maya C. André and Martin Ebinger (Department of Pediatric Hematology/Oncology, University Children's Hospital Tuebingen) for kindly providing pediatric AML PDX samples.

Funding: The work was supported by grants from the European Research Council Consolidator Grant 681524; a Mildred Scheel Professorship by German Cancer Aid; German Research Foundation (DFG) Collaborative Research Center 1243 "Genetic and Epigenetic Evolution of Hematopoietic Neoplasms", projects A05, A06 (to KHM), A07 (to KS) and A14 (to JWB and WE), and associate member (TH); DFG proposal MA 1876/13-1; Bettina Bräu Stiftung and Dr. Helmut Legerlotz Stiftung (all to IJ, if not indicated differently). This work was further supported by the Physician Scientists Grant (G-509200-004) from the Helmholtz Zentrum München to T.H.

Conflict of Interest Disclosures

The authors declare that they have no conflict of interest.

References

1. Dohner H, Weisdorf DJ, Bloomfield CD. Acute Myeloid Leukemia. N Engl J Med. 2015;373(12):1136-1152.

2. Thomas D, Majeti R. Biology and relevance of human acute myeloid leukemia stem cells. Blood. 2017;129(12):1577-1585.

3. Ishikawa F, Yoshida S, Saito Y, et al. Chemotherapy-resistant human AML stem cells home to and engraft within the bone-marrow endosteal region. Nat Biotechnol. 2007;25(11):1315-1321.

4. Saito Y, Uchida N, Tanaka S, et al. Induction of cell cycle entry eliminates human leukemia stem cells in a mouse model of AML. Nat Biotechnol. 2010;28(3):275-280.

5. Hope KJ, Jin L, Dick JE. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. Nat Immunol. 2004;5(7):738-743.

6. Farge T, Saland E, de Toni F, et al. Chemotherapy-Resistant Human Acute Myeloid Leukemia Cells Are Not Enriched for Leukemic Stem Cells but Require Oxidative Metabolism. Cancer Discov. 2017;7(7):716-735.

7. Boyd AL, Aslostovar L, Reid J, et al. Identification of Chemotherapy-Induced Leukemic-Regenerating Cells Reveals a Transient Vulnerability of Human AML Recurrence. Cancer Cell. 2018;34(3):483-498.e5.

8. Ebinger S, Ozdemir EZ, Ziegenhain C, et al. Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia. Cancer Cell. 2016;30(6):849-862.

9. Vick B, Rothenberg M, Sandhofer N, et al. An advanced preclinical mouse model for acute myeloid leukemia using patients' cells of various genetic subgroups and in vivo bioluminescence imaging. PLoS One. 2015;10(3):e0120925.

10. de Boer B, Prick J, Pruis MG, et al. Prospective Isolation and Characterization of Genetically and Functionally Distinct AML Subclones. Cancer Cell. 2018;34(4):674-689.

11. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood. 2016;127(20):2391-2405.

12. Takizawa H, Regoes RR, Boddupalli CS, Bonhoeffer S, Manz MG. Dynamic variation in cycling of hematopoietic stem cells in steady state and inflammation. J Exp Med. 2011;208(2):273-284.

13. Bagnoli JW, Ziegenhain C, Janjic A, et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. Nat Commun. 2018;9(1):2937.

14. Cheung WH, Rai KR, Sawitsky A. Characteristics of cell proliferation in acute leukemia. Cancer Res. 1972;32(5):939-942.

15. Griessinger E, Vargaftig J, Horswell S, Taussig DC, Gribben J, Bonnet D. Acute myeloid leukemia xenograft success prediction: Saving time. Exp Hematol. 2018;59:66-71.

16. Gupta PB, Fillmore CM, Jiang G, et al. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. Cell. 2011;146(4):633-644.

Figure Legends

Figure 1 AML PDX cells contain a rare subpopulation of low-cycling cells

A Experimental procedure; primary patients' AML cells were transplanted into NSG mice, resulting PDX cells were genetically engineered, sorted, and amplified. At advanced disease stage, $mCherry^+$ AML PDX cells were isolated, stained with CFSE, and re-transplanted. At different time points, AML cells were re-isolated from mouse bone marrow, enriched, and CFSE content measured by flow cytometry, to detect CFSE-positive, low-cycling label-retaining cells (LRC), and CFSE-negative, proliferating non-LRC (nLRC). NSG: NOD.Cg-*Prkdcscid Il2rgtm1Wjl*/SzJ; EF1α: elongation factor 1-alpha promoter; Luc: enhanced firefly luciferase.

B Gating strategy; bone marrow cells depleted of murine cells by MACS were gated on (i) leukocytes, (ii) $DAPI^-$ $mCherry^+$ AML PDX cells, and (iii) separated into LRC and non-LRC according to their CFSE content. Maximum CFSE MFI was measured at day two after cell injection or *in vitro* cultivation, and divided by factor two to model cell divisions (dotted lines); upon less than three divisions, cells were considered as low-cycling LRC, upon more than seven divisions as proliferating non-LRC; days indicate time after cell injection.

C,D Growth of AML-393 cells monitored by *in vivo* imaging (**C**) or by quantifying PDX cells re-isolated from mouse bone marrow using flow cytometry (n=21) (**D**); each square represents data from one mouse.

E,F A rare subpopulation of AML PDX cells retains CFSE upon prolonged *in vivo* growth. AML-393 cells from different time points in **D** were analyzed by flow cytometry for CFSE using the gating strategy described in **B**; representative dot plots (**E**) and percentage of LRC cells among all isolated PDX cells are shown (**F**); each square represents data of one mouse.

G,H Gene expression analysis of LRC and non-LRC. LRC and non-LRC were isolated from mice carrying AML-393 (n=4) or AML-491 (n=4) ten or fourteen days after cell injection, respectively and subjected to RNA sequencing. Technical replicates were analyzed in 6 of 8 samples, resulting in a total of 24 samples analyzed.

G Heatmap of top differentially regulated genes (false discovery rate [FDR] ≤0.05) between LRC (green) and nLRC (black) of AML-393 and AML-491.

H LRC of AML-393 and AML-491 show significant overlap with the previously published LRC signature of acute lymphoblastic leukemia (ALL).

See supplemental Figure **S1** and **S2** for additional data.

Figure 2 Low-cycling AML PDX cells are treatment resistant *in vivo*

A Experimental procedure; groups of mice were injected with CFSE labeled AML-PDX cells and treated with PBS (ctrl.) or a combination of 20 mg/kg DaunoXome® (DNX) on day 7 and 150 mg/kg cytarabine (Ara-C) on days 7 to 9; PDX cells were re-isolated from murine bone marrow on day 10 and analyzed as described in Figure **1B**.

B Tumor load was monitored by *in vivo* imaging in AML-393.

C Total number of isolated PDX cells is shown of control and treated mice as mean+/-SD of AML-393 (n=8), AML-491 (n=6), AML-372 (n=10) and AML-388 (n=7) (**C**); each dot/square represents one mouse.

D, Representative dot plots (AML-393)

E, F Absolute number (**E**) and percentage (**F**) of non-LRC and LRC among all isolated PDX cells are shown from the same mice as in **C**; Log2 fold reduction for each subpopulation is displayed.

* $p<0.05$

Figure 3 AML PDX cells display reversible growth behavior, independently from stemness potential

A Experimental procedure; AML-393 cells were isolated from advanced disease donor mice (n=5 in three independent experiments), labeled with CFSE, and re-transplanted into first recipient mice. Ten days after injection, cells were re-isolated and sorted into LRC and non-LRC (nLRC) using the gates as described in Figure **1B** and re-injected into secondary recipient mice.

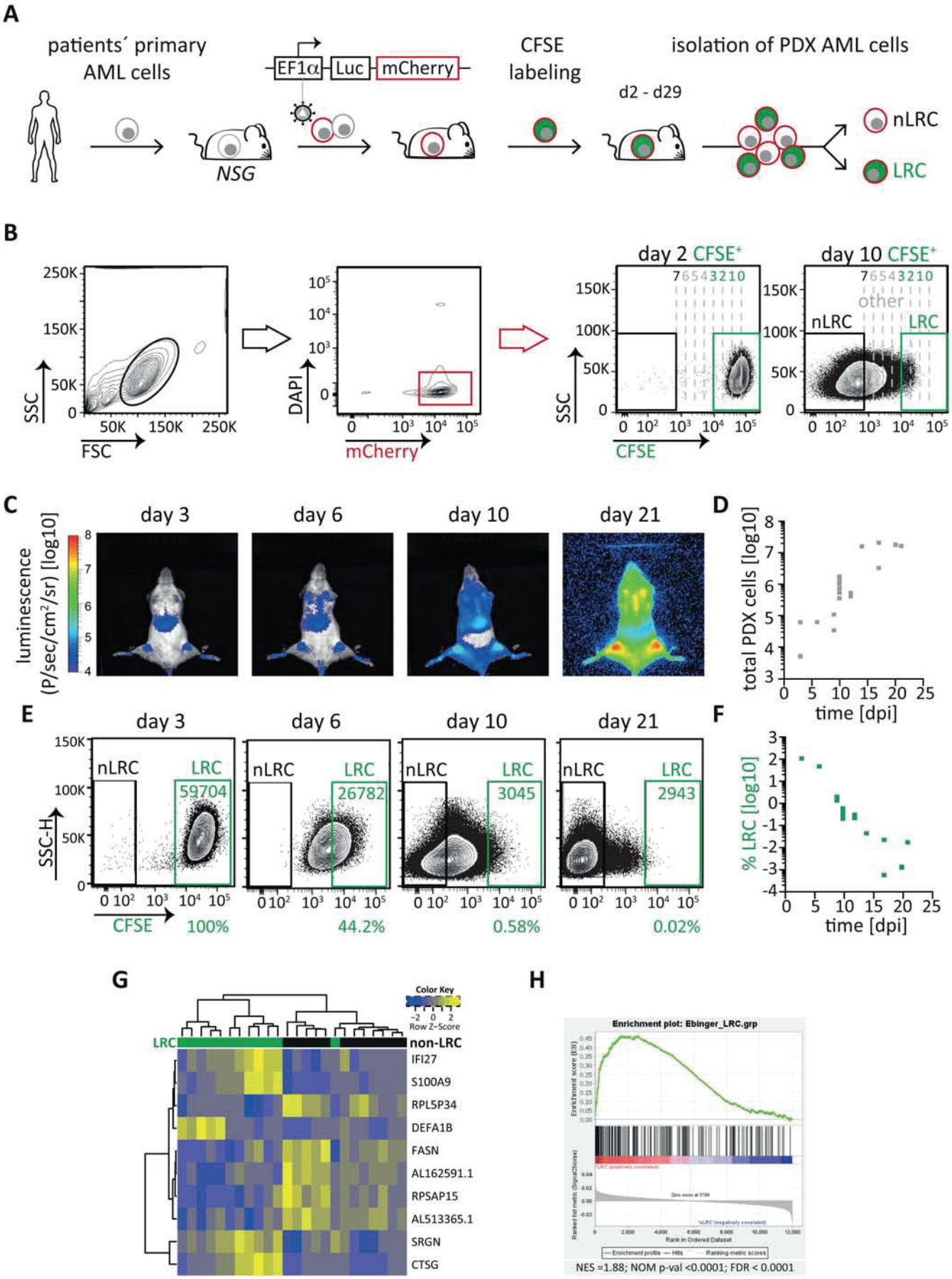
B Secondary recipient mice receiving either 300 LRC or 300 non-LRC (n=5) were monitored by *in vivo* imaging.

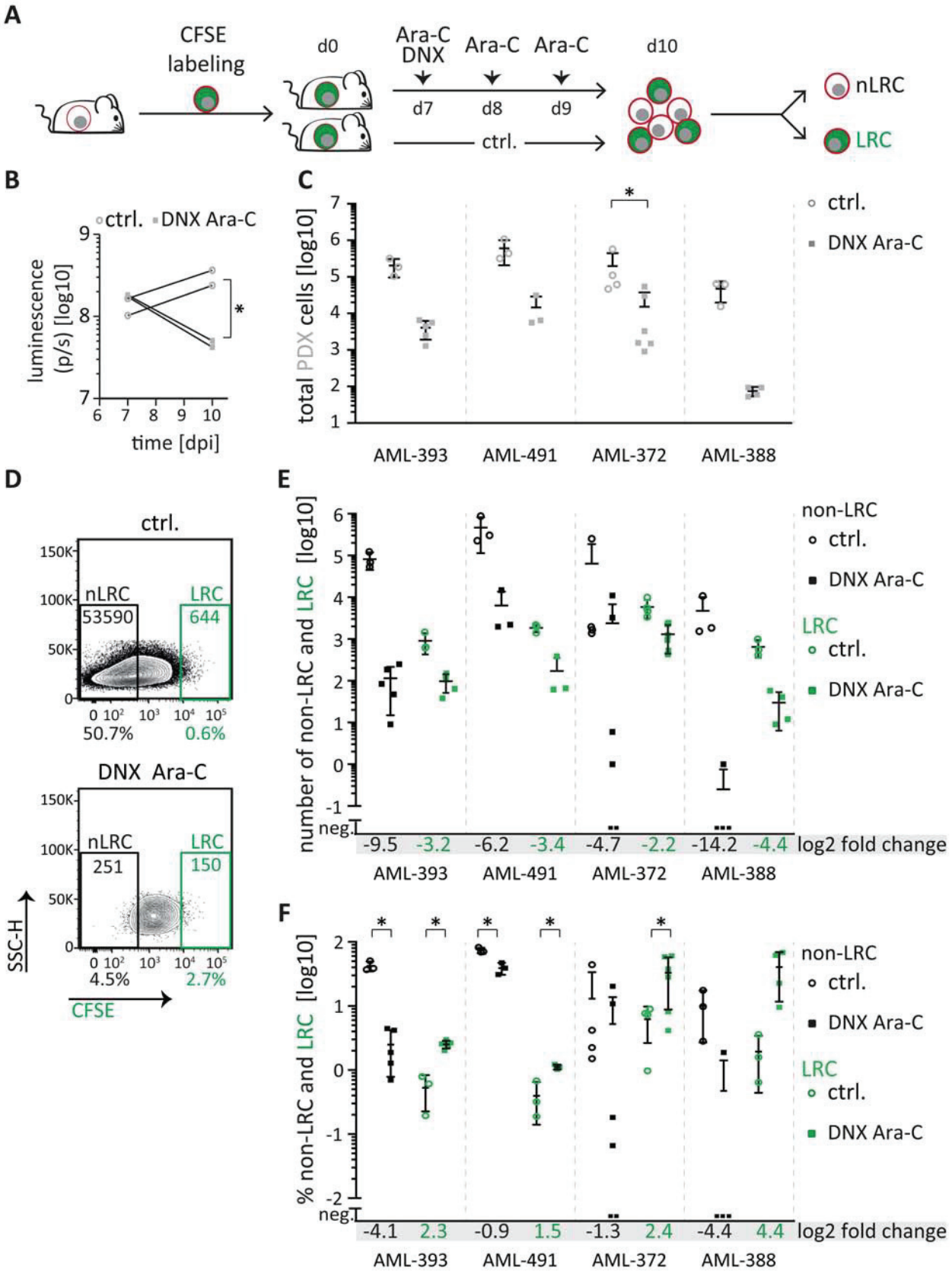
C LRC and non-LRC were re-injected into secondary recipient mice (n=38) in limiting dilutions at numbers indicated in Table **S3**. Positive engraftment of PDX cells was determined by *in vivo* imaging and/or flow cytometry. LIC frequency was calculated using the ELDA software and is depicted +/- 95% confidence interval. No statistically significant difference between LIC frequency of LRC and non-LRC was found according to chi-square test (p=0.0638).

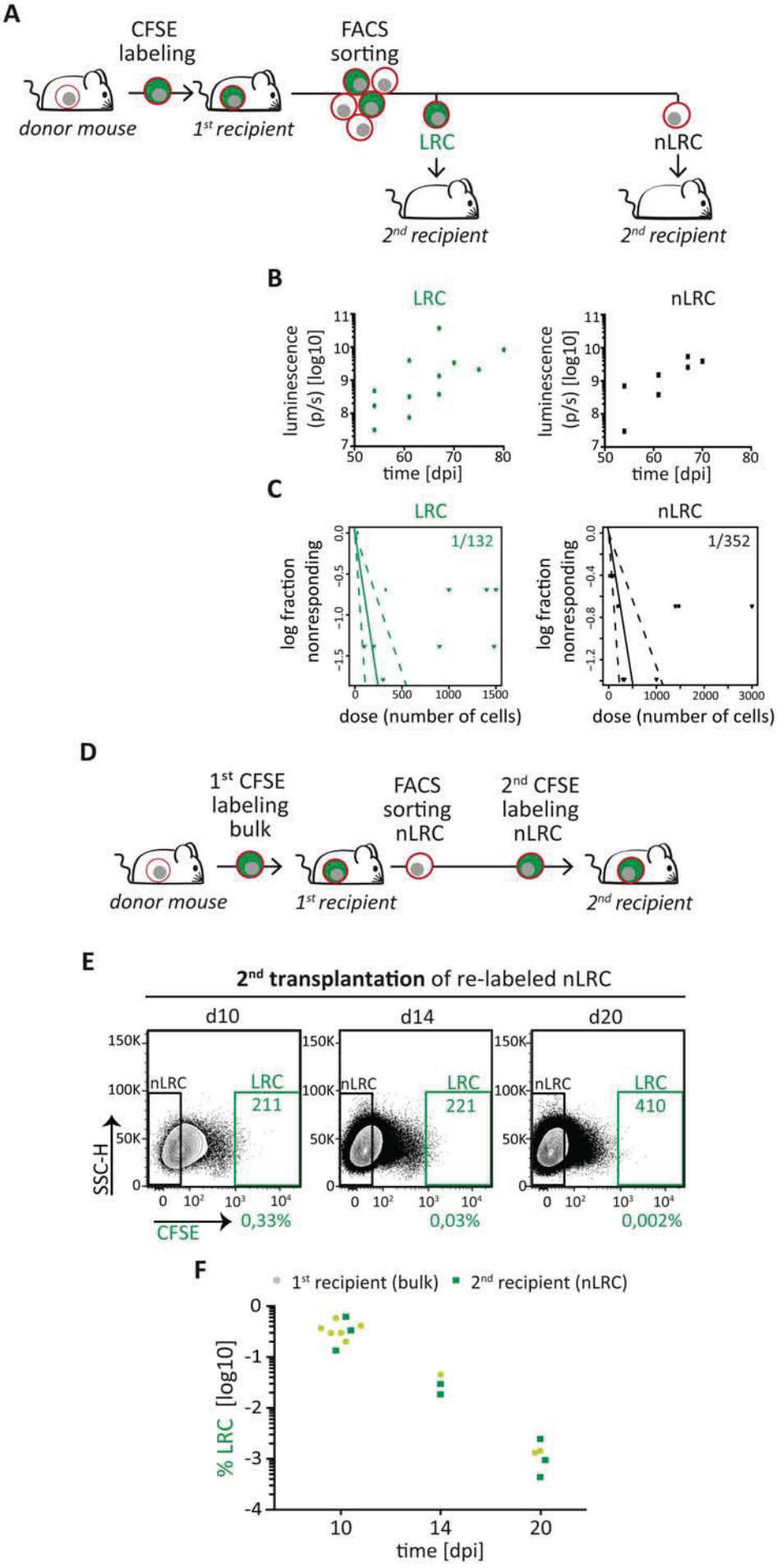
D Experimental procedure; from first recipient mice (n=2 in 2 independent experiments) harboring CFSE stained cells, non-LRC were isolated at day 21, re-stained with CFSE and $3.6*10^6$ cells were injected into secondary recipients (n=8); cells were re-isolated 10, 14 and 20 days later, and LRC were quantified using gates as described in Figure **1B**. The experiment is technically unfeasible for LRC as the high number of cells needed cannot be generated.

E, F Representative dot plots (**E**) and quantification (**F**) of the percentage of LRC among all PDX cells isolated from secondary recipients is displayed (dark green squares). LRC of first recipient mice as determined in Figure **1E** are shown for comparison (light green dots).

See supplemental Figures **S4** and Table **S3** for additional data.


A
patients´ primary AML cells
EF1α
Luc
mCherry
NSG
CFSE labeling
d2 - d29
isolation of PDX AML cells
nLRC
LRC
B
SSC
FSC
DAPI
mCherry
day 2 CFSE+
day 10 CFSE+
other
CFSE
C
day 3
day 6
day 10
day 21
luminescence (P/sec/cm²/sr) [log10]
D
total PDX cells [log10]
time [dpi]
E
nLRC
LRC
59704
26782
3045
2943
SSC-H
CFSE
100%
44.2%
0.58%
0.02%
F
% LRC [log10]
time [dpi]
G
Color Key
Row Z-Score
LRC
non-LRC
IFI27
S100A9
RPL5P34
DEFA1B
FASN
AL162591.1
RPSAP15
AL513365.1
SRGN
CTSG
H
Enrichment plot: Ebinger_LRC.grp
NES =1.88; NOM p-val <0.0001; FDR < 0.0001


A
CFSE labeling
d0
Ara-C DNX
Ara-C
Ara-C
d10
d7
d8
d9
ctrl.
nLRC
LRC
B
ctrl.
DNX Ara-C
luminescence (p/s) [log10]
time [dpi]
C
total PDX cells [log10]
ctrl.
DNX Ara-C
AML-393
AML-491
AML-372
AML-388
D
ctrl.
nLRC
53590
LRC
644
50.7%
0.6%
DNX Ara-C
nLRC
251
LRC
150
4.5%
2.7%
SSC-H
CFSE
E
number of non-LRC and LRC [log10]
non-LRC
ctrl.
DNX Ara-C
LRC
ctrl.
DNX Ara-C
neg.
-9.5 -3.2 -6.2 -3.4 -4.7 -2.2 -14.2 -4.4 log2 fold change
AML-393
AML-491
AML-372
AML-388
F
% non-LRC and LRC [log10]
non-LRC
ctrl.
DNX Ara-C
LRC
ctrl.
DNX Ara-C
neg.
-4.1 2.3 -0.9 1.5 -1.3 2.4 -4.4 4.4 log2 fold change
AML-393
AML-491
AML-372
AML-388


A
CFSE labeling
FACS sorting
donor mouse
1st recipient
LRC
nLRC
2nd recipient
2nd recipient
B
LRC
nLRC
luminescence (p/s) [log10]
time [dpi]
C
LRC
nLRC
1/132
1/352
log fraction nonresponding
dose (number of cells)
D
1st CFSE labeling bulk
FACS sorting nLRC
2nd CFSE labeling nLRC
donor mouse
1st recipient
2nd recipient
E
2nd transplantation of re-labeled nLRC
d10
d14
d20
nLRC
LRC
211
221
410
SSC-H
CFSE
0,33%
0,03%
0,002%
F
1st recipient (bulk)
2nd recipient (nLRC)
% LRC [log10]
time [dpi]

Plasticity in growth behavior of patients' acute myeloid leukemia stem cells growing in mice

Sarah Ebinger[1], Christina Zeller[1], Michela Carlet[1], Daniela Senft[1], Johannes Bagnoli[2], Wen-Hsin Liu[1], Maja Rothenberg-Thurley[3], Wolfgang Enard[2], Klaus H. Metzeler[3-5], Tobias Herold[1,3-5], Karsten Spiekermann[3-5], Binje Vick[1,4], Irmela Jeremias[1,4,5]

Supplemental Information

This pdf file contains:

Supplemental Methods and References
Supplemental Tables (3)
Supplemental Figures (5)

Supplemental Methods

Patients' acute myeloid leukemia (AML) cells

Bone marrow (BM) or peripheral blood (PB) samples from adult AML patients were obtained from the Department of Internal Medicine III, Ludwig-Maximilians-Universität, Munich, Germany, during the years 2012 and 2014. Specimens were collected for diagnostic purposes before start of treatment. Written informed consent was obtained from all patients. The study was performed in accordance with the ethical standards of the responsible committee on human experimentation (written approval by the Research Ethics Boards of the medical faculty of Ludwig-Maximilians-Universität, Munich, number 068-08 and 222-10) and with the Helsinki Declaration of 1975, as revised in 2000. AML-538 was kindly provided by Claudia Baldus and Lorenz Bastian (Charité Universitätsmedizin Berlin, Germany). Pediatric AML PDX samples were a gift from Maya C. André and Martin Ebinger (University Children's Hospital Tuebingen, Germany), and were described previously.(1) Genetic profiling of AML PDX and primary AML samples was performed by Maja Rothenberg-Thurley and Klaus H. Metzeler, as described previously.(2)

The patient derived xenograft (PDX) mouse model of patients' AML

Xenotransplantation and establishing AML PDX cells in NSG mice (NOD-*scid*-gamma, The Jackson Laboratory, Bar Harbour, ME, USA) was performed as described previously.(3) In the study presented here, only AML PDX cells were applied that re-engrafted in NSG mice over several passages, and lead to a BM chimerism above 90% $hCD33^+$ $hCD45^+$ cells within 16 weeks after transplantation. These requirements precluded the use of primary patient cells, slow engrafters, low engrafters, or samples without the capacity to re-engraft; therefore, the PDX cohort used in this study is enriched for highly aggressive samples. All animal trials were

performed in accordance with the current ethical standards of the official committee on animal experimentation (written approval by Regierung von Oberbayern, tierversuche@reg-ob.bayern.de; 55.2-1-54-2531-95-10, ROB-55.2Vet-2532.Vet_02-15-193, ROB-55.2Vet-2532.Vet_02-16-7 and ROB-55.2Vet-2532.Vet_03-16-56). Accuracy of sample identity was verified by repetitive finger printing using PCR of mitochondrial DNA.(4)

Lentiviral transduction of AML PDX cells and enrichment of transgenic cells

In vitro culture, lentiviral constructs, transduction, and sorting of transgenic PDX cells were performed as described previously.(3) In the study presented here, cells were transduced with a construct expressing enhanced firefly luciferase and mCherry in equimolar amounts (pCDH-EF1a-eFFly-mCherry, available via Addgene #104833). Initial lentiviral transduction efficiencies were between 1% and 44% (AML-346 30%, AML-372 1%, AML-388 2%, AML-393 12%, AML-491 11%, AML-579 1%, AML-661 44%). PDX cells were sorted using a FACSAria III (BD Biosciences, Heidelberg, Germany) to reach a purity of more than 95% of $mCherry^+$ cells. As control, three AML PDX samples without transgenic expression of firefly luciferase and mCherry were applied (AML-356, AML-358, AML-538).

Bioluminescence *in vivo* imaging (BLI)

BLI and quantification of tumor burden was performed as described previously.(3, 5)

Labeling of PDX cells with carboxyfluorescein succinimidyl ester (CFSE)

Labeling of PDX cells with CFSE was performed as described previously.(5) In brief, AML PDX cells were isolated from mice with advanced disease stage, indicated by a BM chimersim of more than 90% $mCherry^+$ PDX cells. Cells were labeled with CFSE *ex vivo* (Life Technologies, Carlsbad, CA, USA) according to manufacturer's

instructions, washed in PBS, and injected into next recipient mice (10^7 $CFSE^+$ PDX cells per mouse). The procedure resulted in CFSE positivity of well above 98% of PDX cells, as validated by flow cytometry. As AML PDX cells are heterogeneous in size, loss of CFSE appears as continuum in flow cytometry, devoid of the distinct peaks known from normal leukocytes.

Enriching and quantifying PDX and label-retaining cells (LRC) from mouse BM

To purify AML PDX cells from mouse BM, bones from hip, femura, tibiae, sternum and spine were crushed with a mortar and pistil, cells were washed once in PBS, and filtered through a cell strainer (EASYSTRAINER 70 µM, Greiner bio-one, Frickenhausen, Germany). Murine cells were depleted using magnetic beads according to manufacturer's instructions (Mouse Cell Depletion Kit, Miltenyi Biotech, Bergisch Gladbach, Germany), with the exception that only 100 µl MicroBeads and two columns were used for one mouse BM suspension. As second step, AML PDX cells were analyzed or sorted by flow cytometry by gating on (i) leukocytes in FSC/SSC, and (ii) $DAPI^-$ living cells and transgenic $mCherry^+$ AML PDX cells using a BD LSRFortessa or FACSAriaIII, respectively (BD Biosciences, Heidelberg, Germany) as shown in Figure **1B**. Sorting of LRC and non-LRC was performed with the precision setting "purity" at the FACSAria. To determine the fraction of low-cycling AML PDX cells, LRC were discriminated from non-LRC using CFSE staining as shown in Figure **1B.** To quantify LRC, CFSE mean fluorescence intensity (MFI) of CFSE labelled PDX cells either incubated for two to three days *ex vivo* or isolated from a mouse two to three days after injection was measured, which defined the starting condition (''0 divisions''). Day two or three CFSE MFI was divided by factor two to calculate putative CFSE bisections mimicking cell divisions. Cells with a high CFSE signal below three bisections of the maximum CFSE MFI were defined as

LRC. Seven CFSE MFI bisections or more were defined as entire loss of the CFSE signal characterizing non-LRC. At late time points, due to high cell numbers and time-dependent issues, after murine cell depletion often 1/10 of cells were analyzed by flow cytometry. Absolute cell numbers were calculated thereof.

Cell cycle analysis

AML PDX cells were isolated from a first donor mouse, labeled with CFSE and 10^7 cells were transplanted into first recipients. Sixteen days after injection, AML PDX cells were re-isolated as described above. After Mouse Cell Depletion Kit, cells were stained with Vibrant DyeCycle Violet (Invitrogen, Eugene, OR, USA) according to manufacturer's instructions. Cells were analyzed by flow cytometry for cell cycle distribution within the LRC and non-LRC compartment.

RNA sequencing and data analysis

Library Preparation of RNA-Seq: 1,000 and 2,000 cells of each individual sample were sorted and lysed in RLT Plus (Qiagen) supplemented with 1% 2-Mercaptoethanol (Sigma Aldrich) and stored at -80°C until processing. A modified SCRB-seq protocol (6, 7) was used for library preparation. Briefly, proteins in the lysate were digested by Proteinase K (Ambion), RNA was cleaned up using SPRI beads (GE, 22% PEG). In order to remove isolated DNA, samples were treated with DNase I for 15 min at RT. cDNA was generated by oligo-dT primers containing well specific (sample specific) barcodes and unique molecular identifiers (UMIs). Unincorporated barcode primers were digested using Exonuclease I (Thermo Fisher). cDNA was pre-amplified using KAPA HiFi HotStart polymerase (Roche) and pooled before Nextera libraries were constructed from 0.8 ng of pre-amplified cleaned up cDNA using Nextera XT Kit (Illumina). 3' ends were enriched with a custom P5 primer (P5NEXTPT5, IDT) and libraries were size selected using 2% E-

Gel Agarose EX Gels (Life Technologies), cut out in the range of 300–800 bp, and extracted using the Monarch DNA Gel Extraction Kit (New England Biolabs) according to manufacturer's recommendations.

Sequencing: Libraries were paired-end sequenced on a rapid flow cell (1.5 lanes, ~208*10^6 reads in total) of an Illumina HiSeq 1500 instrument. Sixteen bases were sequenced within the first read to obtain cellular and molecular barcodes, and 50 bases were sequenced in the second read into the cDNA fragment. An additional eight bases were sequenced to obtain the i7 barcode.

Data processing and differential gene expression and pathway analysis: All raw fastq data was demultiplexed using deML (8) and further processed with zUMIs (9). Mapping was performed using STAR 2.6.0a (10) against the concatenated human (hg38) and mouse genome (mm10). Gene annotations were obtained from Ensembl (GRCh38.84/GRCm38.75). Samples were identified via the cellular barcode, with initial phred score filtering allowing one base below 20. UMI phred filtering allowed one bases below phred 20. Differential gene expression of LRC and nLRC was calculated using the DESeq2 package following recommended workflows.(11) Pathway analysis using the MSigDB Collection "hallmark of cancer" and "KEGG" (v7.0) was conducted using default setting.(12, 13) Sequencing data are available at the NCBI Gene Expression Omnibus (GEO accession number: GSE141627).

In vivo treatment trials

AML PDX cells were injected into groups of mice (10^7 $CFSE^+$ PDX cells per mouse). Seven days after cell injection, mice were treated with a combination of Cytarabine (150 mg/kg dissolved in PBS, i.p.) on days seven, eight, and nine, and one dose of DaunoXome (20 mg/kg i.v.) on day seven (see scheme in Figure **2A**). Body weight was measured daily. Tumor burden was monitored on days seven and ten by BLI. At

day ten, mice were sacrificed, BM was collected, and AML PDX cells were isolated and analyzed for CFSE label retention as described above.

Analysis of plasticity of LRC and non-LRC

AML PDX cells were isolated from a first donor mouse, labeled with CFSE and transplanted into first recipient mice as described above. Ten (AML-393) or 15 (AML-491) days after injection, AML PDX cells were re-isolated and purely sorted into LRC and non-LRC fractions as described above (see also scheme in Figures **3** and **S4**). Limiting dilutions of sorted cells were re-injected into secondary recipient mice (between 30 and 3000 cells per mouse, see **Table S3**). Tumor outgrowth was analyzed by BLI and compared between the groups. Engraftment was determined by positive bioluminescence in vivo imaging signal, analysis of $hCD33^+/hCD45^+$ cells in peripheral blood (PB), and/or analysis of $hCD33^+/hCD45^+$ cells in BM by FACS staining. If no AML PDX cells were detectable within 150 days after injection via BLI, in PB or in BM, mice were counted as non-engrafters. LIC frequencies were determined according to Poisson statistics, using the ELDA software application (http://bioinf.wehi.edu.au/software/elda/).(14)

To determine if highly proliferative cells convert into low-cycling cells, non-LRC from a primary recipient mouse were isolated at day 20 (AML-393) or 21 (AML-491), sorted, re-labeled with CFSE, and re-injected into secondary recipient mice ($3.6*10^6$ $CFSE^+$ AML-393 non-LRC, n=8, or $1.9*10^6$ $CFSE^+$ AML-491 non-LRC, n=5). Cells were re-isolated at different time points after injection, distribution of LRC and non-LRC was analyzed, and compared to the distribution within first recipient mice.

For this analysis, many cells are needed for the re-injection into secondary recipient mice. The minute numbers of LRC that can be re-isolated after ten days from first recipient mice cannot be enriched from secondary recipient mice after re-

transplantation; therefore, it is technically unfeasible to perform this analysis with the LRC fraction of cells.

Analysis of CD34 and CD38 immunophenotype

AML PDX cells were isolated from a first donor mouse, labeled with CFSE and 10^7 cells were transplanted into first recipients. Ten (AML-393) or 14 (AML-491) days after injection, AML PDX cells were re-isolated as described above. After Mouse Cell Depletion Kit, 19/20 of cells were stained with 10 µl PC7-conjugated CD34 monoclonal antibody 581 (Beckman Coulter,Marseille, France) and 10 µl APC-conjugated CD38 monoclonal antibody HIT2 (BioLegend, San Diego, CA, USA). In the remaining 1/20 of cells, antibody-reactivity was controlled using 5 µl isotype-matched control-antibodies. For detection of CD34+/CD38- cells within the LRC and non-LRC compartment, cells were analyzed by flow cytometry.

Statistics

Statistical analyses were calculated using GraphPad Prism 7 software. To compare groups after drug treatment, we first checked normality in the control and treatment group of each sample using the Shapiro-Wilk normality test. If normality assumption was rejected, the Mann Whitney U test was applied. Otherwise, variance homogeneity was tested using the F test, and based on these results, we applied the students or welchs t-test as appropriate. Due to the explorative nature and the limited statistical power based on small sample size, we decided not to correct for multiple testing with respect to tumor samples. ELDA software was used to test differences in LIC frequency by chi-square test (http://bioinf.wehi.edu.au/software/elda/).(14)

References

1. Woiterski J, Ebinger M, Witte KE, Goecke B, Heininger V, Philippek M, et al. Engraftment of low numbers of pediatric acute lymphoid and myeloid leukemias into NOD/SCID/IL2Rcgammanull mice reflects individual leukemogenecity and highly correlates with clinical outcome. International journal of cancer. 2013 Oct 1;133(7):1547-56.

2. Metzeler KH, Herold T, Rothenberg-Thurley M, Amler S, Sauerland MC, Görlich D, et al. Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. Blood. 2016;128(5):686.

3. Vick B, Rothenberg M, Sandhofer N, Carlet M, Finkenzeller C, Krupka C, et al. An advanced preclinical mouse model for acute myeloid leukemia using patients' cells of various genetic subgroups and in vivo bioluminescence imaging. PloS one. 2015;10(3):e0120925.

4. Hutter G, Nickenig C, Garritsen H, Hellenkamp F, Hoerning A, Hiddemann W, et al. Use of polymorphisms in the noncoding region of the human mitochondrial genome to identify potential contamination of human leukemia-lymphoma cell lines. The hematology journal : the official journal of the European Haematology Association. 2004;5(1):61-8.

5. Ebinger S, Ozdemir EZ, Ziegenhain C, Tiedt S, Castro Alves C, Grunert M, et al. Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia. Cancer Cell. 2016 Dec 12;30(6):849-62.

6. Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. bioRxiv. 2014:003236.

7. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Molecular cell. 2017 Feb 16;65(4):631-43.e4.

8. Renaud G, Stenzel U, Maricic T, Wiebe V, Kelso J. deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. Bioinformatics (Oxford, England). 2015 Mar 1;31(5):770-2.

9. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. GigaScience. 2018 Jun 1;7(6).

10. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England). 2013 Jan 1;29(1):15-21.

11. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.

12. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics. 2003 Jul;34(3):267-73.

13. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005 Oct 25;102(43):15545-50.

14. Hu Y, Smyth GK. ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. Journal of Immunological Methods. 2009 2009/08/15/;347(1):70-8.

Table S1. Clinical characteristics of AML patients

Sample	disease stage*	age[1] [years]	sex	cytogenetics	mutations[2]
AML-346	R1	1	f	int. del(5q)(13q)	CKIT
AML-356	R1	5	m	*ND*	*ND*
AML-358	R2	9	m	*ND*	FLT3-TKD
AML-372	R1	42	m	complex, incl. -17	KRAS, TP53
AML-388	ID	57	m	KMT2A-AF6	KRAS, CEBPZ
AML-393	R1	47	f	KMT2A-AF10	BCOR, KRAS
AML-491	R1	53	f	del(7)(q2?1)	DNMT3A, BCOR, NRAS, KRAS, ETV6, PTPN11, RUNX1
AML-538	R1	68	f	CN	DNMT3A, IDH1
AML-579	R1	51	m	CN	NPM1, FLT3-ITD, DNMT3A, IDH1
AML-661	R2	55	f	del(7)(q2?1)	DNMT3A, BCOR, NRAS, ETV6, PTPN11, RUNX1, EZH2

[1]when the primary AML sample was obtained; [2]mutations detected by targeted re-sequencing in PDX cells; ID = initial diagnosis; R1 = 1st relapse; R2 = 2nd relapse; int = interstitial; del = deletion; CN = cytogenetically normal; f = female; m = male; *ND* = not determined

Table S2: LIC frequencies of different AML PDX samples (Related to Figure **S1D**)

Sample	# of cells*	# of mice injected / engrafted	LIC frequency (95% CI)
AML-372	100,000	3 / 3	**1/5,125** (1/1,926 - 1/13,640)
	30,000	3 / 3	
	10,000	3 / 3	
	3,000	3 / 1	
	1,000	3 / 0	
AML-388	72,000	1 / 1	**1/3,665** (1/939 - 1/14,300)
	24,000	1 / 1	
	21.870	1 / 1	
	7,290	1 / 1	
	2,430	2 / 1	
	710	1 / 0	
	270	2 / 0	
	90	1 / 0	
	30	2 / 0	
AML-346	100,000	4 / 4	**1/2,337** (1/898 - 1/6,093)
	20,000	3 / 3	
	10,000	4 / 4	
	2,000	3 / 2	
	1,000	4 / 1	
	100	4 / 0	
AML-491	10,000	3 / 3	**1/1,799** (1/945 - 1/3,426)
	5,400	2 / 2	
	2,000	2 / 1	
	1,800	2 / 0	
	1,200	6 / 6	
	1,000	2 / 1	
	600	5 / 0	
	200	3 / 0	
	100	4 / 0	
AML-393	20,000	3 / 3	**1/507** (1/194-1/1,325)
	2,000	3 / 3	
	666	3 / 1	
	200	3 / 2	
	66	3 / 1	
AML-579	72,900	1 / 1	**1/351** (1/77.6-1/1,590)
	24,300	2 / 2	
	7,100	1 / 1	
	2,700	2 / 2	
	900	1 / 1	
	300	2 / 1	
AML-661	8,100	1 / 1	**1/546** (1/230 - 1/1,403)
	2,700	1 / 1	
	900	1 / 3	
	300	1 / 3	
	100	2 / 4	
	33	2 / 4	
	11	0 / 3	

*Cells from different AML samples were transplanted into recipient mice in limiting dilutions at numbers indicated; bioluminescence *in vivo* imaging, blood measurement or bone marrow FACS staining was performed to determine engraftment; LIC frequency was calculated using the ELDA software; 95% confidence interval (CI).

Table S3: LIC frequencies of LRC and nLRC (Related to Figure **3B** and **S4B**)

Sample	group	# of cells	# of mice injected / engrafted	LIC frequency (95% CI)
AML-393	non-LRC	3,000	1 / 1	**1/352** (1/157 - 1/788)
		1,480	2 / 1	
		1,400	1 / 1	
		1,000	2 / 2	
		330	2 / 2	
		300	2 / 2	
		200	2 / 1	
		100	3 / 1	
		30	3 / 1	
	LRC	1,500	1 / 1	**1/132** (1/59 - 1/294)
		1,480	2 / 2	
		1,400	1 / 1	
		1,000	1 / 1	
		900	2 / 2	
		330	2 / 1	
		300	3 / 3	
		200	2 / 2	
		100	2 / 2	
		30	3 / 0	
AML-491	non-LRC	2,000	1 / 1	**1/1,080** (1/336 - 1/3,474)
		1,200	2 / 2	
		950	1 / 0	
		600	1 / 0	
	LRC	2,000	1 / 1	**1/1,021** (1/324 - 1/3,225)
		1,200	2 / 2	
		600	2 / 0	
		200	1 / 0	

*LRC and non-LRC from first recipient mice were sorted and were transplanted into secondary recipient mice in limiting dilutions at numbers indicated; bioluminescence *in vivo* imaging, blood measurement or bone marrow FACS staining was performed to determine engraftment; LIC frequency was calculated using the ELDA software; 95% confidence interval (CI).

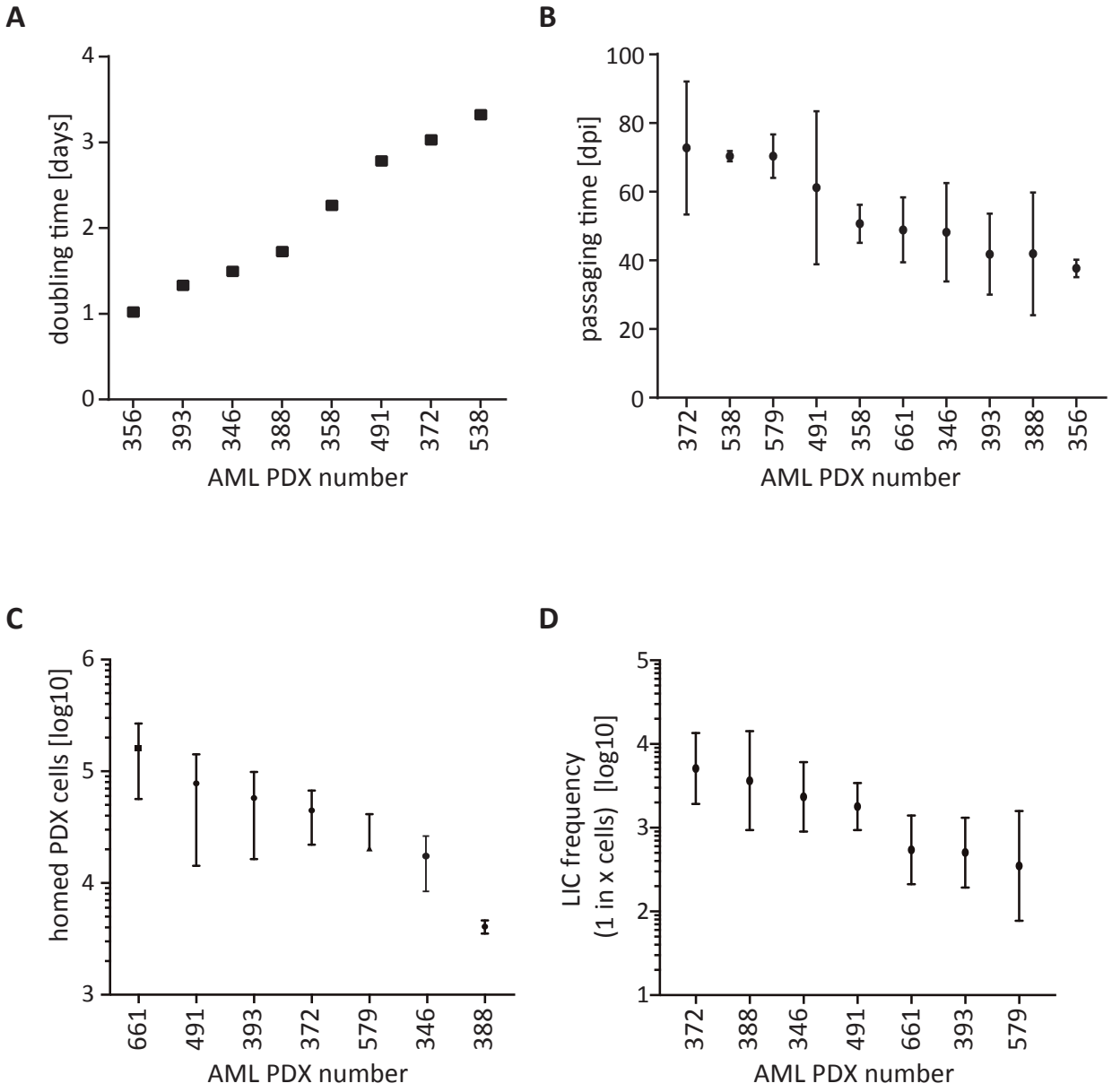


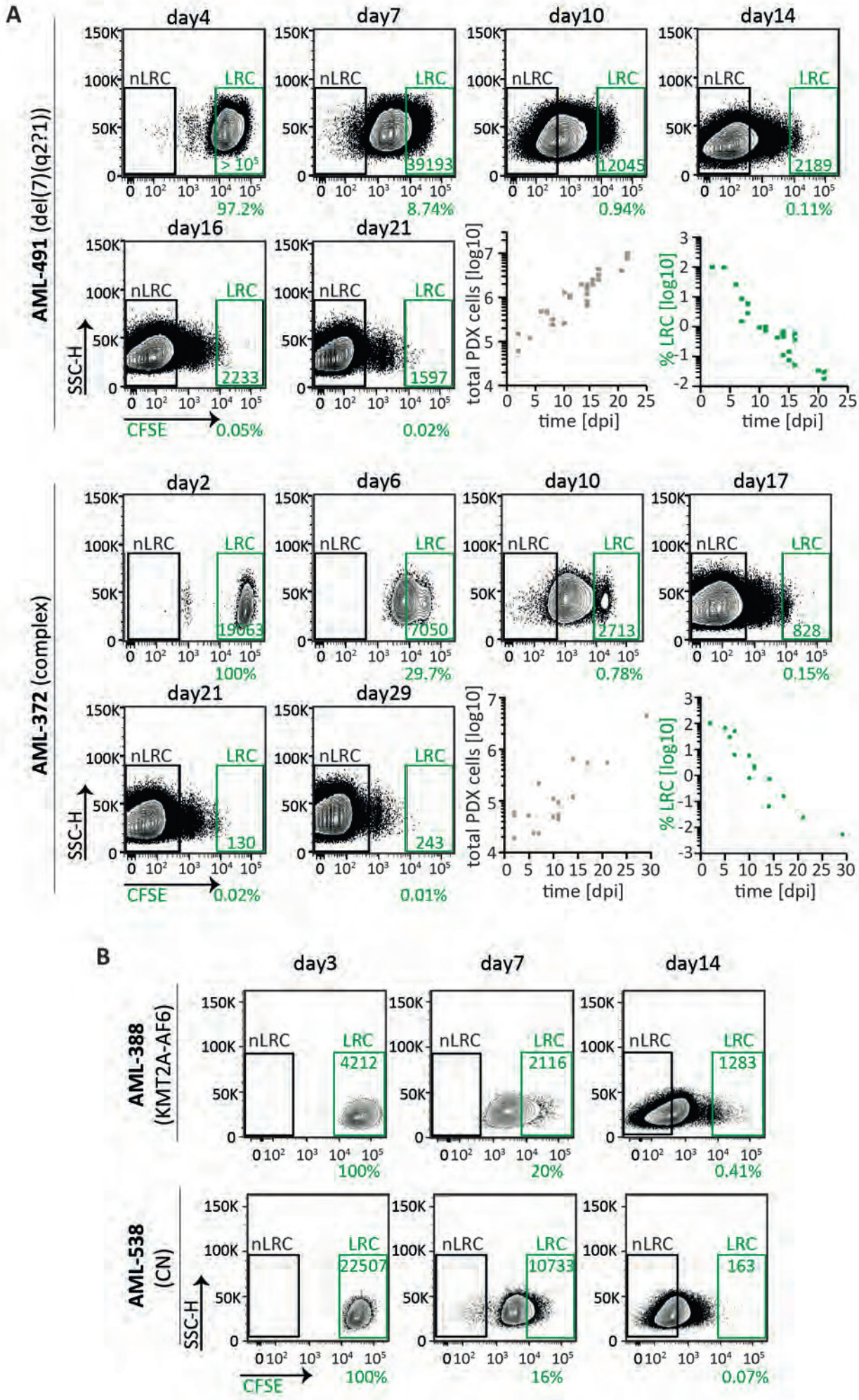
Figure S1 AML PDX cells display heterogeneity regarding *in vivo* proliferation, homing and LIC frequency (related to Figure **1**).

A *In vivo* doubling times were calculated out of growth curves measured as in Figures **1D** and **S2**.

B Passaging times from injection until overt leukemia, $5x10^5$ to $5x10^6$ AML PDX cells were injected per mouse; mean +/- SD of at least 4 and up to 100 mice per sample is depicted. dpi=days post injection.

C Number of AML PDX cells homing to the BM was determined 2 or 3 days following injection of 10^7 cells; mean +/- SD of at least 3 mice is shown.

D Bulk cells from different AML samples were transplanted into recipient mice in limiting dilutions at numbers indicated in Table **S2**. LIC frequency was calculated using the ELDA software and mean +/- 95%CI is depicted.


A
AML-491 (del(7)(q2?1))
day4
day7
day10
day14
day16
day21
nLRC
LRC
> 10⁵
39193
12045
2189
2233
1597
97.2%
8.74%
0.94%
0.11%
0.05%
0.02%
SSC-H
CFSE
total PDX cells [log10]
% LRC [log10]
time [dpi]
AML-372 (complex)
day2
day6
day10
day17
day21
day29
19063
7050
2713
828
130
243
100%
29.7%
0.78%
0.15%
0.02%
0.01%
B
AML-388 (KMT2A-AF6)
day3
day7
day14
4212
2116
1283
100%
20%
0.41%
AML-538 (CN)
22507
10733
163
100%
16%
0.07%

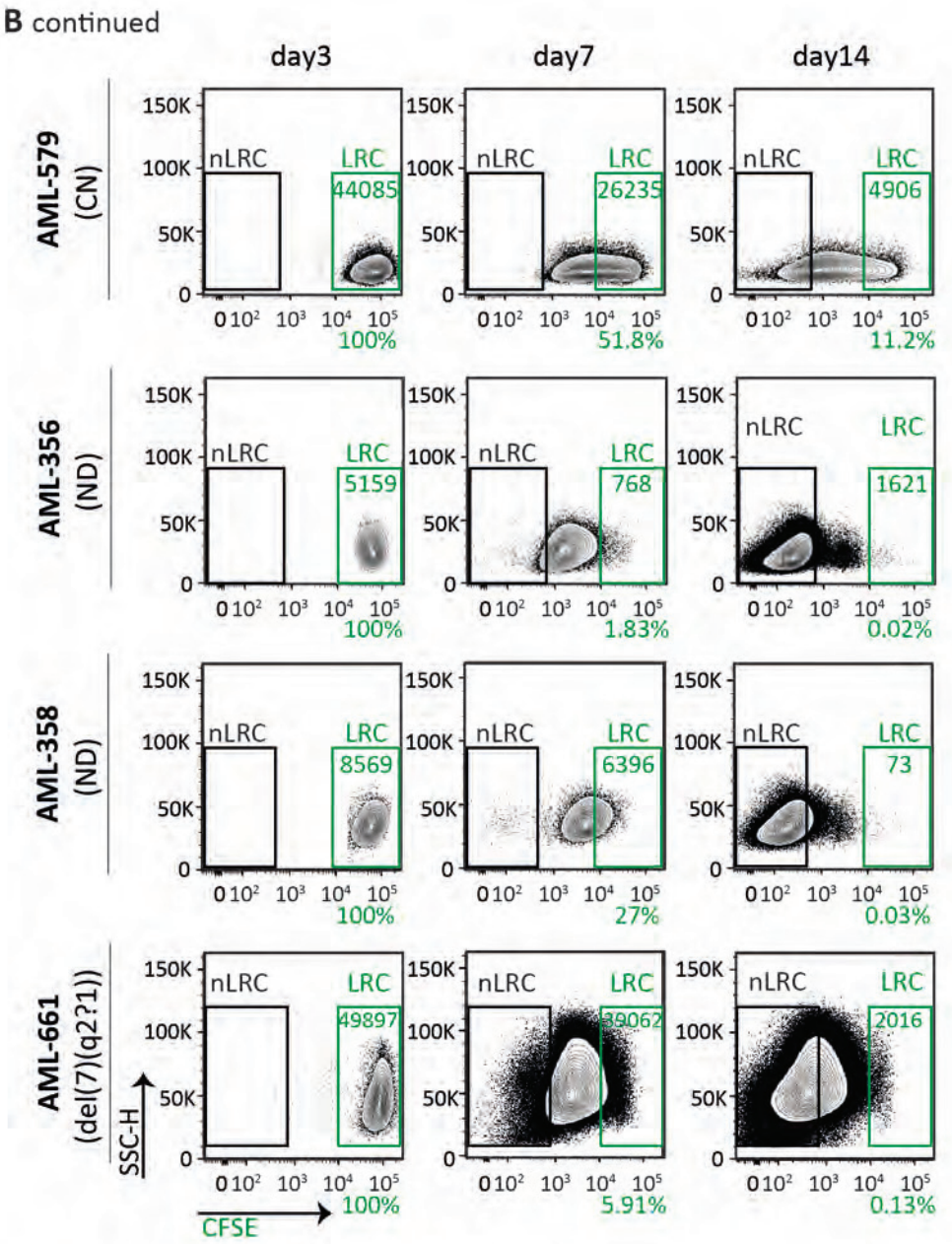

B continued
day3
day7
day14
AML-579 (CN)
AML-356 (ND)
AML-358 (ND)
AML-661 (del(7)(q2?1))
nLRC
LRC
44085
26235
4906
5159
768
1621
8569
6396
73
49897
39062
2016
100%
51.8%
11.2%
100%
1.83%
0.02%
100%
27%
0.03%
100%
5.91%
0.13%
SSC-H
CFSE
C
day1
day 3
day 5
day 7
day 9
day 15
AML-346 (int.del.(5q)(13q))
6206
6930
947
50
2
8
0
100%
83.4%
2.9%
0.12%
0.001%
0.003%
0%

Figure S2 AML PDX cells contain a rare subpopulation of lowly-cycling cells (related to Figure **1,** additional samples).

Experiments were performed and depicted identically as in Figure **1**. In brief, 10^7 CFSE labeled AML PDX cells were injected into groups of mice; cells were isolated at different time points and analyzed by flow cytometry for CFSE content. FACS plots for representative mice are shown. Total number of isolated PDX cells and percentage of LRC cells among all isolated PDX cells are shown in (**A**). Total number of mice studied was (**A**) 30 for AML-491, 18 for AML-372, (**B**) 5 for AML-388, 3 for AML-538, 8 for AML-579, 3 for AML-356, 3 for AML-358, 8 for AML-661 and (**C**) 10 for AML-346.

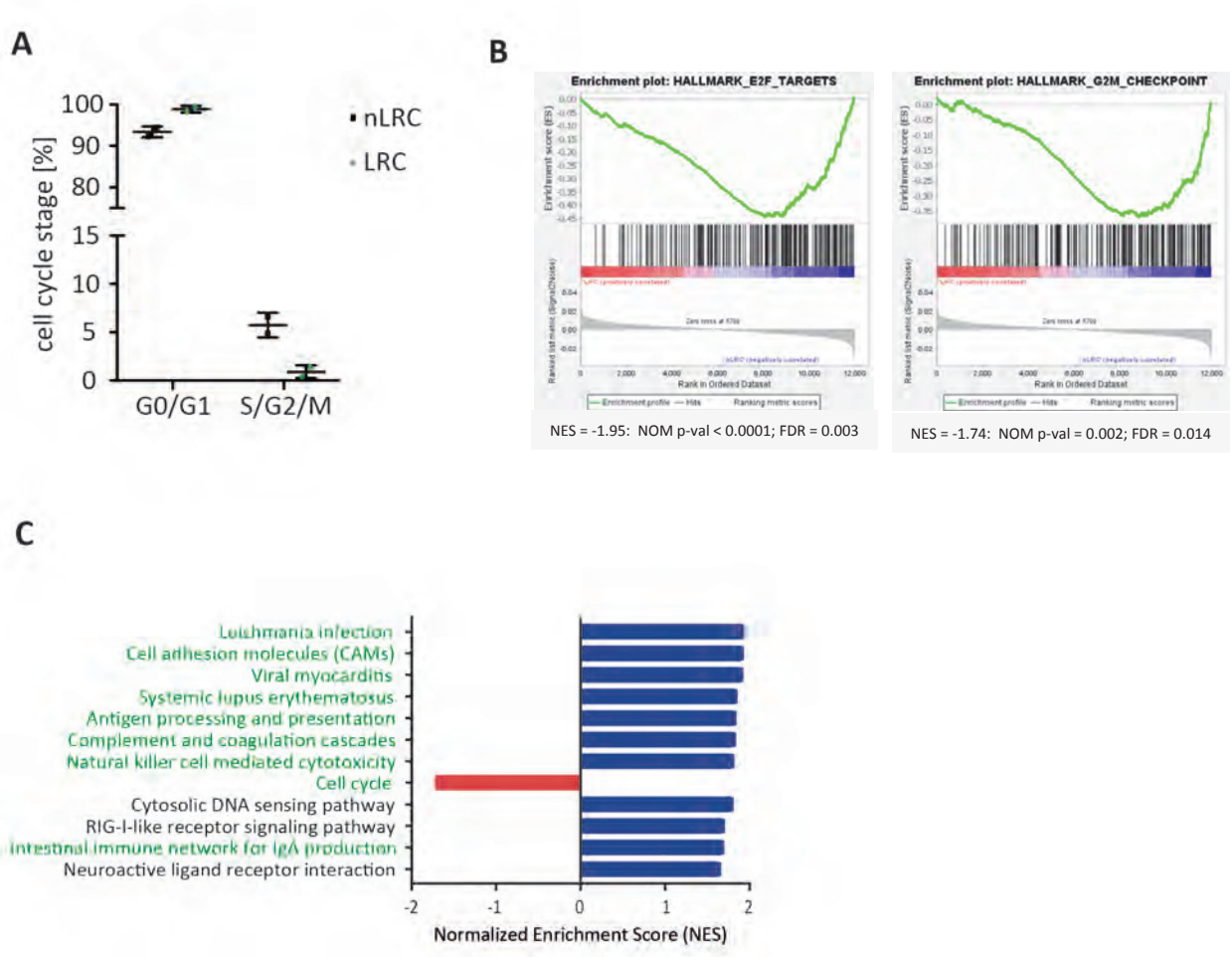


Figure S3 AML LRC are low cycling and resemble LRC from acute lymphoblastic leukemia (related to Figure **1**).

A Cell cycle analysis: 10^7 CFSE labeled AML-491 PDX cells were injected into two mice; cells were isolated at day 16, stained with Vibrant DyeCycle Violet and analyzed for CFSE content and cell cycle distribution by flow cytometry.

B Gene set enrichment analysis (GSEA): Comparison of LRC versus non-LRC by GSEA (hallmarks of cancer) demonstrate down regulation of the cell cycle activity pathways E2F target genes and G2M Checkpoint, indicating reduced proliferation.

C Significantly enriched KEGG pathways (nominal p-value ≤0.01) comparing LRC and nLRC of AML-393 and AML-491; KEGG pathways with concordant enrichment in both AML (new data here) and ALL (Ebinger et al, 2016) are marked in green.

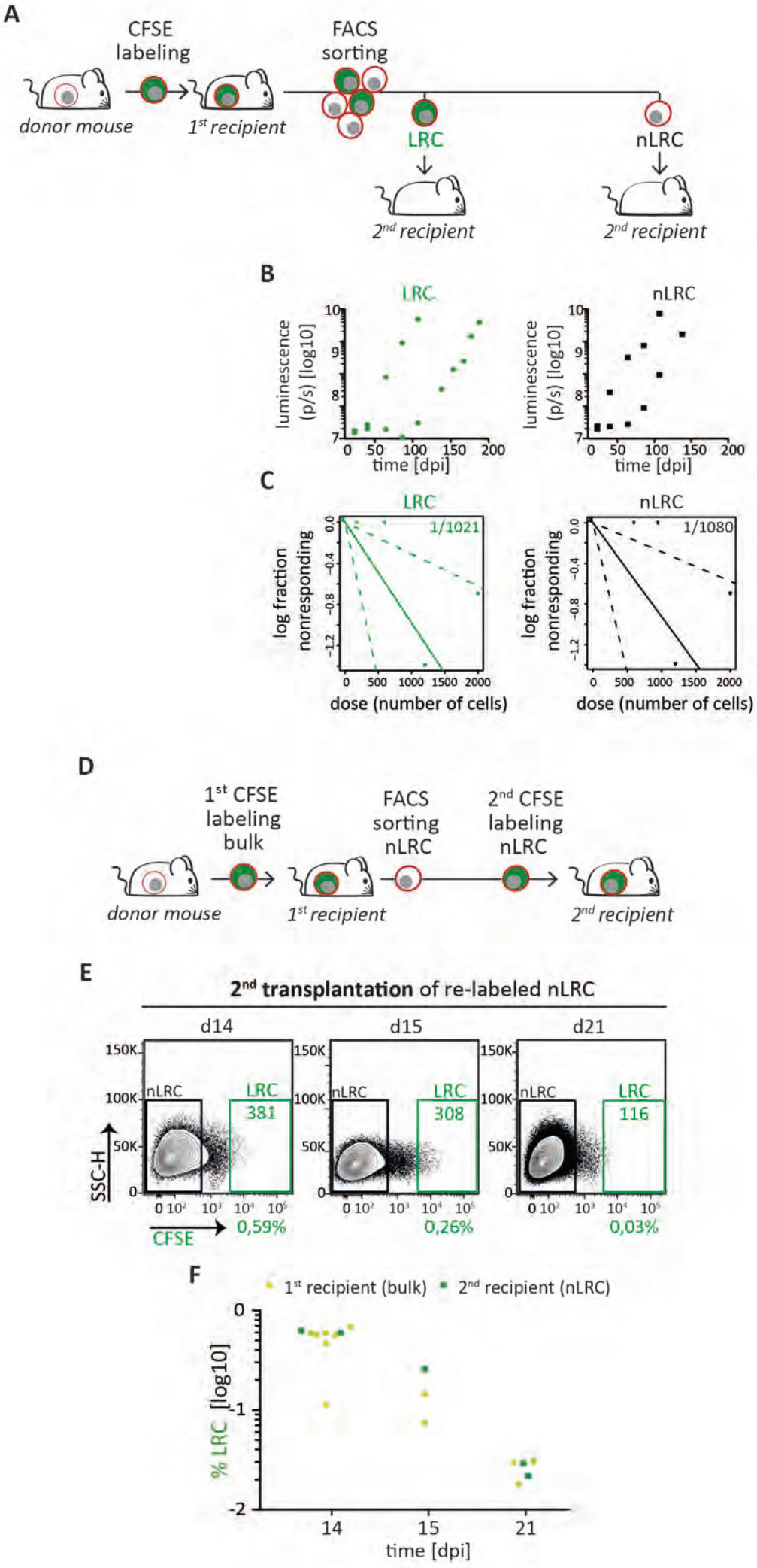

A
CFSE labeling
FACS sorting
donor mouse
1st recipient
LRC
nLRC
2nd recipient
2nd recipient
B
LRC
nLRC
luminescence (p/s) [log10]
time [dpi]
C
LRC
nLRC
1/1021
1/1080
log fraction nonresponding
dose (number of cells)
D
1st CFSE labeling bulk
FACS sorting nLRC
2nd CFSE labeling nLRC
donor mouse
1st recipient
2nd recipient
E
2nd transplantation of re-labeled nLRC
d14
d15
d21
nLRC
LRC
381
308
116
SSC-H
CFSE
0,59%
0,26%
0,03%
F
1st recipient (bulk)
2nd recipient (nLRC)
% LRC [log10]
time [dpi]

Figure S4 **AML PDX cells display reversible growth behavior, independently from stemness potential** (related to Figure **3,** additional sample AML-491)**.**

Experiments were performed and data depicted identically as in Figure **3**;

A Experimental procedure; AML-491 PDX cells were isolated from advanced disease donor mice (n=2 in two independent experiments), labeled with CFSE, and re-transplanted into first recipient mice. Fifteen days after injection, cells were re-isolated and sorted into LRC and non-LRC (nLRC) using the gates as described in Figure **1B** and re-injected into secondary recipient mice.

B Secondary recipient mice receiving either 1200 LRC or 1200 non-LRC (n=4) were monitored by in vivo imaging.

C LRC and non-LRC were re-injected into secondary recipient mice (n=11) in limiting dilutions at numbers as indicated in Table **S3**. Positive engraftment of PDX cells was determined by in vivo imaging and/or flow cytometry. LIC frequency was calculated using the ELDA software and is depicted +/- 95% confidence interval. No statistically significant difference between LIC frequency of LRC and non-LRC was found according to chi-square test (p=0.95).

D Experimental procedure; from first recipient mice (n=2 in 2 independent experiments) harboring CFSE stained cells, non-LRC were isolated at day 21, re-stained with CFSE and $1.9*10^6$ cells injected into secondary recipients; (n=5); Cells were re-isolated 14, 15 and 21 days later and LRC were quantified using gates as described in Figure **1B**. The experiment is technically unfeasible for LRC as the high number of cells needed cannot be generated.

E,F Representative dot plots (**E**) and quantification (**F**) of the percentage of LRC among all PDX cells isolated from secondary recipients is displayed (green squares). LRC of first recipient mice as determined in Figure **S2A** are shown for comparison (dots in light green).

See supplemental Table **S3** for additional data.

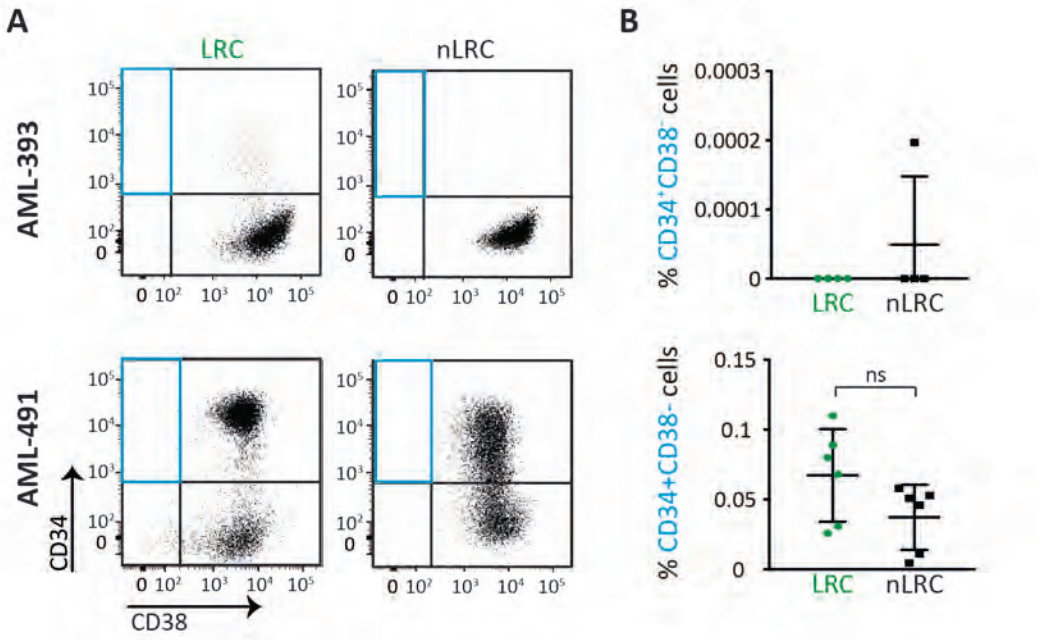


Figure S5 **Low-cycling cells are not enriched in immature cells**.

AML-393 and AML-491 PDX cells were isolated from full-blown donor mice, labeled with CFSE, and 10^7 cells were re-transplanted into first recipient mice. Ten (AML-393; n=4) or fourteen (AML-491; n=6) days after injection, cells were re-isolated, stained with CD34 and CD38 antibodies and percentage of CD34 and CD38 positive cells within the LRC and non-LRC compartment were analyzed by flow cytometry.

A representative dot plots

B percentage of CD34-positive and CD38-negative cells within the LRC and non-LRC compartment is depicted as mean+/-SD. Each dot/square represents one mouse;

Improving Single-Cell RNA Sequencing Technology

Sensitive and Powerful Single-Cell RNA Sequencing Using mcSCRB-Seq.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-05347-6 OPEN

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

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

Single-cell RNA sequencing (scRNA-seq) has emerged as a central genome-wide method to characterize cellular identities and processes. Consequently, improving its sensitivity, flexibility, and cost-efficiency can advance many research questions. Among the flexible plate-based methods, single-cell RNA barcoding and sequencing (SCRB-seq) is highly sensitive and efficient. Here, we systematically evaluate experimental conditions of this protocol and find that adding polyethylene glycol considerably increases sensitivity by enhancing cDNA synthesis. Furthermore, using Terra polymerase increases efficiency due to a more even cDNA amplification that requires less sequencing of libraries. We combined these and other improvements to develop a scRNA-seq library protocol we call molecular crowding SCRB-seq (mcSCRB-seq), which we show to be one of the most sensitive, efficient, and flexible scRNA-seq methods to date.

[1] Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Großhaderner Straße 2, 82152 Martinsried, Germany. [2] Present address: Department of Cell & Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden. [3] Present address: Max Planck Institute for Biology of Ageing, 50931 Cologne, Germany. These authors contributed equally: Johannes W. Bagnoli, Christoph Ziegenhain, Aleksandar Janjic. Correspondence and requests for materials should be addressed to W.E. (email: enard@bio.lmu.de)

Whole transcriptome single-cell RNA sequencing (scRNA-seq) is a transformative tool with wide applicability to biological and biomedical questions[1,2]. Recently, many scRNA-seq protocols have been developed to overcome the challenge of isolating, reverse transcribing, and amplifying the small amounts of mRNA in single cells to generate high-throughput sequencing libraries[3,4]. However, as there is no optimal, one-size-fits all protocol, various inherent strengths and trade-offs exist[5–7]. Among flexible, plate-based methods, single-cell RNA barcoding and sequencing (SCRB-seq)[8] is one of the most powerful and cost-efficient[6], as it combines good sensitivity, the use of unique molecular identifiers (UMIs) to remove amplification bias and early cell barcodes to reduce costs. Here, we systematically optimize the sensitivity and efficiency of SCRB-seq and generate molecular crowding SCRB-seq (mcSCRB-seq), one of the most powerful and cost-efficient plate-based methods to date (Fig. 1a).

Results

Systematic optimization of SCRB-seq. We started to test improvements to SCRB-seq by optimizing the cDNA yield and quality generated from universal human reference RNA (UHRR)[9] in a standardized SCRB-seq assay (see Supplementary Fig. 1a and Methods). By including the barcoded oligo-dT primers in the lysis buffer, we increased cDNA yield by 10% and avoid a time-consuming pipetting step during the critical phase of the protocol (Supplementary Fig. 1b). Next, we compared the performance of nine Moloney murine leukemia virus (MMLV) reverse transcriptase (RT) enzymes that have the necessary template-switching properties. Especially at input amounts below 100 pg, Maxima H- (Thermo Fisher) performed best closely followed by SmartScribe (Clontech) (Supplementary Fig. 1c). In order to reduce the costs of the reaction, we showed that cDNA yield and quality is not measurably affected when we reduced the enzyme (Maxima H-) by 20%, reduced the oligo-dT primer by 80%, or used the cheaper unblocked template-switching oligo (Supplementary Fig. 2). Next, we evaluated the effect of $MgCl_2$, betaine and trehalose, as these led to the increased sensitivity of the Smart-seq2 protocol[10]. Since both Smart-seq2 and SCRB-seq generate cDNA by oligo-dT priming, template switching, and PCR amplification, we were surprised that these additives decreased cDNA yield for SCRB-seq (Supplementary Fig. 3a). Apparently, the interactions between enzymes and buffer conditions are complex and optimizations cannot be easily transferred from one protocol to another.

Molecular crowding significantly increases sensitivity. An additive that has not yet been explored for scRNA-seq protocols is polyethylene glycol (PEG 8000). It makes ligation reactions more efficient[11] and is thought to increase enzymatic reaction rates by mimicking (macro)molecular crowding, i.e., by reducing the effective reaction volume[12]. As small reaction volumes can increase the sensitivity of scRNA-seq protocols[5,13], we tested whether PEG 8000 can also increase the cDNA yield of SCRB-seq. Indeed, we observed that PEG 8000 increased cDNA yield in a concentration-dependent manner up to tenfold (Supplementary Fig. 3b). However, at higher PEG concentrations, unspecific DNA fragments accumulated in reactions without RNA (Supplementary Fig. 3d) and therefore we chose 7.5% PEG 8000 as an optimal concentration balancing yield and specificity (Supplementary

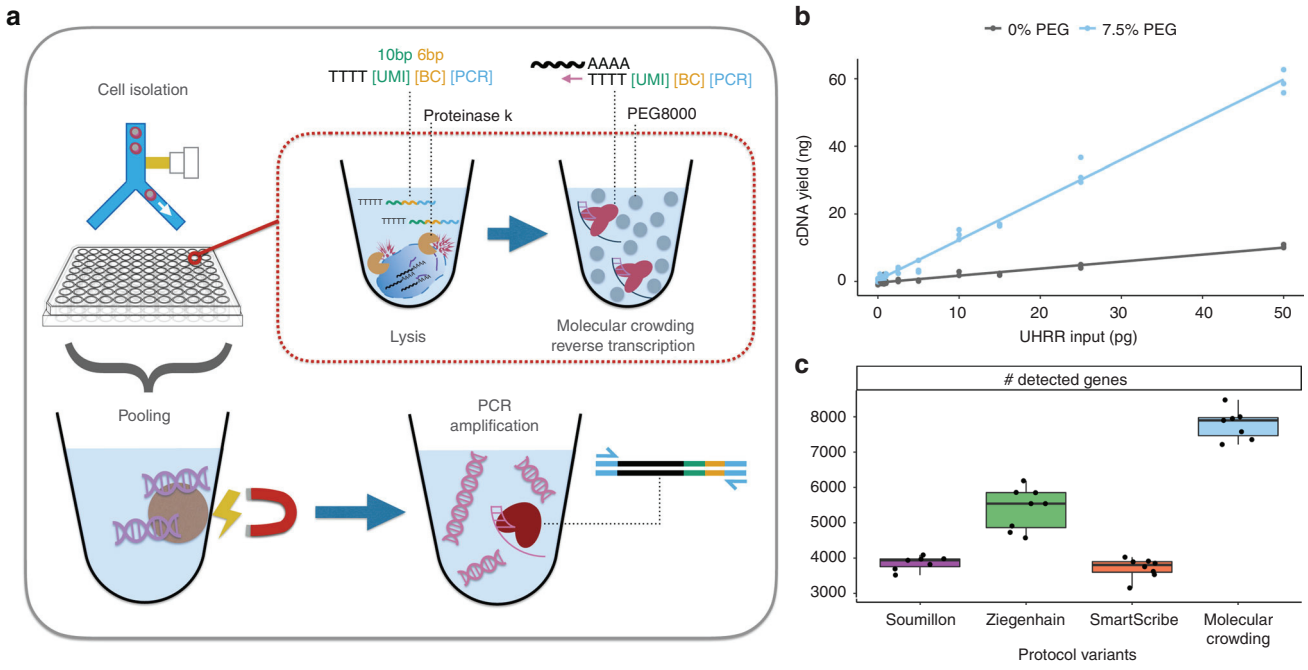


Fig. 1 mcSCRB-seq workflow and the effect of molecular crowding. **a** Overview of the mcSCRB-seq protocol workflow. Single cells are isolated via FACS in multiwell plates containing lysis buffer, barcoded oligo-dT primers, and Proteinase K. Reverse transcription and template switching are carried out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

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

To test whether these increases in cDNA yield indeed correspond to increases in sensitivity, we generated and sequenced 32 RNA-seq libraries from 10 pg of total RNA (UHRR) using eight replicates for each of the following four SCRB-seq protocol variants (Supplementary Tables 1, 2): the original SCRB-seq protocol[8] ("Soumillon"; with Maxima H- as RT and Advantage2 as PCR enzyme), the slightly adapted protocol benchmarked in Ziegenhain et al.[6] ("Ziegenhain"; with Maxima H- and KAPA), the same protocol with SmartScribe as the RT enzyme ("SmartScribe") and our optimized protocol ("molecular crowding"; with Maxima H-, KAPA, 7.5% PEG, 80% less oligo-dT, and 20% less Maxima H-). As expected, the molecular crowding protocol yielded the most cDNA, while variant "Soumillon" yielded the least, confirming our systematic optimization (Supplementary Fig. 4a). After sequencing, we processed data using *zUMIs*[14] and downsampled each of the 32 libraries to one million reads per sample, which has been suggested to correspond to reasonable saturation for single-cell RNA-seq experiments[5,6]. Of the 32 libraries, 31 passed quality control with a median of 71% of the reads mapping to exons (range: 50–77%), 12% to introns (9–15%), 13% to intergenic regions (10–31%), and 4% (3–7%) to no region in the human genome (Supplementary Fig. 4b). Of note, we observe that a higher proportion of reads are mapping to intergenic regions for the "molecular crowding" condition (Supplementary Fig. 4b). As UHRR is provided as DNAse-digested RNA, these reads are likely derived from endogenous transcripts, but why their proportion is increased in the molecular crowding protocol is unclear. In any case, we assessed the sensitivity of the protocols by the number of detected genes per cell (>=1 exonic read), representing a conservative estimate for the molecular crowding protocol with its higher fraction of intergenic reads (Fig. 1c). This sensitivity measure correlates fairly well with cDNA yield (Supplementary Fig. 4a). Hence, it shows that Maxima H- is indeed more sensitive than SmartScribe (5542 detected genes per sample in "Ziegenhain" vs. 3805 in "SmartScribe", $p = 3 \times 10^{-5}$, Welch two sample *t*-test) and that the molecular crowding protocol is the most sensitive one (7898 vs. 5542 detected genes, $p = 7 \times 10^{-7}$, Welch two sample *t*-test). In summary, we can show that our optimized SCRB-seq protocol, in particular due to the addition of PEG 8000, increases the sensitivity compared to previous protocol variants at reduced costs.

Terra retains more complexity during cDNA amplification. Next, we aimed to increase the efficiency of this protocol by optimizing the cDNA amplification step. Depending on the number of cycles, reaction conditions, and polymerases, substantial noise and bias is introduced when the small amounts of cDNA molecules are amplified by PCR[15,16]. While UMIs allow for the correction of these effects computationally, scRNA-seq methods that have less amplification bias require fewer reads to obtain the same number of UMIs and hence are more efficient[6,17]. As a first step, we evaluated 12 polymerases for cDNA yield and found KAPA, SeqAmp, and Terra to perform best (Supplementary Fig. 5a). We disregarded SeqAmp because of a decreased median length of the amplified cDNA molecules (Supplementary Fig. 5b) as well as the higher cost of the enzyme and continued to compare the amplification bias of KAPA and Terra polymerases. To this end, we sorted 64 single mouse embryonic stem cells (mESCs) and generated cDNA using our optimized molecular crowding protocol. Two pools of cDNA from 32 cells were amplified with KAPA or Terra polymerase (18 cycles) and used to generate libraries. After sequencing and downsampling each transcriptome to one million raw reads[14], we found that amplification using Terra yielded twice as much library complexity (UMIs) than when using KAPA (Supplementary Fig. 5c). This is in agreement with a recent study that optimized the scRNA-seq protocol Quartz-seq2, which also found Terra to retain a higher library complexity[17]. In addition to choosing Terra for cDNA amplification, we also reduced the number of cycles from 19 in the original SCRB-seq protocol to 14, as fewer cycles are expected to decrease amplification bias further[15] and 14 cycles still generated sufficient amounts of cDNA (~1.6–2.4 ng/µl) from mouse ESCs to prepare libraries with Nextera XT (~0.8 ng needed). Depending on the investigated cells, which may have a lower or higher RNA content than ESCs, the cycle number might need to be adapted to generate enough cDNA while avoiding overcycling.

With the final improved version of the molecular crowding protocol (mcSCRB-seq), we tested to what extent cross-contamination occurs. For example, chimeric PCR products may occur following the pooling of cDNA[18] and we assessed whether this might potentially be influenced by PEG that is present during cDNA synthesis before pooling. To this end, we sorted 96 cells of a mixture of mESCs and human-induced pluripotent stem cells, synthesized cDNA according to the mcSCRB-seq protocol with and without the addition of PEG and generated libraries for each of the two conditions. After mapping the sequenced reads to the joint human and mouse reference genomes, each barcode/well could be clearly classified into human or mouse cells, indicating that no doublets were sorted into wells, as may be expected for a fluorescence-activated cell sorting (FACS)-based cell isolation (Supplementary Fig. 6a). Importantly, the median number of reads mapping best to the wrong species is less than 2000 per cell (<0.4% of all reads or <1.5% of uniquely mapped reads). This is not influenced by the addition of PEG, as may be expected, since PEG is only present during cDNA generation (Supplementary Fig. 6b; two-sided *t*-test, p value = 0.81). In summary, we developed an optimized protocol, mcSCRB-seq, that has higher sensitivity, a less biased amplification and little crosstalk of reads across cells.

mcSCRB-seq increases sensitivity 2.5-fold more than SCRB-seq. To directly compare the entire mcSCRB-seq protocol to the previously benchmarked SCRB-seq protocol used in Ziegenhain et al.[6] (Supplementary Table 2), we sorted for each method 48 and 96 single mESCs from one culture into plates, and added ERCC spike-ins[19]. Following sequencing, we filtered cells to discard doublets/dividing cells, broken cells, and failed libraries (see Methods). The remaining 249 high-quality libraries all show a similar mapping distribution with ~50% of reads falling into exonic regions (Supplementary Fig. 7). When plotting the number of detected endogenous mRNAs (UMIs) against sequencing depth, mcSCRB-seq clearly outperforms SCRB-seq and detects 2.5 times as many UMIs per cell at depths above 200,000 reads (Fig. 2a and Supplementary Fig. 8a). At two million reads, mcSCRB-seq detected a median of 102,282 UMIs per cell and a median of 34,760 ERCC molecules, representing 48.9% of all spiked in ERCC molecules (Supplementary Fig. 8b). Assuming that the efficiency of detecting ERCC molecules is representative of the efficiency to detect endogenous mRNAs, the median content per mESC is 227,467 molecules (Supplementary Fig. 8c and 8d), which is very similar to previous estimates using mESCs and STRT-seq, a 5′ tagged UMI-based scRNA-seq protocol[20]. As expected, the higher number of UMIs in mcSCRB-seq also results in a higher number of detected genes. For instance, at 500,000 reads, mcSCRB-seq detected 50,969 UMIs that corresponded to

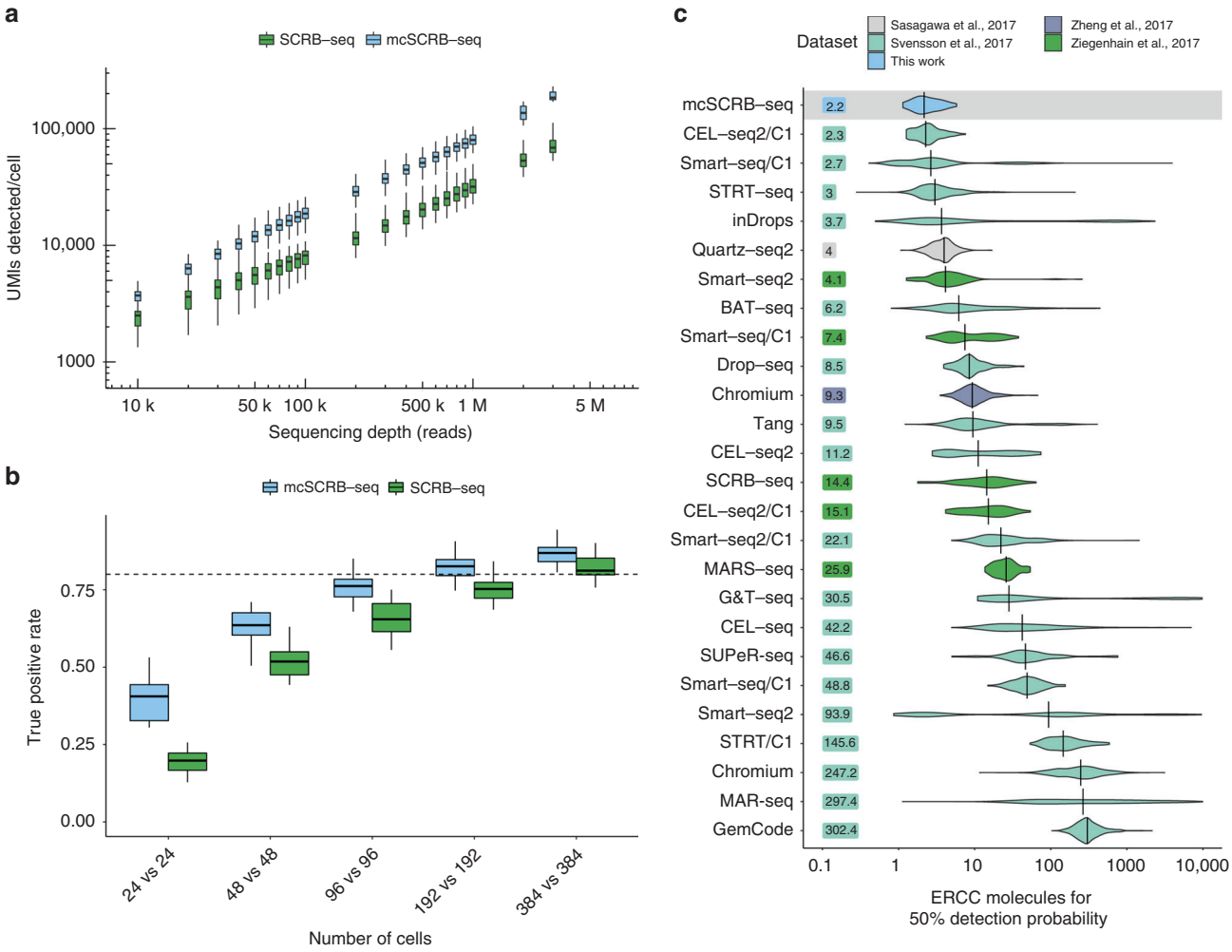


Fig. 2 Comparison of mcSCRB-seq to SCRB-seq and other protocols. **a** Number of UMIs detected in libraries generated from 249 single mESCs using SCRB-seq or mcSCRB-seq when downsampled to different numbers of raw sequence reads. Each box represents the median and first and third quartiles per cell, sequencing depth and method. Whiskers indicate the most extreme data point that is no more than 1.5 times the length of the box away from the box. **b** The true positive rate of mcSCRB-seq and SCRB-seq estimated by power simulations using the powsimR package[22]. The empirical mean–variance distribution of the 10,904 genes that were detected in at least 10 cells in either mcSCRB-seq or SCRB-seq (500,000 reads) was used to simulate read counts when 10% of the genes are differentially expressed. Boxplots represent the median and first and third quartiles of 25 simulations with whiskers indicating the most extreme data point hat is no more than 1.5 times the length of the box away from the box. The dashed line indicates a true positive rate of 0.8. The matching plot for the false discovery rate is shown in Supplementary Fig. 11d. **c** Sensitivity of mcSCRB-seq and other protocols, calculated as the number of ERCC molecules needed to reach a 50% detection probability as calculated in Svensson et al[5]. Per-cell distributions are shown using violin plots with vertical lines and numbers indicating the median per protocol

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

Decisively, we find by using power simulations[6,22] that mcSCRB-seq requires approximately half as many cells as SCRB-seq to detect differentially expressed genes between two groups of cells (Fig. 2b and Supplementary Fig. 11d). Hence, the higher sensitivity and lower noise of mcSCRB-seq compared to SCRB-seq, as measured in parallelly processed cells, indeed matters for quantifying gene expression levels and can be quantified as a doubling of cost-efficiency. Furthermore, we have reduced the reagent costs from about 1.70 € per cell for SCRB-seq[6] to less than 0.54 € for mcSCRB-seq (Supplementary Fig. 12a and Supplementary Table 3). Together, this makes mcSCRB-seq sixfold more cost-efficient than SCRB-seq. Moreover, owing to an optimized workflow, we could reduce the library preparation time to one working day with minimal hands-on time (Supplementary Fig. 12b and Supplementary Table 4). As SCRB-seq was already one of the most cost-efficient protocols in our recent benchmarking study[6], this likely makes mcSCRB-seq the most cost-efficient plate-based method available.

Benchmarking by ERCCs. The widespread use of ERCC spike-ins also allows us to estimate and compare the absolute sensitivity across many scRNA-seq protocols using published data[5]. As in Svensson et al.[5], we used a binomial logistic regression to estimate the number of ERCC transcripts that are needed on average to reach a 50% detection probability (Supplementary Fig. 13a).

mcSCRB-seq reached this threshold with 2.2 molecules, when ERCCs are sequenced to saturation (Supplementary Fig. 13b). When comparing this to a total of 26 estimates for 20 different protocols obtained from two major protocol comparisons[5,6] as well as additional relevant protocols[17,23], mcSCRB-seq has the highest sensitivity among all protocols compared to date (Fig. 2c). It should be noted that the data show large amounts of variation within protocols, even for well-established, sensitive methods like Smart-seq2. This is the case, especially in Svensson et al.[5], because the data were generated from many varying cell types sequenced in numerous labs. Similarly, mcSCRB-seq sensitivity estimates could be variable across labs and conditions. Nevertheless, the average ERCC detection efficiency is the most representative measure to compare sensitivities across many protocols.

mcSCRB-seq detects biological differences in complex tissues. Finally, we applied mcSCRB-seq to peripheral blood mononuclear cells (PBMCs), a complex cell population with low mRNA amounts, to test whether it is efficient in recapitulating biological differences. We obtained PBMCs from one healthy donor, FACS-sorted cells in four 96-well plates and prepared libraries using mcSCRB-seq with a more stringent lysis condition (see Methods; Fig. 3a). We sequenced ~203 million reads for the resulting pool, of which ~189 million passed filtering criteria in the *zUMIs* pipeline (see Methods). Next, we filtered low-quality cells (<50,000 raw reads or mapping rates <75%; Supplementary Fig. 14a), leaving 349 high-quality cells for further analysis (Supplementary Fig. 14b). Using the Seurat package[24], we clustered the expression data and obtained five clusters that could be easily attributed to expected cell types: B cells, Monocytes, NK cells, and T cells (Fig. 3b). Rare cell types, such as dendritic cells or megakaryocytes that are known to occur in PBMCs at frequencies of ~0.5–1%, could not be detected, as expected from the low power to cluster 2–3 cells. For the detected cell types, known marker gene expression fits closely to previously described results[23] (Fig. 3c, d). Overall, we show that mcSCRB-seq is a powerful tool to highlight biological differences, already when a low number of cells are sequenced.

Discussion

In this work, we developed mcSCRB-seq, a scRNA-seq protocol utilizing molecular crowding. Based on benchmarking data generated from mouse ES cells, we show that mcSCRB-seq considerably increases sensitivity and decreases amplification bias due to the addition of PEG 8000 and the use of Terra polymerase, respectively. Furthermore, it shows no indication of bias for GC content and transcript lengths, and has low levels of crosstalk between cell barcodes, which has been seen especially in droplet-based RNA-seq approaches[23,25]. Compared to the previous SCRB-seq protocol, mcSCRB-seq increases the power to quantify gene expression twofold. Additionally, optimized reagents and workflows reduce costs by a factor of three. Qualitatively, we validate our protocol by sequencing PBMCs, a complex mixture of different cell types. We show that mcSCRB-seq can identify the different subpopulations and marker gene expression correctly and distinctively detect the major cell types present in the population.

In this context, we found that it was necessary to use different lysis conditions for the PBMCs than for mESCs. In our experience, some cell types may require a more stringent lysis buffer to stabilize mRNA, which might be a result of internal RNAses and/or lower RNA content. Therefore, we also provide an alternative lysis strategy for mcSCRB-seq to deal with more difficult cell types or samples.

Taken together, mcSCRB-seq is—to the best of our knowledge—not only the most sensitive protocol when benchmarked using ERCCs, it is also the most cost-efficient and flexible plate-based protocol currently available, and could be a valuable methodological addition to many laboratories, in particular as it requires no specialized equipment and reagents.

Methods

cDNA yield assay. For all optimization experiments, universal human reference RNA (UHRR; Agilent) was utilized to exclude biological variability. Unless otherwise noted, 1 ng of UHRR was used as input per replicate. Additionally, Proteinase K digestion and desiccation were not necessary prior to reverse transcription. In order to accommodate all the reagents, the total volume for reverse transcription was increased to 10 µl. All concentrations were kept the same, with the exception that we added the same total amount of reverse transcriptase (25 U), thus lowering the concentration from 12.5 to 2.5 U/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), Enz-Script (Biozym), ProtoScript II (New England Biolabs), Superscript II (Thermo Fisher), GoScript (Promega), Revert UP II (Biozym), and M-MLV Point Mutant (Promega), were compared to determine which enzyme yielded the most cDNA. Several dilutions ranging from 1 to 1000 pg of universal human reference RNA (UHRR; Agilent) were used as input for the RT reactions.

RT reactions contained final concentrations of 1 × 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.

All reactions were amplified using 25 PCR cycles to be able to detect low inputs.

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

Effect of reaction enhancers. In order to improve the efficiency of the RT, we tested the addition of reaction enhancers, including $MgCl_2$, 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 the presence or absence of 1 M betaine (Sigma-Aldrich). Furthermore, the addition of 1 M betaine and 0.6 M trehalose (Sigma-Aldrich) was compared to the standard RT protocol. Lastly, increasing concentrations of PEG 8000 (0, 3, 6, 9, 12, and 15% W/V) were also tested.

Comparison of PCR DNA polymerases. The following 12 DNA polymerases were evaluated in preamplification: KAPA HiFi HotStart (KAPA Biosystems), SeqAmp (Clontech), Terra direct (Clontech), Platinum SuperFi (Thermo Fisher), Precisor (Biocat), Advantage2 (Clontech), AccuPrime Taq (Invitrogen), Phusion Flash (Thermo Fisher), AccuStart (QuantaBio), PicoMaxx (Agilent), FideliTaq (Affymetrix), and Q5 (New England Biolabs). For each enzyme, at least three replicates of 1 ng UHRR were reverse transcribed using the optimized molecular crowding reverse transcription in 10 µl reactions. Optimal concentrations for dNTPs, reaction buffer, stabilizers, and enzyme were determined using the manufacturer's recommendations. For all amplification reactions, we used the original SCRB-seq PCR cycling conditions[8].

Cell culture of mouse embryonic stem cells. J1[26] and JM8[27] mouse embryonic stem cells (mESCs) were provided by the Leonhardt lab (LMU Munich) and originally provided by Kerry Tucker (Ruprecht-Karls-University,Heidelberg) and by the European Mouse Mutant Cell repository (JM8A3; www.eummcr.org), respectively. They were used for the comparison of KAPA vs. Terra PCR amplification (Supplementary Fig. 5c) and the comparison of SCRB-seq and mcSCRB-seq, respectively. Both 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 µg/ml streptomycin (Thermo Fisher), 2 mM L-glutamine (Thermo Fisher), 1 × 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).

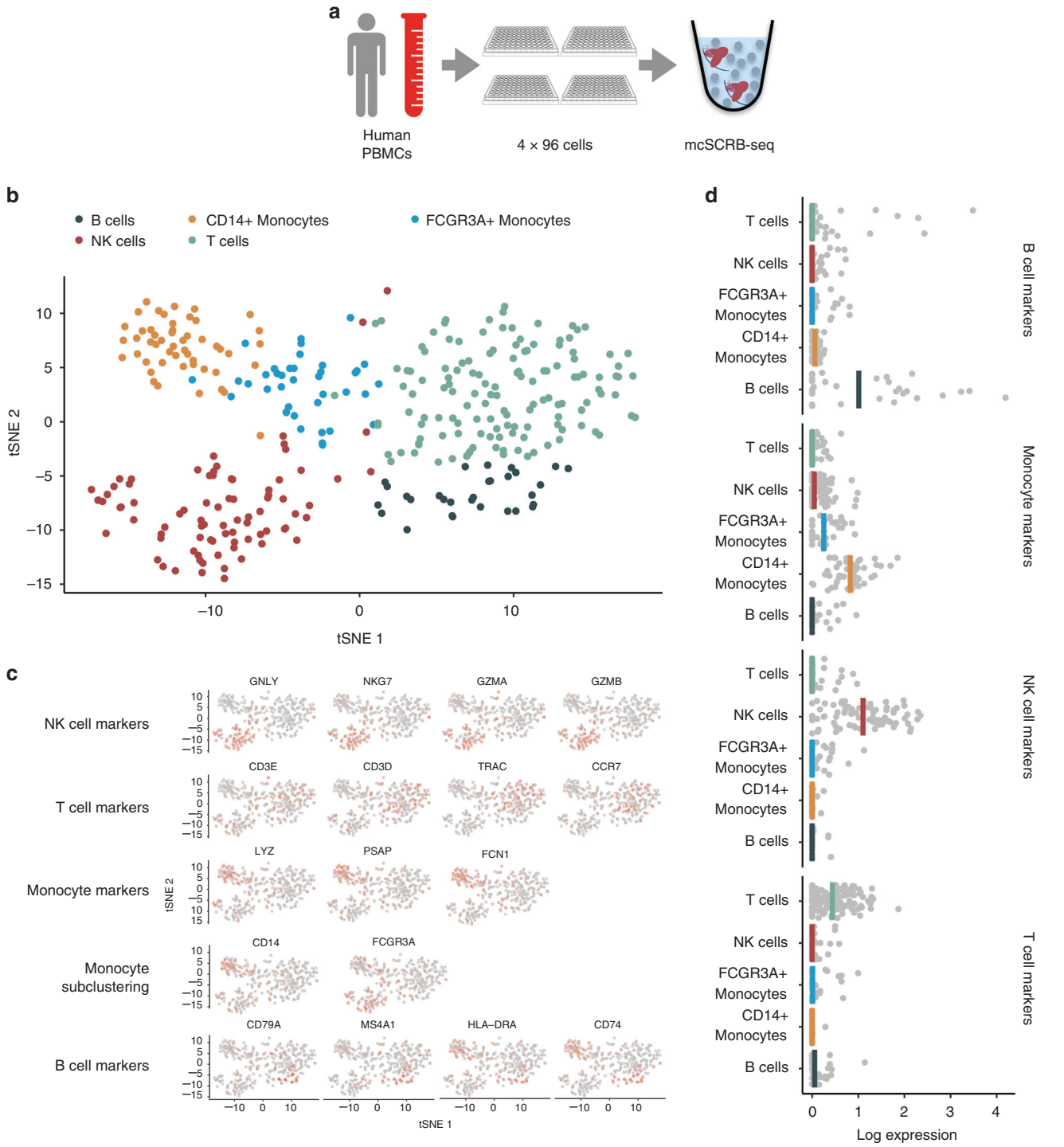


Fig. 3 mcSCRB-seq distinguishes cell types of peripheral blood mononuclear cells. **a** PBMCs were obtained from a healthy male donor and FACS sorted into four 96-well plates. Using the mcSCRB-seq protocol, sequencing libraries were generated. **b** tSNE projection of PBMC cells ($n = 349$) that were grouped into five clusters using the Seurat package[24]. Colors denote cluster identity. **c** tSNE projection of PBMC cells ($n = 349$) where each cell is colored according to its expression level of various marker genes for the indicated cell types. Expression levels were log-normalized using the Seurat package. **d** Marker gene expression from **c** was summarized as the mean log-normalized expression level per cell. B-cell markers: *CD79A*, *CD74*, *MS4A1*, *HLA-DRA*; Monocyte markers: *LYZ*, *PSAP*, *FCN1*, *CD14*, *FCGR3A*; NK-cell markers: *GNLY*, *NKG7*, *GZMA*, *GZMB*; T-cell markers: *CD3E*, *CD3D*, *TRAC*, *CCR7*

mESC cultures were confirmed to be free of mycoplasma contamination by a PCR-based test[28].

Cell culture of human-induced pluripotent stem cells. Human-induced pluripotent stem cells were generated using standard techniques from renal epithelial cells obtained from a healthy donor with written informed consent in accordance with the ethical standards of the responsible committee on human experimentation (216–08, Ethikkommission LMU München) and with the current (2013) version of the Declaration of Helsinki. hiPSCs were cultured under feeder-free conditions on Geltrex (Thermo Fisher)-coated dishes in StemFit medium (Ajinomoto) supplemented with 100 ng/ml recombinant human basic FGF (Peprotech) and 100 U/ml penicillin, 100 µg/ml streptomycin (Thermo Fisher). Cells were routinely passaged using 0.5 mM EDTA. Whenever cells were dissociated into single cells using 0.5 × TrypLE Select (Thermo Fisher), the culture medium was supplemented with 10 µM Rho-associated kinase (ROCK) inhibitor Y27632 (BIOZOL) to prevent apoptosis.

hiPSC cultures were confirmed to be free of mycoplasma contamination by a PCR-based test[28].

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 previously described[6,8]. 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. A full step-by-step protocol for mcSCRB-seq has been deposited in the protocols.io repository[29]. Briefly, cells were dissociated using trypsin and resuspended in PBS. Single cells ("3 drops" purity mode) were sorted into 96-well DNA LoBind plates (Eppendorf) containing 5 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 min. Proteinase K was then heat inactivated for 10 min at 80 °C. Next, 5 µl reverse transcription master mix consisting of 20 units Maxima H- enzyme (Thermo Fisher), 2 × 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 min 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) 1.66 × 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 s at 98 °C, 30 s 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. Sixteen bases were sequenced with the first read to obtain cellular and molecular barcodes and 50 bases were sequenced in the second read into the cDNA fragment. When several libraries were multiplexed on sequencing lanes, an additional 8 base i7 barcode read was done.

Primary data processing. All raw fastq data were processed using zUMIs together with STAR to efficiently generate expression profiles for barcoded UMI data[14,30]. For UHRR experiments, we mapped to the human reference genome (hg38) while mouse cells were mapped to the mouse genome (mm10) concatenated with the ERCC reference. Gene annotations were obtained from Ensembl (GRCh38.84 or GRCm38.75). Downsampling to fixed numbers of raw sequencing reads per cell were performed using the "-d" option in zUMIs.

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

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

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

Complex tissue analysis. PBMCs were obtained from a healthy male donor with written informed consent in accordance with the ethical standards of the responsible committee on human experimentation (216–08, Ethikkommission LMU München) and with the current (2013) version of the Declaration of Helsinki. Cells were sorted into 96-well plates containing 5 µl lysis buffer using a Sony SH800 sorter (Sony Biotechnology; 100 µm chip). Lysis buffer consisted of 5 M Guanidine hydrochloride (Sigma-Aldrich), 1% 2-mercaptoethanol (Sigma-Aldrich) and a 1:500 dilution of Phusion HF buffer (New England Biolabs). Before library preparation, each well was cleaned up using SPRI beads and resuspended in a mix of 5 µl reverse transcription master mix (see above) and 4 µl ddH_2O. After the addition of 1 µl 2 µM barcoded oligo-dT primer (E3V6NEXT, IDT), cDNA was synthesized according to the mcSCRB-seq protocol (see above). Pooling was performed by adding SPRI bead buffer. Sequencing and primary data analysis was performed as described above using the human reference genome (hg38). We retained only high-quality cells with at least 50,000 reads and a mapping rate above 75%. Furthermore, we discarded potential doublets that contained more than 40,000 UMIs and 5000 genes. Next, we used Seurat[24] to perform normalization (LogNormalize) and scaling. We selected the most variable genes using the "FindVariableGenes" command (1108 genes). Next, we performed dimensionality reduction with PCA and selected components with significant variance using the "JackStraw" algorithm. Statistically significant components were used for shared nearest-neighbor clustering (FindClusters) and tSNE visualization (RunTSNE). Log-normalized expression values were used to plot marker genes.

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

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

For public data from Svensson et al.[5], we used their published molecular abundances calculated using the same logistic regression model obtained from Supplementary Table 2 (https://www.nature.com/nmeth/journal/v14/n4/extref/nmeth.4220-S3.csv). For Quartz-seq2[17], we obtained expression values for ERCCs from Gene Expression Omnibus (GEO; GSE99866), sample GSM2656466; for Chromium[23] we obtained expression tables from the 10 × 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[6], 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[22]. Parameter estimation of the negative binomial distribution was done using scran normalized counts at 500,000 raw reads per cell[31]. Next, we simulated two-group comparisons with 10% differentially expressed genes. Log2 fold-changes were drawn from a normal distribution with a mean of 0 and a standard deviation of 1.5. In each of the 25 simulation iterations, we draw equal sample sizes of 24, 48, 96, 192 and 384 cells per group and test for differential expression using ROTS[32] and scran normalization[31].

Batch effect analysis. In order to detect genes differing between batches of one scRNA-seq protocol, data were normalized using scran[31]. Next, we tested for differentially expressed genes using limma-voom[33,34]. Genes were labeled as significantly differentially expressed between batches with Benjamini–Hochberg adjusted p values <0.01.

Code availability. Analysis code to reproduce major analyses can be found at https://github.com/cziegenhain/Bagnoli_2017.

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

Further data including cDNA yield of optimization experiments is available on GitHub (https://github.com/cziegenhain/Bagnoli_2017). A detailed step-by-step protocol for mcSCRB-seq has been submitted to the protocols.io repository (mcSCRB-seq protocol 2018). All other data available from the authors upon reasonable request.

Received: 22 December 2017 Accepted: 26 June 2018
Published online: 26 July 2018

References

1. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
2. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
3. 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).
4. Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I. & Enard, W. Quantitative single-cell transcriptomics. *Brief. Funct. Genomics* https://doi.org/10.1093/bfgp/ely009 (2018).
5. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
6. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 (2017).
7. Menon, V. Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Brief. Funct. Genomics* https://doi.org/10.1093/bfgp/ely001 (2018).
8. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. Preprint at https://doi.org/10.1101/003236 (2014).
9. 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).
10. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
11. Zimmerman, S. B. & Pheiffer, B. H. Macromolecular crowding allows blunt-end ligation by DNA ligases from rat liver or *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **80**, 5852–5856 (1983).
12. Rivas, G. & Minton, A. P. Macromolecular crowding in vitro, in vivo, and in between. *Trends Biochem. Sci.* **41**, 970–981 (2016).
13. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
14. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs - a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, giy059 (2018).
15. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 25533 (2016).
16. Quail, M. A. et al. Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* **9**, 10–11 (2012).
17. Sasagawa, Y. et al. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, 29 (2018).
18. Dixit, A. Correcting chimeric crosstalk in single cell RNA-seq experiments. Preprint at https://doi.org/10.1101/093237 (2016).
19. Baker, S. C. et al. The external RNA controls consortium: a progress report. *Nat. Methods* **2**, 731–734 (2005).
20. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
21. Phipson, B., Zappia, L. & Oshlack, A. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res.* **6**, 595 (2017).
22. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33**, 3486–3488 (2017).
23. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
24. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
25. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. Preprint at https://doi.org/10.1101/303727 (2018).
26. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).
27. Pettitt, S. J. et al. Agouti C57BL/6N embryonic stem cells for mouse genetic resources. *Nat. Methods* **6**, 493–495 (2009).
28. Young, L., Sung, J., Stacey, G. & Masters, J. R. Detection of mycoplasma in cell cultures. *Nat. Protoc.* **5**, 929–934 (2010).
29. Bagnoli, J., Ziegenhain, C., Janjic, A., Wange, L. E. & Vieth, B. mcSCRB-seq protocol. *protocols.io* https://doi.org/10.17504/protocols.io.nrkdd4w (2018).
30. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
31. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
32. Seyednasrollah, F., Rantanen, K., Jaakkola, P. & Elo, L. L. ROTS: reproducible RNA-seq biomarker detector—prognostic markers for clear cell renal cell cancer. *Nucleic Acids Res.* **44**, e1 (2015).
33. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
34. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

Acknowledgements

We thank Ines Bliesener for expert technical assistance. We are grateful to Magali Soumillon and Tarjei Mikkelsen for providing the original SCRB-seq protocol and to Stefan Krebs and Helmut Blum for sequencing. We would like to thank Elena Winheim for the PBMC sample. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through LMUexcellent and the SFB1243 (Subproject A14/A15).

Author contributions

C.Z. and W.E. conceived the study. J.W.B., C.Z., A.J. and L.E.W. performed experiments and prepared sequencing libraries. J.G. and J.W.B. cultured mouse ES and human iPS cells. Sequencing data were processed by S.P. and C.Z. J.W.B., C.Z., A.J. and B.V. analyzed the data. J.W.B., C.Z., A.J., I.H. and W.E. wrote the manuscript.

Additional information

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

Competing interests: The authors declare no competing interests.

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

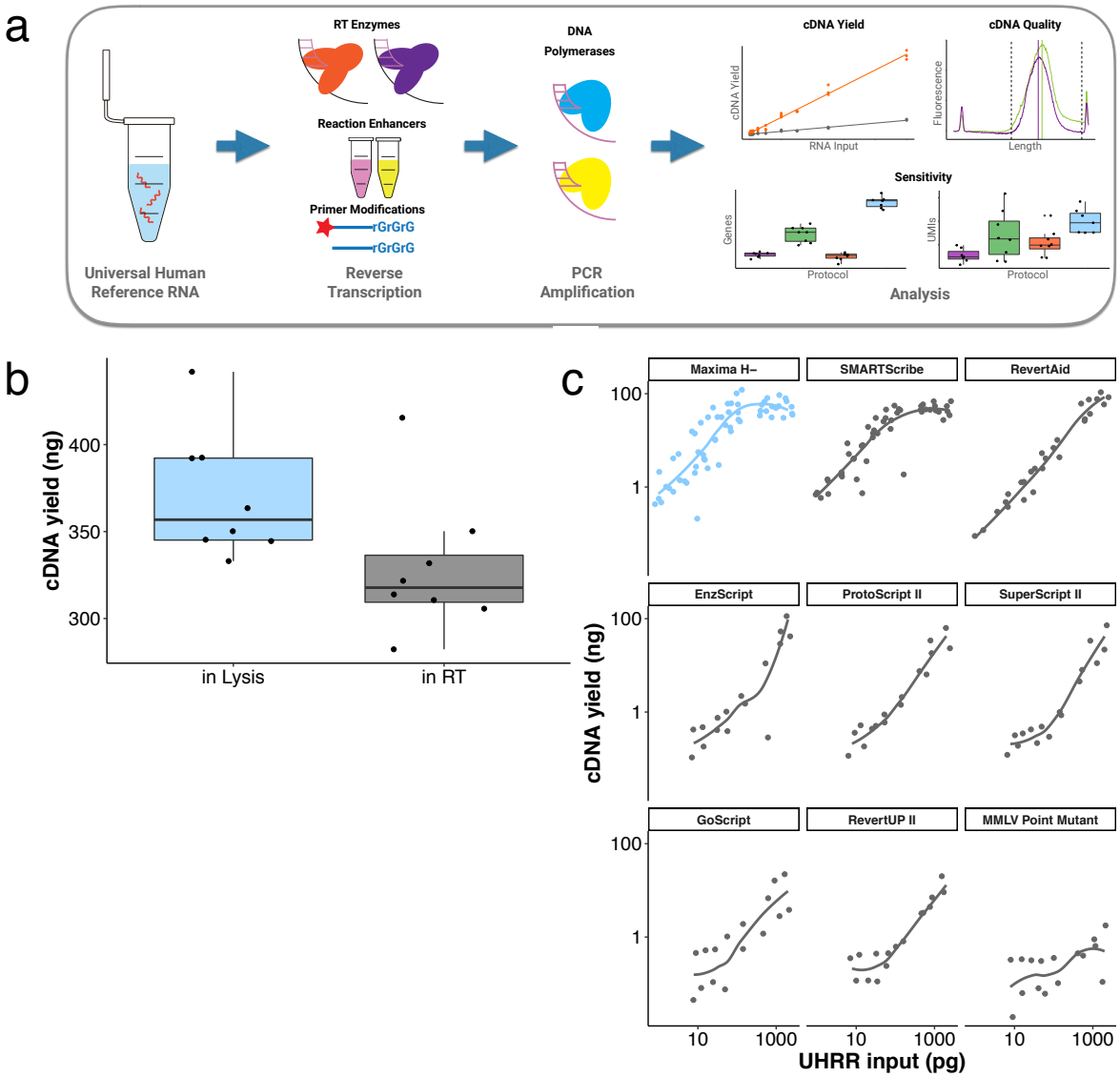
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information

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

Bagnoli et al.

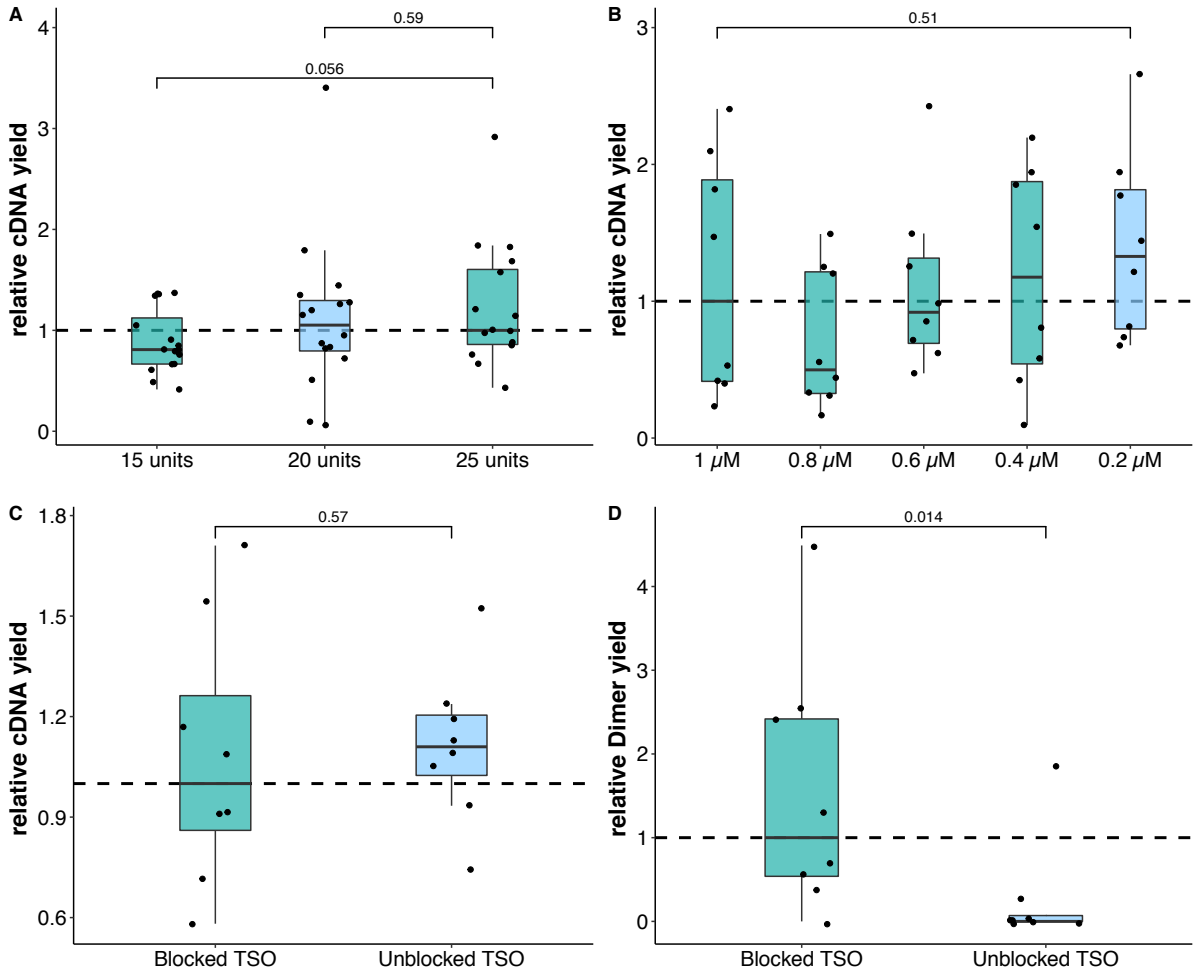
Supplementary Figure 1



Supplementary Figure 1: Schematic overview and optimization of reverse transcription

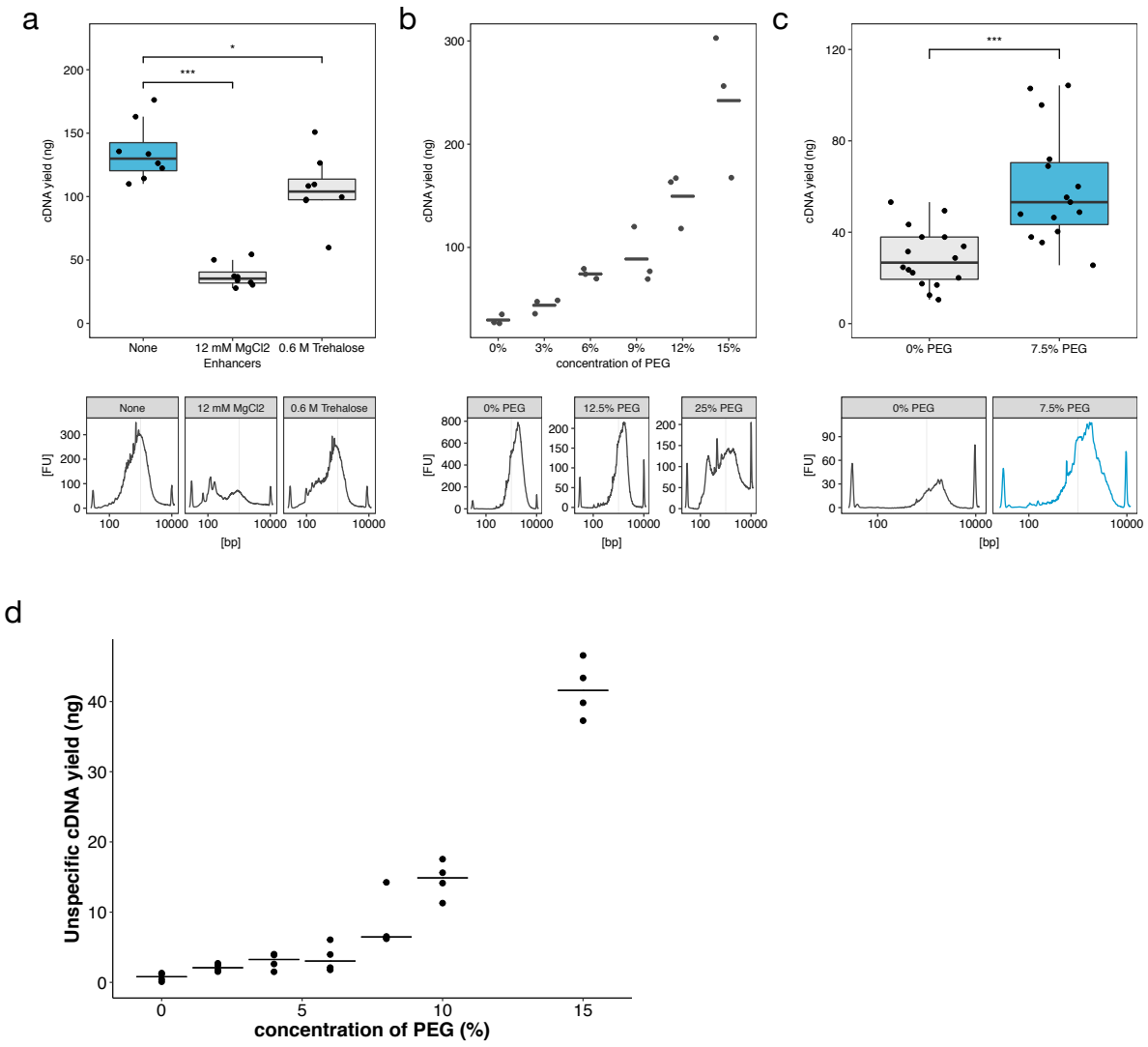
a) Low amounts (1-1000pg) of universal human reference RNA (UHRR) were used in optimization experiments. We assessed components affecting reverse transcription and PCR amplification with respect to cDNA yield and cDNA quality and verified effects on gene and transcript sensitivity by sequencing scRNA-seq libraries to develop the mcSCRB-seq protocol.
b) cDNA yield (ng) after reverse transcription with oligo-dT primers already in the lysis buffer ("in Lysis") or separately added before reverse transcription ("in RT"). Each dot represents a replicate and each box represents the median and first and third quartiles. The condition selected for the final mcSCRB-seq protocol is highlighted in blue.
c) cDNA yield (ng) dependent on varying UHRR input using 9 different RT enzymes. Each dot represents a replicate. Lines were fitted using local regression. The condition selected for the final mcSCRB-seq protocol is highlighted in blue.

Supplementary Figure 2



Supplementary Figure 2: Optimization of reverse transcription conditions.
Shown are relative cDNA yields after reverse transcription and PCR amplification of UHRR using:
a) varying amounts of reverse transcriptase enzyme (15-25 units, Maxima H-; 1 ng UHRR input per replicate)
b) varying amounts of oligo-dT primer (E3V6; 1 ng UHRR input per replicate)
c) blocked or unblocked Template switching oligo (TSO, E5V6; 10 pg UHRR per replicate)
d) relative primer dimer yield using blocked or unblocked Template switching oligo (TSO, E5V6) estimated using no-input controls (see Methods).
All values are relative to the median of the condition used in the original SCRB-seq protocol[1], which is indicated by a dashed horizontal line. Each dot represents a replicate and each box represents the median and first and third quartiles method. Numbers above boxes indicate p-values (Welch Two Sample t-test).
Optimized conditions selected for the mcSCRB-seq protocol are marked in blue.

Supplementary Figure 3



Supplementary Figure 3: Reverse transcription yield is increased by molecular crowding.

cDNA yield as well as representative length distributions (Bioanalyzer traces, bottom) using various additives in the reverse transcription and template switching reaction.

Each dot represents a replicate, lines represent the median and boxes the first and third quartile. Stars above boxes indicate p-values < 0.05 (Welch Two Sample t-test)

a) Influence of MgCl2 and Trehalose on cDNA synthesis (1 ng UHRR input per replicate; 21 PCR cycles).

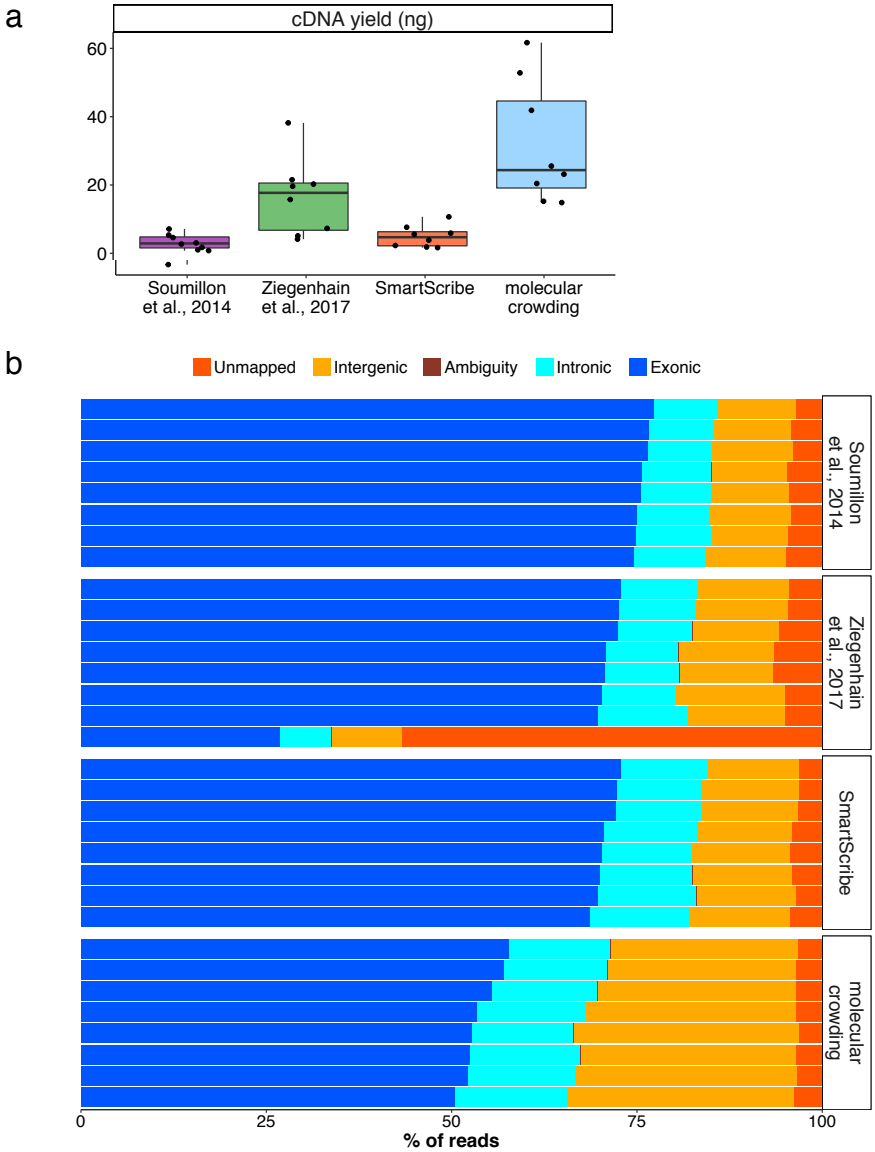
b) Concentration-dependent influence of PEG 8000 on cDNA yield (100 pg UHRR input per replicate; 23 PCR cycles).

c) Effect of 7.5% PEG 8000 (100 pg UHRR input per replicate; 23 PCR cycles).

d) Concentration-dependent generation of unspecific reverse transcription products (0 pg UHRR input per replicate; 23 PCR cycles).

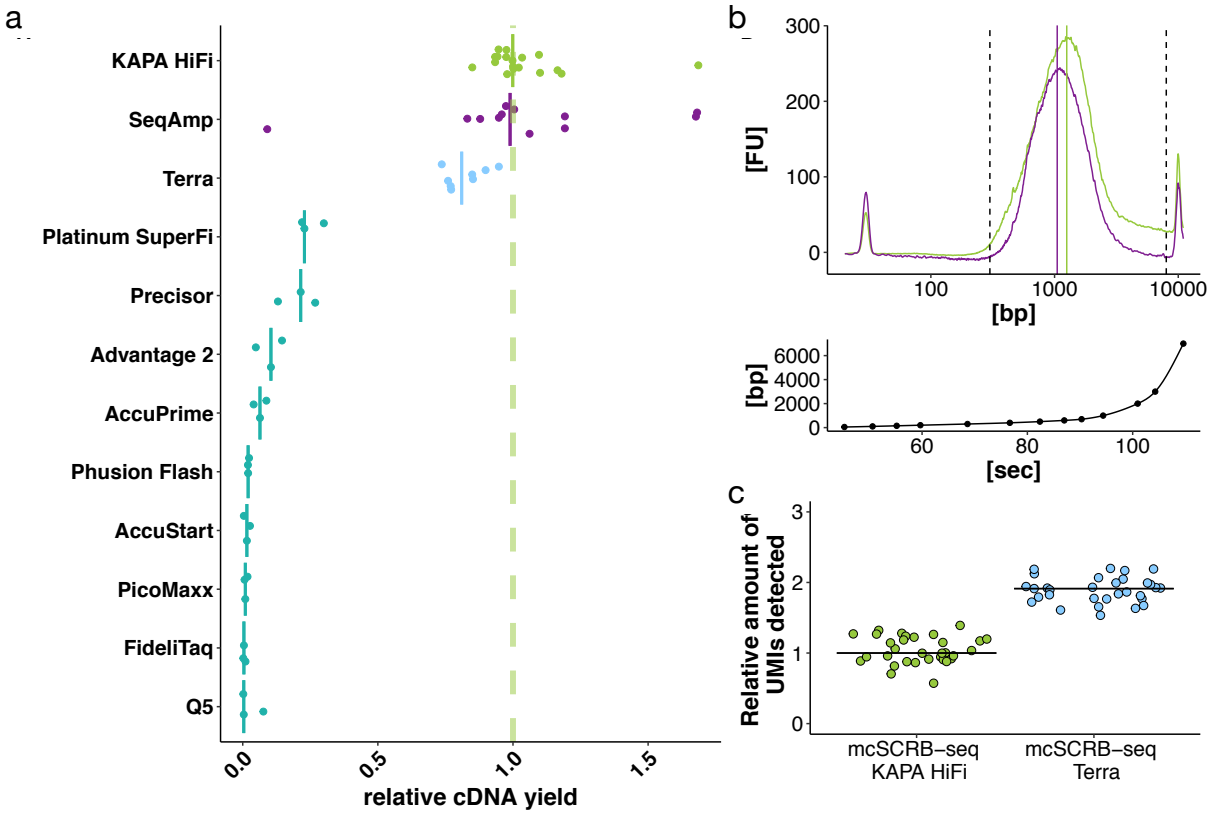
The conditions selected for the final mcSCRB-seq protocol are highlighted in blue.

Supplementary Figure 4



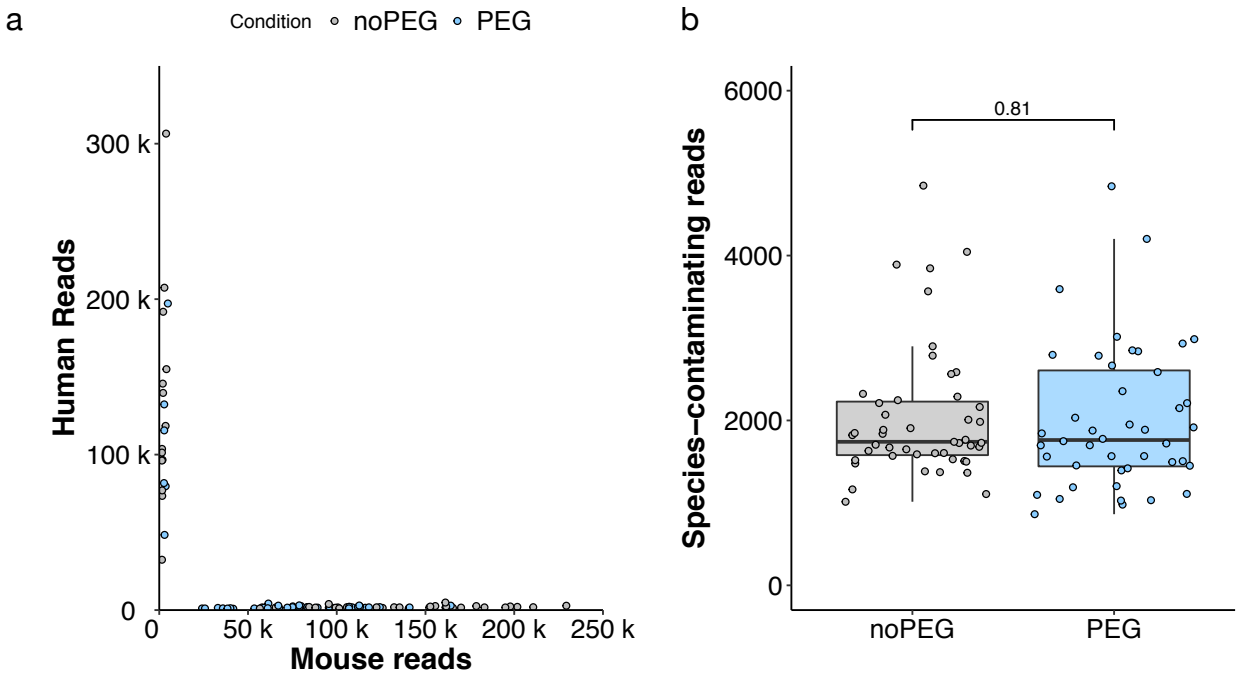
Supplementary Figure 4: Sequencing of UHRR samples.
10 pg of UHRR where used as input for eight replicates for each of the four protocol variants (Supplementary Table 1).
a) cDNA yield (ng) after PCR amplification per method. Each dot represents a replicate and each box represents the median and first and third quartiles per method.
b) Libraries were generated and sequenced from the above cDNA, downsampled to one million reads per library and mapped. Shown are the percentage of sequencing reads that cannot be mapped to the human genome (red), mapped to ambiguous genes (brown), mapped to intergenic regions (orange), inside introns (teal) or inside exons (blue).
Note the higher fraction of reads mapping to intergenic regions, especially in the molecular crowding condition. As UHRR is provided as DNAse-digested RNA, these reads are likely derived from endogenous transcripts, although it is unclear why these are proportionally more detected than annotated transcripts only in the molecular crowding protocol. This is also not generally observed for molecular crowding conditions, as SCRB-seq and mcSCRB-seq protocols have the same fraction (~25%) of intergenic reads mapped when single mouse ES cells are used (Supplementary Figure 7c).

Supplementary Figure 5



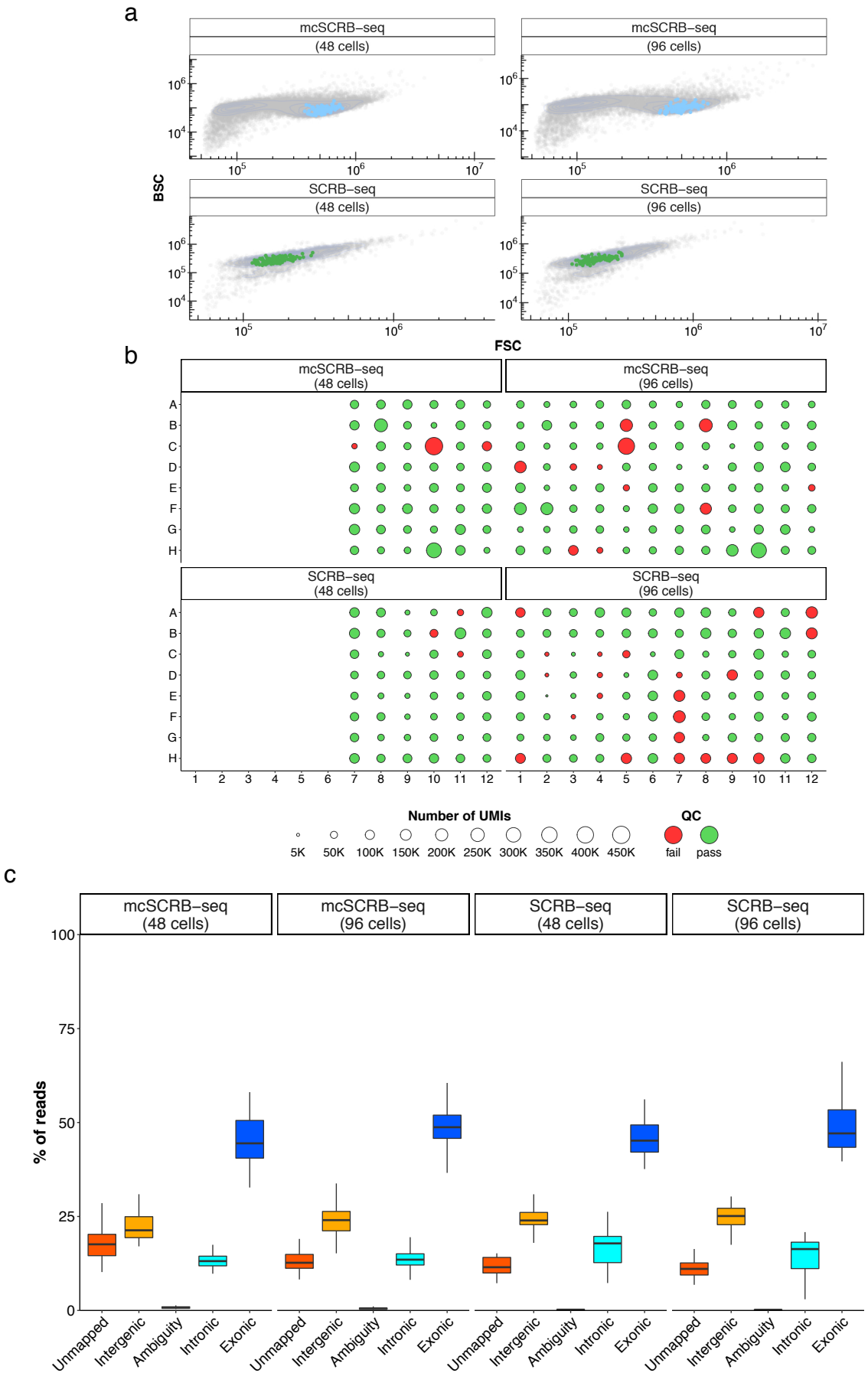
Supplementary Figure 5: Optimization of PCR amplification.
a) Relative cDNA yield after reverse transcription of 1 ng UHRR and amplification using different polymerase enzymes or ready mixes. All values are relative to the median of KAPA HiFi which is indicated by a dashed vertical line, as this was used in the SCRB-seq protocol variant of Ziegenhain et al.[2]. Solid vertical lines indicate the median for each polymerase.
b) Top: Representative length quantification of cDNA libraries amplified with KAPA HiFi (green) or SeqAmp (purple) as quantified by capillary gel electrophoresis (Agilent Bioanalyzer). Solid vertical lines depict the ranked mean length for each library within the region marked with dashed vertical lines. Bottom: Depiction of time length model (spline fit) used to analyze capillary gel electrophoresis via the ladder. Each dot represents a ladder peak with known length (bp) and measurement time (sec).
c) Relative amount of detected UMIs in single mESCs (J1) downsampled to 1 million reads using KAPA-HiFi or Terra for cDNA amplification. For both conditions, molecular crowding conditions (7.5% PEG 8000) were used during reverse transcription. Each dot represents a cell and horizontal lines indicate the median per polymerase.

Supplementary Figure 6



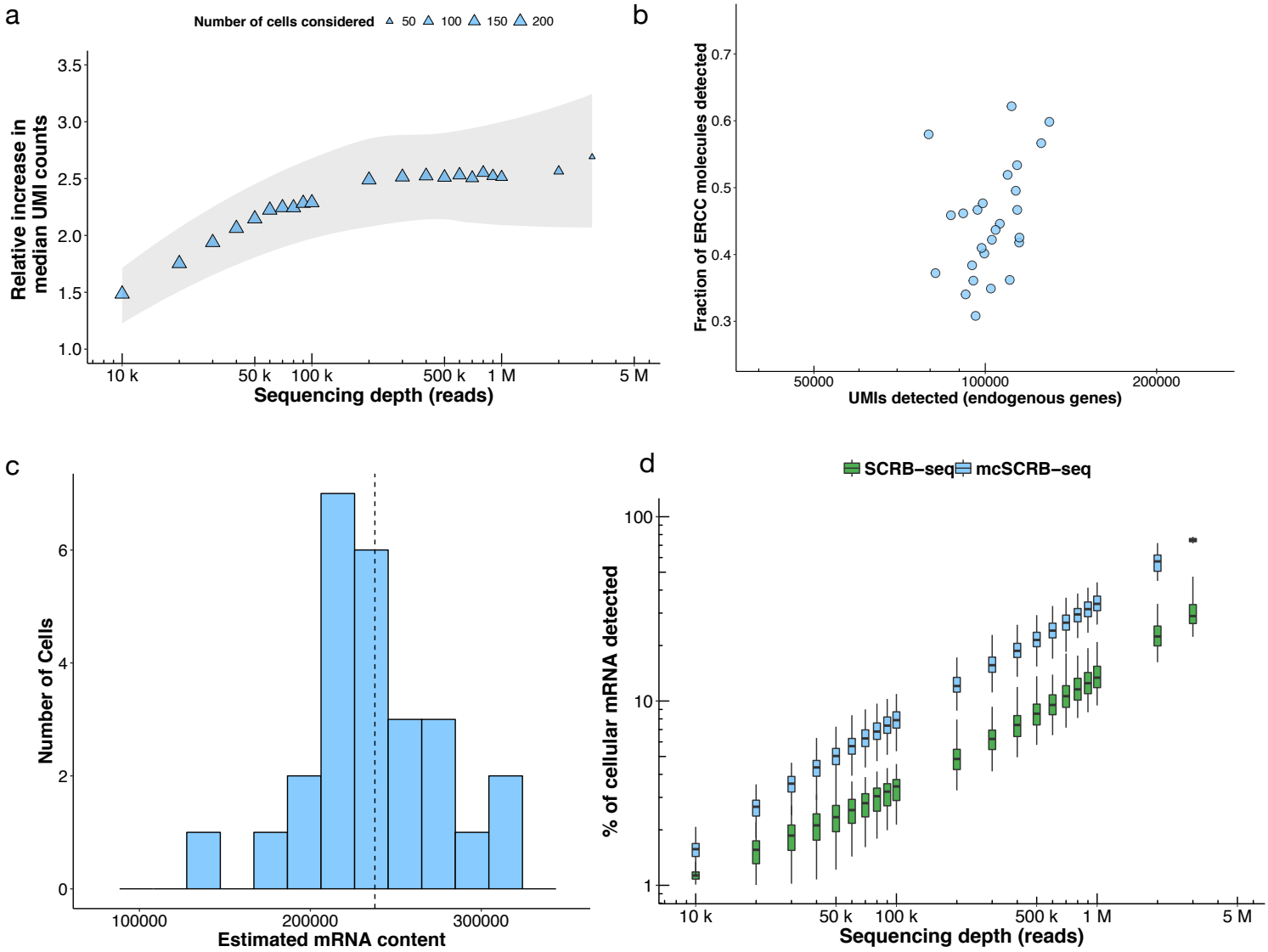
Supplementary Figure 6: Species mixing experiment for mcSCRB-seq
Human induced pluripotent stem cells and Mouse embryonic stem cells were mixed and sorted in a 96-well plate. cDNA was synthesized using the mcSCRB-seq protocol in absence and presence of PEG.
a) For each cell barcode, uniquely aligning reads to human or mouse gene features are shown in a dot plot. No doublets were observed, as expected from single-cell purity FACS sorting.
b) Each cell barcode was classified to be a human or mouse cell. Shown are the number of reads aligning to the wrong species for each of the cell barcodes. There is no significant difference between the protocols with and without PEG (two-sided t-test, p-value=0.81).

Supplementary Figure 7



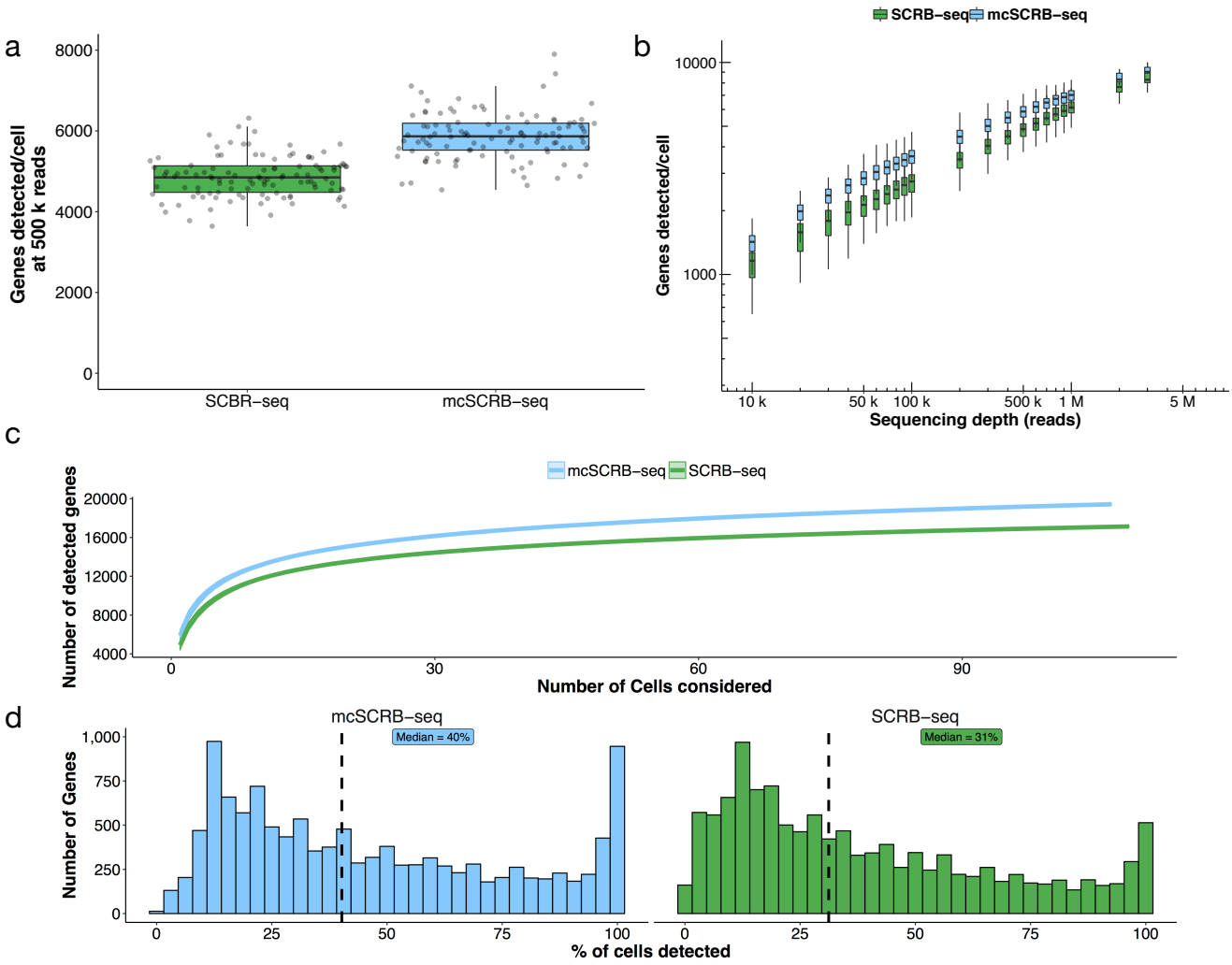
Supplementary Figure 7: Libraries from single mESCs generated with mcSCRB-seq and SCRB-seq protocols.
a) Scatter plots showing FACS data with forward (FS(c) and backward (BS(c) scatter intensities of one vial of mESCs (JM8) resuspended in PBS (mcSCRB-seq) or resuspended in RNAProtect Cell Reagent (SCRB-seq). Each dot represents an event. Coloured dots represent events that were sorted for scRNA-seq libraries in the four plates as depicted in **b**.
b) UMI counts for each cell by method (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).
c) Percentage of reads that cannot be mapped to the human genome (red), are mapped ambiguously (brown), are mapped to intergenic regions (orange), inside introns (teal) or inside exons (blue). Each box represents the median and first and third quartiles of cells that passed QC for each method.

Supplementary Figure 8



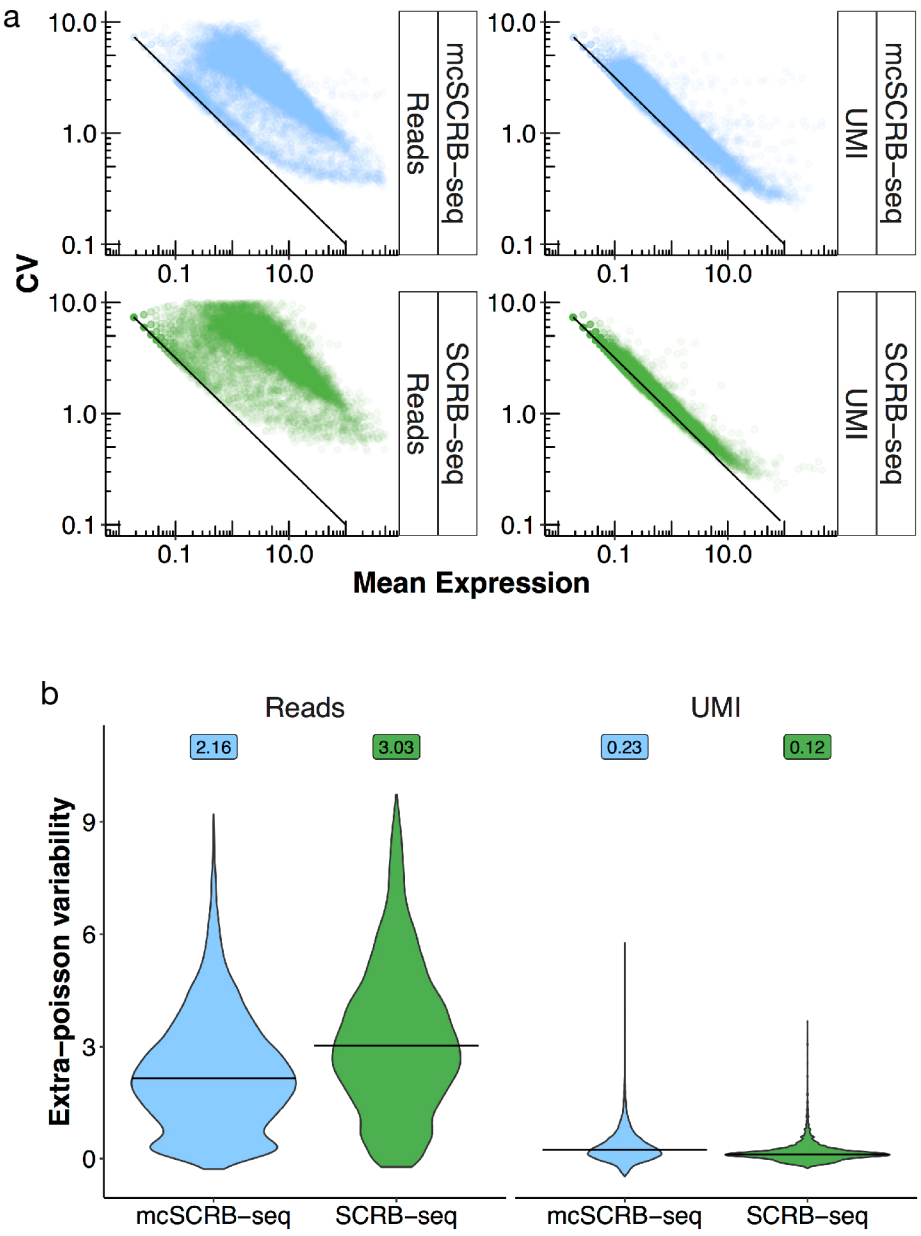
Supplementary Figure 8: Sensitivity of SCRB-seq and mcSCRB-seq protocols.
a) Relative increase in the median of detected UMIs dependent on raw sequencing depth (reads) using mcSCRB-seq compared to SCRB-seq. Each symbol represents the median over all cells at the given sequencing depth. The size of symbols depicts the number of cells (SCRB-seq + mcSCRB-seq) that were considered to calculate the median. The 95% confidence interval of a local regression model is depicted by the shaded area.
b) For each mcSCRB-seq cell that could be downsampled to 2 million reads, the number of UMIs from endogenous genes is plotted on the x axis (median at 102,282 UMIs per cell) and the fraction of UMI- ERCCs from the total amount of spiked-in ERCCs (70,000) is plotted on the y-axis (median 0.49). These values where used to calculate the histogram shown in
c) where for each cell the number of endogenous UMIs is divided by the fraction of ERCCs that were detected in that cell. Using the median of this distribution (dotted line) was set at 100% for the graph in
d) in which the percentage of cellular mRNAs is plotted for each cell at different sequencing depths.

Supplementary Figure 9



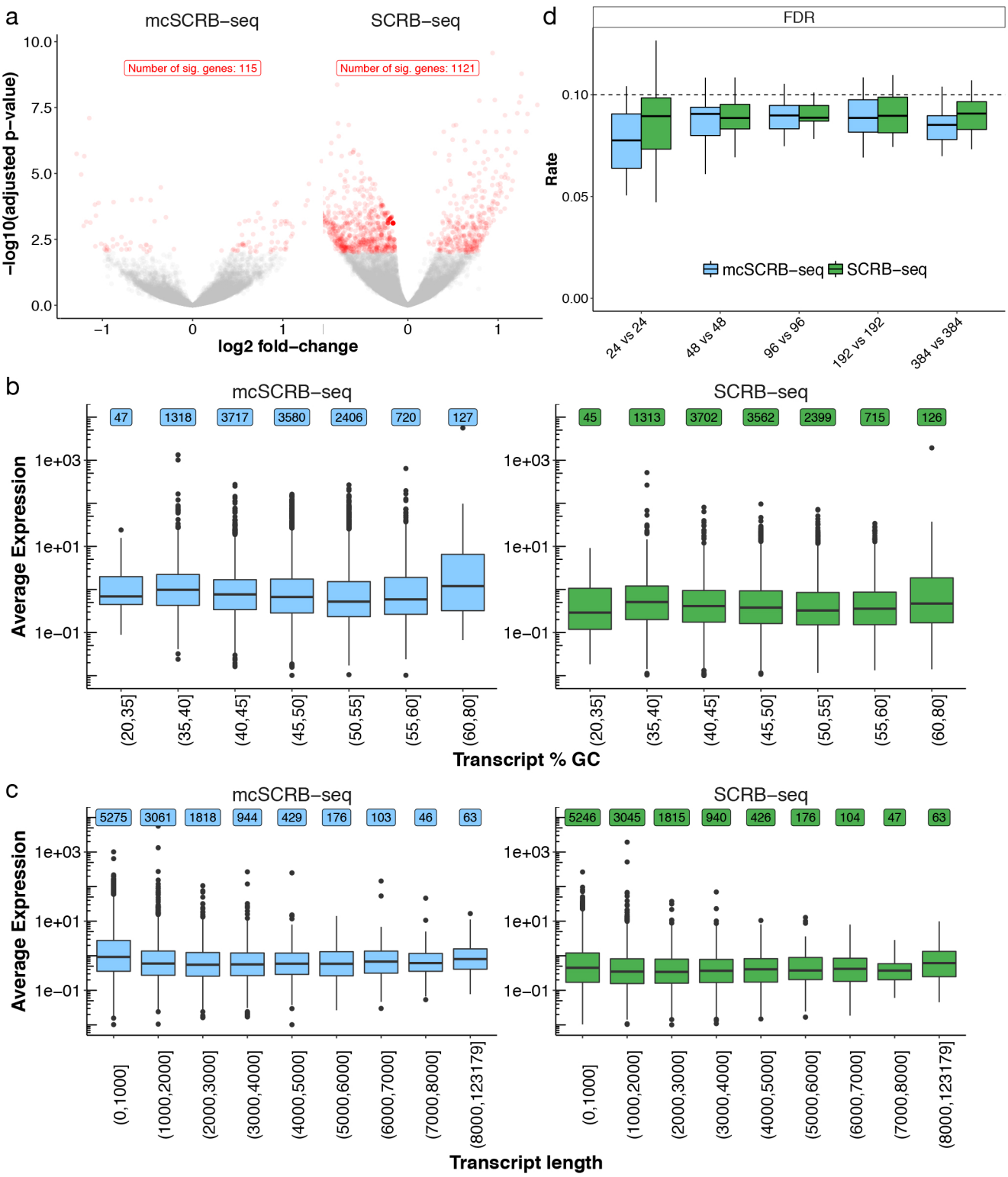
Supplementary Figure 9: Sensitivity of SCRB-seq and mcSCRB-seq protocols by genes.
a) Number of detected genes per cell and method (SCRB-seq/mcSCRB-seq) at a sequencing depth of 500,000 reads per cell (downsampled). Each dot represents a cell and each box represents the median and first and third quartiles.
b) Number of detected genes per cell and method (SCRB-seq/mcSCRB-seq) dependent on sequencing depth (reads). Each box represents the median and first and third quartiles per sequencing depth and method. Sequencing depths and genes are plotted on a logarithmic axis (base 10).
c) Number of detected genes at a sequencing depth of 500,000 reads per cell (downsampled) dependent on the number of cells considered.
d) Gene detection reproducibility is displayed as the fraction of cells detecting a given gene. Dashed line and label indicate the median of the distribution.

Supplementary Figure 10



Supplementary Figure 10: Variation parameters of SCRB-seq and mcSCRB-seq protocols by genes.
Variation and mean were calculated for each gene and method in cells downsampled to 500,000 reads using either UMIs per gene or reads per gene.
a) Gene-wise mean and coefficient of variation (standard deviation/mean) from all cells are shown as scatterplots for all methods based on read counts or UMIs. The black line indicates variance according to the poisson distribution.
b) Extra-Poisson variability across 12,086 reliably detected genes (detected in > 10% of cells) was calculated by subtracting the expected amount of variation due to Poisson sampling from the coefficient of variation (CV) measured in read-count or UMI quantification. Distributions are shown as violin plots and medians are shown as bars. Numbers indicate the median for each distribution.

Supplementary Figure 11



Supplementary Figure 11: Batch effects, biases and power analysis of SCRB-seq and mcSCRB-seq protocols

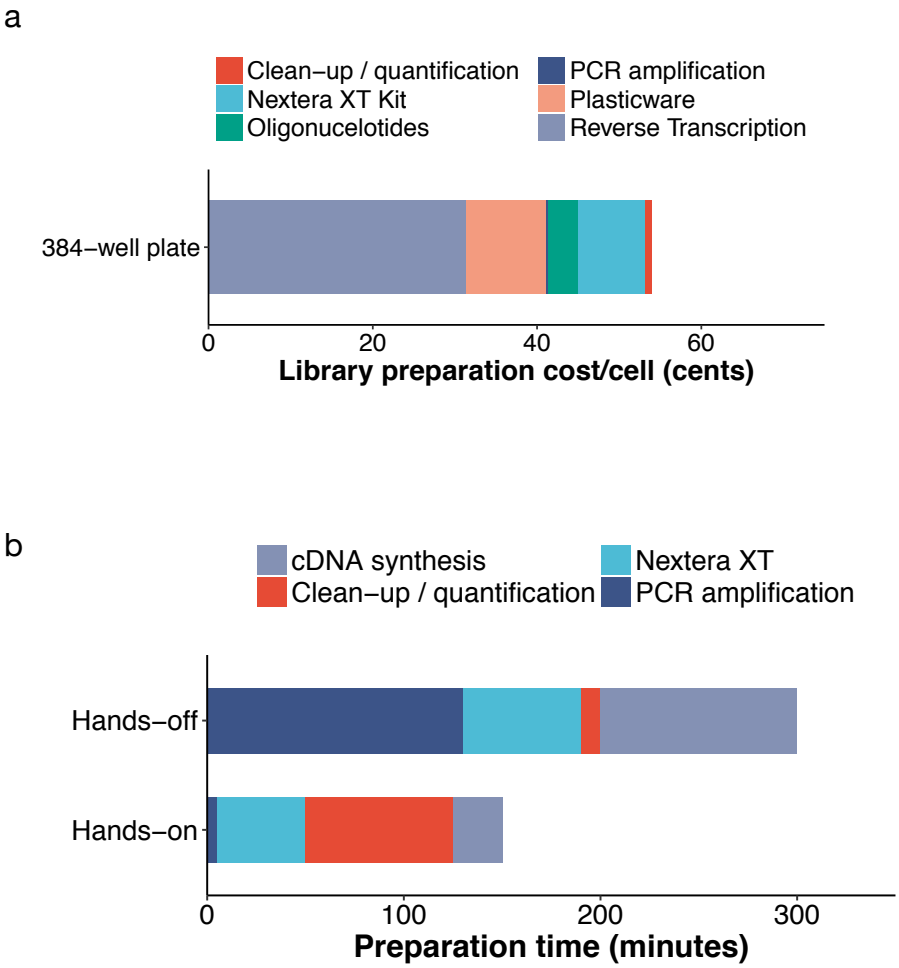
a) Volcano plots show differentially expressed genes between plates for each method. Points in red depict significantly differentially expressed genes (limma-voom; FDR < 0.01). Red labels show the number of differentially expressed genes between batches.

b) Average detected gene-wise expression levels (log normalized UMI) dependent on GC content of each transcript. Transcripts are grouped in 7 bins of GC content. Each dot represents an outlier and each box represents the median and first and third quartiles.

c) Average detected gene-wise expression levels (log normalized UMI) dependent on transcript length. Transcripts lengths are grouped in 7 bins and number of genes in each bin are indicated. Each dot represents an outlier and each box represents the median and first and third quartiles.

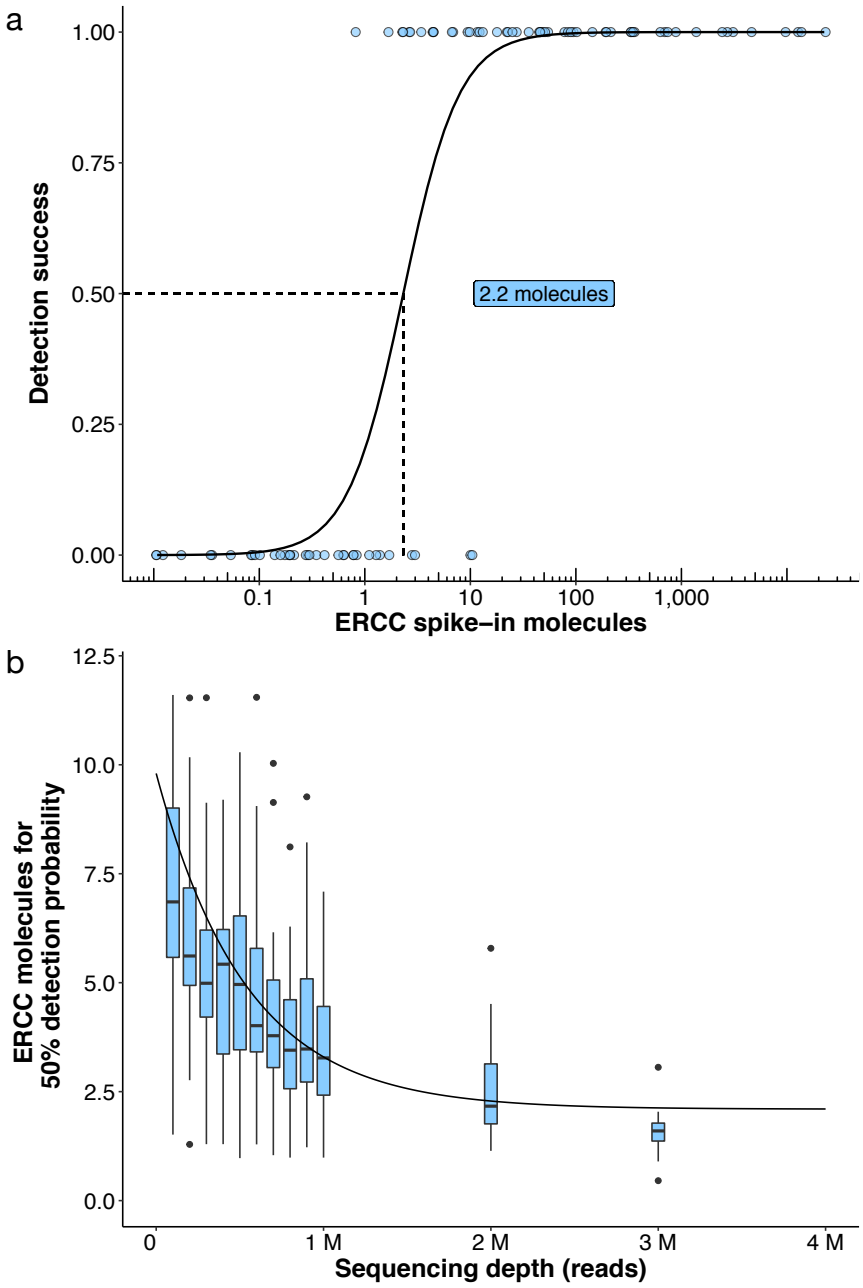
d) Power simulations were performed using the powsimR package[3] from empirical parameters estimated at 500,000 raw reads per cell. For SCRB-seq and mcSCRB-seq, we simulated n-cell two-group differential gene expression experiments with 10% differentially expressed genes. Shown is the false discovery rate ("FDR") for sample sizes
n = 24, n = 48, n = 96, n = 192 and n = 384 per group. The corresponding true positive rate is shown in Figure 2b. Boxplots represent the median and first and third quartiles of 25 simulations. Dashed lines indicate the desired nominal level.

Supplementary Figure 12



Supplementary Figure 12: Costs and preparation time of mcSCRB-seq
a) Library preparation costs (Eurocents) per cell. Colors indicate the consumable type based on list prices (see Supplementary Table 3). Costs also apply if four 96-well plates are pooled for PCR amplification and Nextera
b) Library preparation time for one 96-well plate of mcSCRB-seq libraries was measured for bench times ("Hands-on") and incubation times ("Hands-off"). Colors indicate the library preparation step. The total time was 7.5 hours. (see Supplementary Table 4)

Supplementary Figure 13

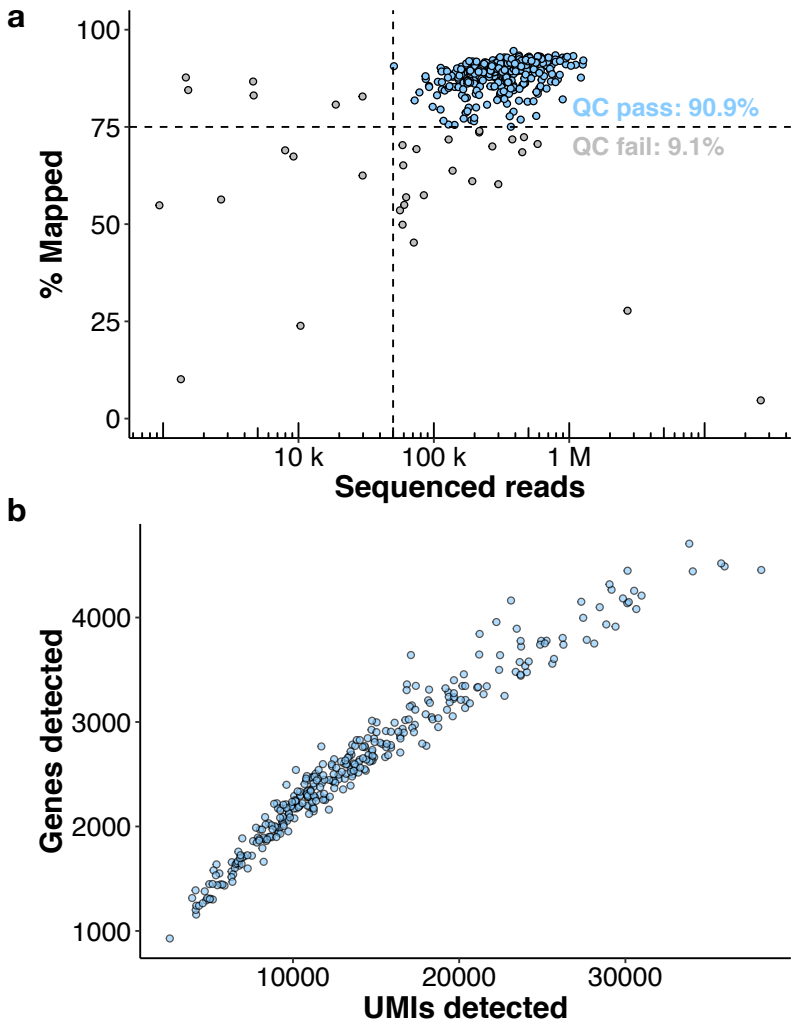


Supplementary Figure 13 : Comparison of mcSCRB-seq to other scRNA-seq data based on ERRCC spike-in detection probability

a) Shown is the detection (0 or 1) of the 92 ERCC transcripts in an average cell processed with mcSCRB-seq at 2 million reads coverage. Points and solid line represent the ERCC genes with their logistic regression model. Dashed lines and label indicate the number of ERCC molecules required for a detection probability of 50%.

b) Number of ERCC molecules required for 50% detection probability dependent on the sequencing depth (reads) for mcSCRB-seq. Each each box represents the median, first and third quartiles of cells per sequencing depth with dots marking outliers. A non-linear asymptotic fit is depicted as a solid black line.

Supplementary Figure 14



Supplementary Figure 14: Quality control of PBMC data

a) Scatter plot shows each of the 384 sequenced PBMC cells with the number of sequenced reads and the % of those reads mapped to the human genome. Dashed lines indicate quality filtering cut-offs chosen. Colors indicate QC passed cells (blue) or discarded cells (grey).

b) Cell-wise detected genes (>=1 UMI) and detected UMIs are shown for all cells that passed quality control (n=349).

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

Supplementary Table 1: 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 €

Supplementary Table 2: Overview of the key differences between SCRB-seq as used in Ziegenhain et al.[2] and mcSCRB-seq (this work).

Supplementary Table 3

consumable	price/unit	# 384 plates	price/384 plate
Barcode oligo-dT	24.000,00 €	5000	4,80 €
TSO E5V6unblocked	453,40 €	50	9,07 €
Maxima RT	554,00 €	5	110,80 €
Exonuclease I	327,00 €	1000	0,33 €
Clontech Terra	551,00 €	800	0,69 €
Nextera XT	3.002,00 €	96	31,27 €
dNTPs	1.236,00 €	125	9,89 €
Beads	20,00 €	10	2,00 €
Picogreen	542,00 €	400	1,36 €
PCR Seal	500,00 €	1000	0,50 €
PCR Plate/96	140,00 €	0	0,00 €
PCR Plate/384	195,00 €	25	7,80 €
Tips/96	36,50 €	0	0,00 €
Robotic tips/384	290,00 €	10	29,00 €
Total			207,50 €
Total/cell			**0,54 €**

Supplementary Table 3. Detailed overview of costs for mcSCRB-seq.

Supplementary Table 4

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

Supplementary Table 4. Detailed overview of hands-on and hands-off time necessary to create a sequenceable mcSCRB-seq library from one single cell plate.

Supplementary References

1. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv* (2014). doi:10.1101/003236

2. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* 65**,** 631–643.e4 (2017)

3. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: Power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx435

Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects.

nature biotechnology

ANALYSIS

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

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

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

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

Single-cell genomics provides an unprecedented view of the cellular makeup of complex and dynamic systems. Single-cell transcriptomic approaches in particular have led the technological advances that allow unbiased charting of cell phenotypes[1]. The latest improvements in scRNA-seq allow these technologies to scale to thousands of cells per experiment, providing comprehensive profiling of tissue composition[2,3]. This has led to the identification of new cell types[4–6] and the fine-grained description of cell plasticity in dynamic systems, such as development[7,8]. Recent large-scale efforts, such as the Human Cell Atlas (HCA) project[9], are attempting to produce cellular maps of entire cell lineages, organs and organisms[10,11] by conducting phenotyping at the single-cell level. The HCA project aims to advance our understanding of tissue function and to serve as a reference for defining variation in human health and disease. In addition to methods that capture the spatial organization of tissues[12,13], the main approach being used is scRNA-seq analysis of dissociated cells. Therefore, tissues are disaggregated and individual cells captured either by cell sorting or using microfluidic systems[1]. In sequential processing steps, cells are lysed, the RNA is reverse transcribed to complementary DNA, amplified and processed to sequencing-ready libraries.

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

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

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

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

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

Results

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

For a systematic comparison of protocols, we designed a reference sample containing human peripheral blood mononuclear cells (PBMCs) and mouse colon, which are tissue types with highly heterogeneous cell populations, as determined by previous single-cell sequencing studies[18,19]. In addition to the well-defined cell types, the tissues contain cells in transition states (for example, colon transit-amplifying (TA) or enterocyte progenitor cells) that show transcriptional differences during their differentiation trajectory[20]. The reference sample also included a wide range of cell sizes (for example, B cells: ~7 µm; HEK293 cells: ~15 µm) and RNA content, which are key parameters that affect performance in cell capture and library preparation. Interrogation of tissues from different species allowed us to pool a large variety of cell types in a single reference sample to maximize complexity while minimizing variability introduced during sample preparation. In addition to the intra-tissue complexity, the fluorescence-labeled, spiked-in cell lines allowed us to monitor cell-type composition during sample processing, and to identify batch effects and biases introduced during cell capture and library preparation.

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

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

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

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

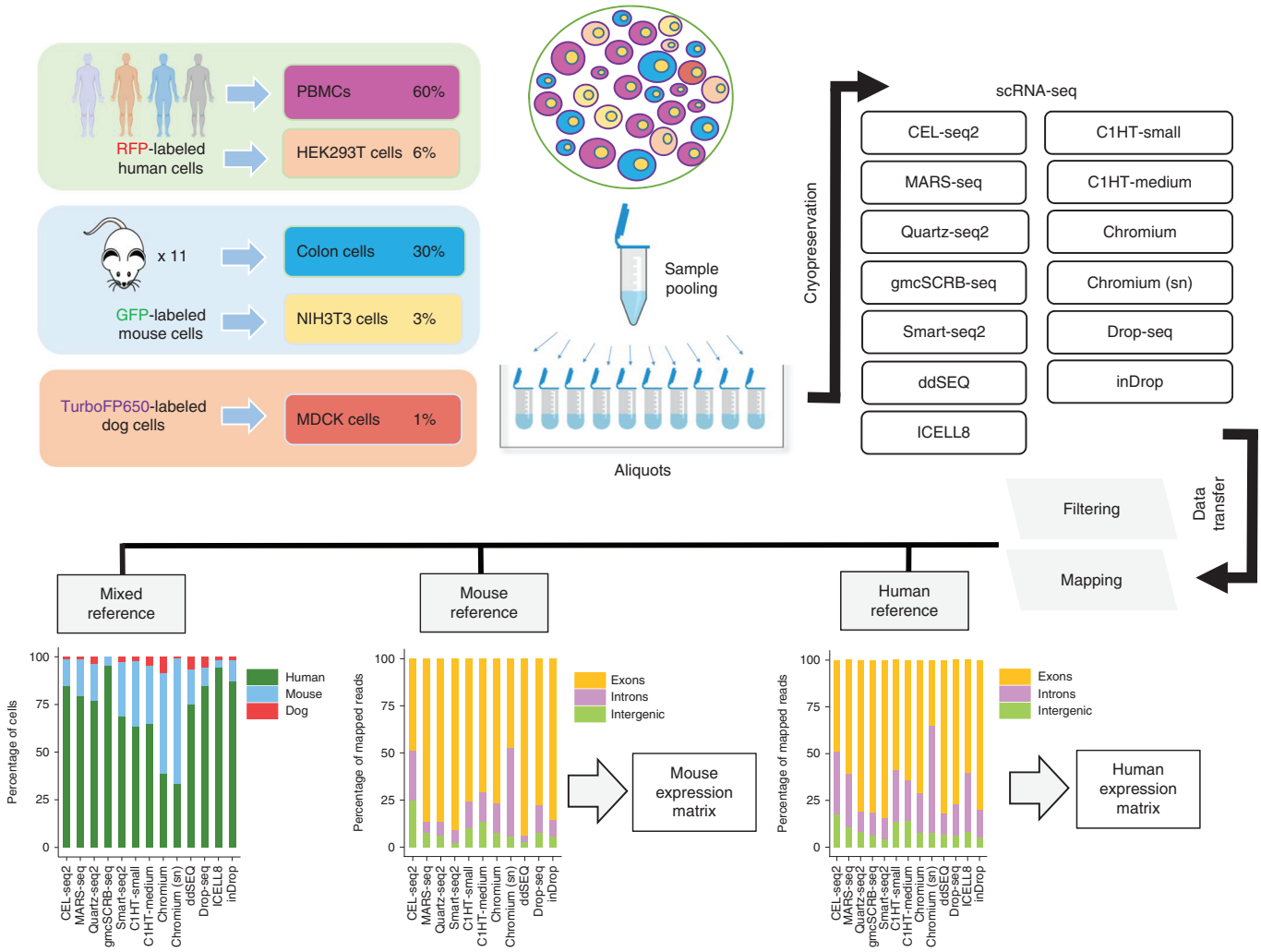


Fig. 1 | Overview of the experimental design and data processing. The reference sample consists of human PBMCs (60%), and HEK293T (6%), mouse colon (30%), NIH3T3 (3%) and dog MDCK cells (1%). The sample was prepared in one single batch, cryopreserved and sequenced by 13 different sc/snRNA-seq methods.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

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

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

Molecule-capture efficiency and library complexity. We produced reference datasets by analyzing 30,807 human and 19,749 mouse cells (Chromium v.2; Fig. 2a–c). The higher cell number allowed us to annotate the major cell types in our reference sample, and to extract population-specific markers (see Supplementary Table 2). It was noteworthy that the reference samples solely provided the basis to assign cell identities and gene marker sets, and were not used to quantify the method's performance. This strategy ensured that the choice of technology for deriving the reference does not influence downstream analyses. Cell clustering and reference-based cell annotation showed high agreement (average 83%; see Supplementary Table 3), and only cells with consistent annotations were used subsequently for comparative analysis at the cell-type level. The PBMCs (human) and colon cells (mouse) represented two largely different scenarios. Although the differentiated PBMCs clearly separated into subpopulations (for example, T/B cells, monocytes; Fig. 2b, and see Supplementary Figs. 3a and 4a–d), colon cells were ordered as a continuum of cell states that differentiate from intestinal stem cells into the main functional units of the colon (that is, absorptive enterocytes and secretory cells; Fig. 2c, and see Supplementary Figs. 3b and 5a–d). Notably, the subpopulation structure of our references was largely consistent with that of published datasets for human PBMCs[18] and mouse colon cells[22] (see Supplementary Figs. 6 and 7). After identifying major subpopulations and their respective markers in our reference sample, we clustered the cells of each sc/snRNA-seq protocol and annotated cell types using matchSCore2 (see Methods). This algorithm allows a gene marker-based projection of single cells (cell by cell) on to a

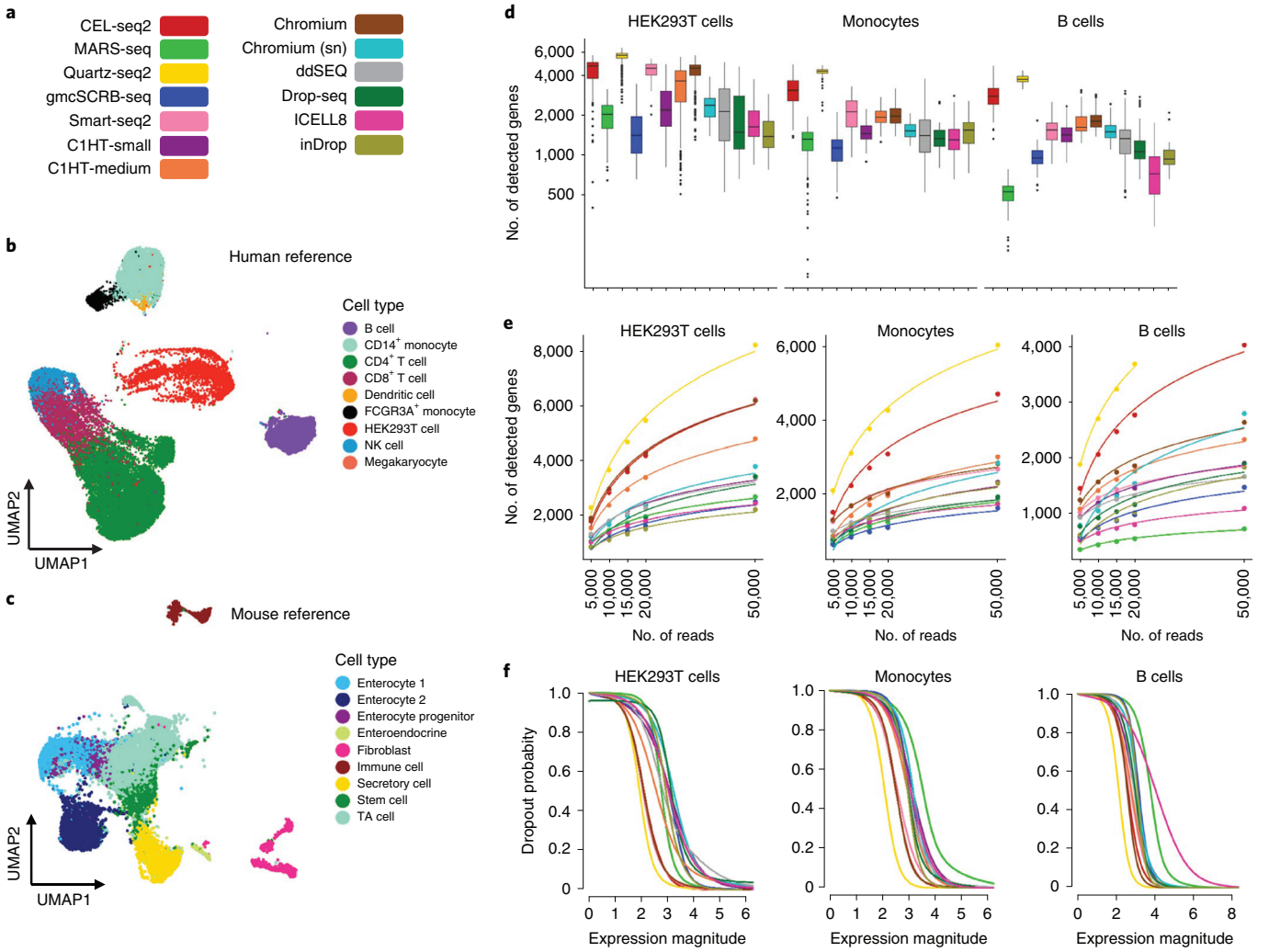


Fig. 2 | Comparison of 13 sc/snRNA-seq methods. a, Color legend of sc/snRNA-seq protocols. **b**, UMAP of 30,807 cells from the human reference sample (Chromium) colored by cell-type annotation. **c**, UMAP of 19,749 cells from the mouse reference (Chromium) colored by cell-type annotation. **d**, Boxplots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

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

To compare the efficiency of messenger RNA capture between protocols, we down-sampled the sequencing reads per cell to a common depth and stepwise-reduced fractions. Stochasticity introduced during down-sampling did not affect the reproducibility of the results (see Supplementary Fig. 10). Library complexity was determined separately for largely homogeneous cell types with markedly different cell properties and function, namely human HEK293T cells, monocytes and B cells (Fig. 2d,e), and mouse colon secretory and TA cells (see Supplementary Fig. 11a,b). We observed large differences in the number of detected genes and molecules across the protocols, with consistent trends across cell types and gene quantification strategies (see Supplementary Fig. 11c,d). Notably, some protocols, such as Smart-seq2 and Chromium v.2, performed better with higher RNA quantities (HEK293T cells) compared with lower starting amounts (monocytes and B cells), suggesting an input-sensitive optimum. Considering the different assay versions and application types of the Chromium system, a dedicated analysis showed increased detection of molecules and genes from nuclei to intact cells and toward the latest protocol versions (see Supplementary Fig. 12). Consistent with the variable library complexity, the protocols presented large differences in dropout probabilities (Fig. 2f), with Quartz-seq2, Chromium v.2 and CEL-seq2 showing consistently lower probability. Note that, despite the considerable differences between protocols, we observed a generally high technical reproducibility within the methods (see Supplementary Fig. 13).

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

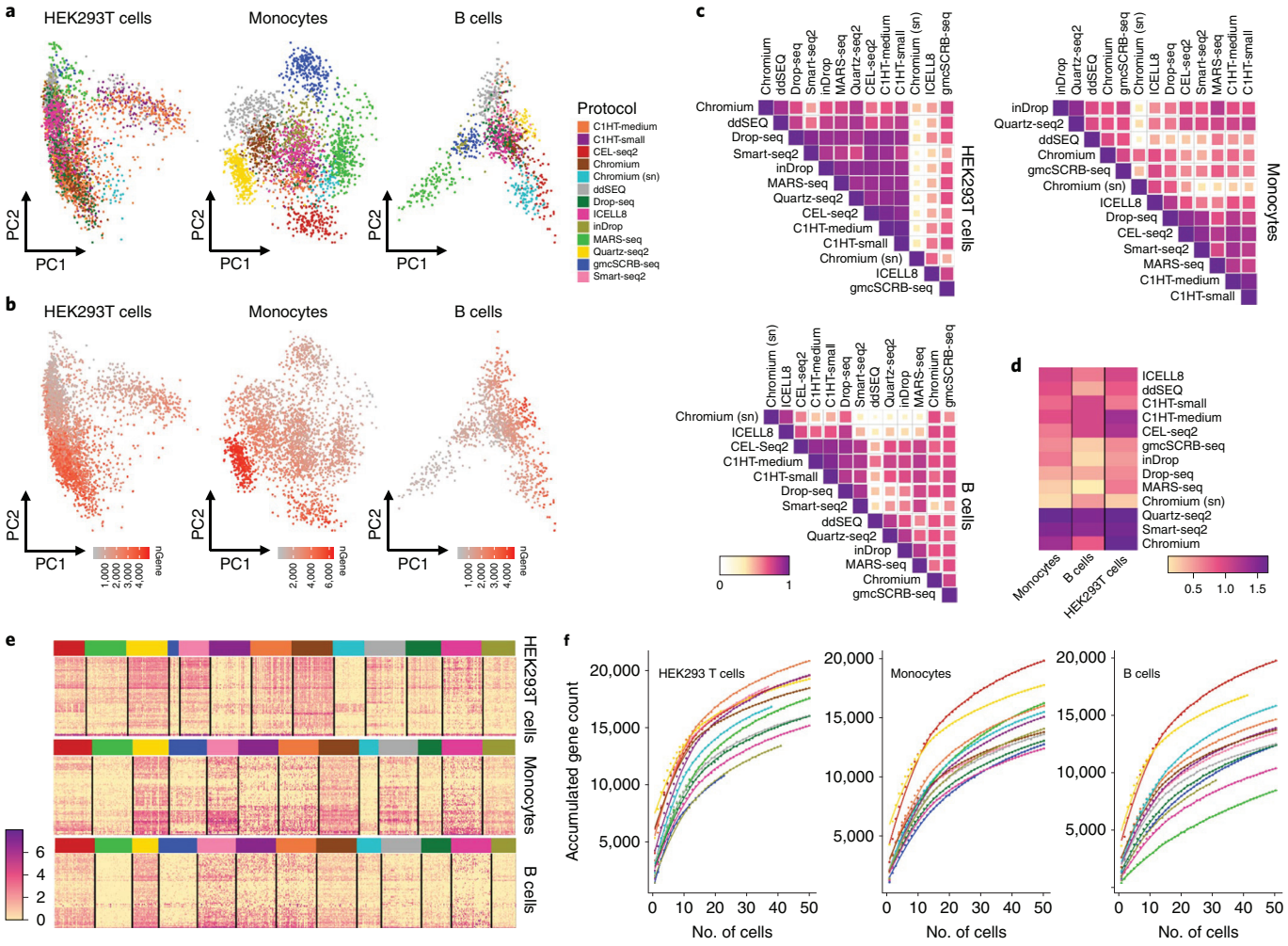


Fig. 3 | Similarity measures of sc/snRNA-seq methods. a,**b**, Principal component analysis on down-sampled data (20,000) using highly variable genes between protocols, separated into HEK293T cells, monocytes and B cells, and color coded by protocol (**a**) and number of detected genes per cell (**b**).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

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

Next, we quantified the similarities in information content of the protocols. Again, we used the down-sampled datasets and commonly expressed genes and calculated the correlation between methods in average transcript counts across multiple cells, thus compensating for the sparseness of single-cell transcriptome data. For the three human cell types, we observed a broad spectrum of correlation across technologies, with generally lower correlation for smaller cell types (Fig. 3c). Although the transcriptome representation was generally conserved (Fig. 3a), the snRNA-seq protocol resulted in a notable outlier when correlating the expression levels of common genes across protocols, possibly driven by decreased correlation of immature transcripts. Restricting the correlation analysis to population-specific marker genes, we observed less variation between protocols (Pearson's $r=0.5$–0.7), which underlines that the expression of these markers is largely conserved across the methods (see Supplementary Fig. 16).

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

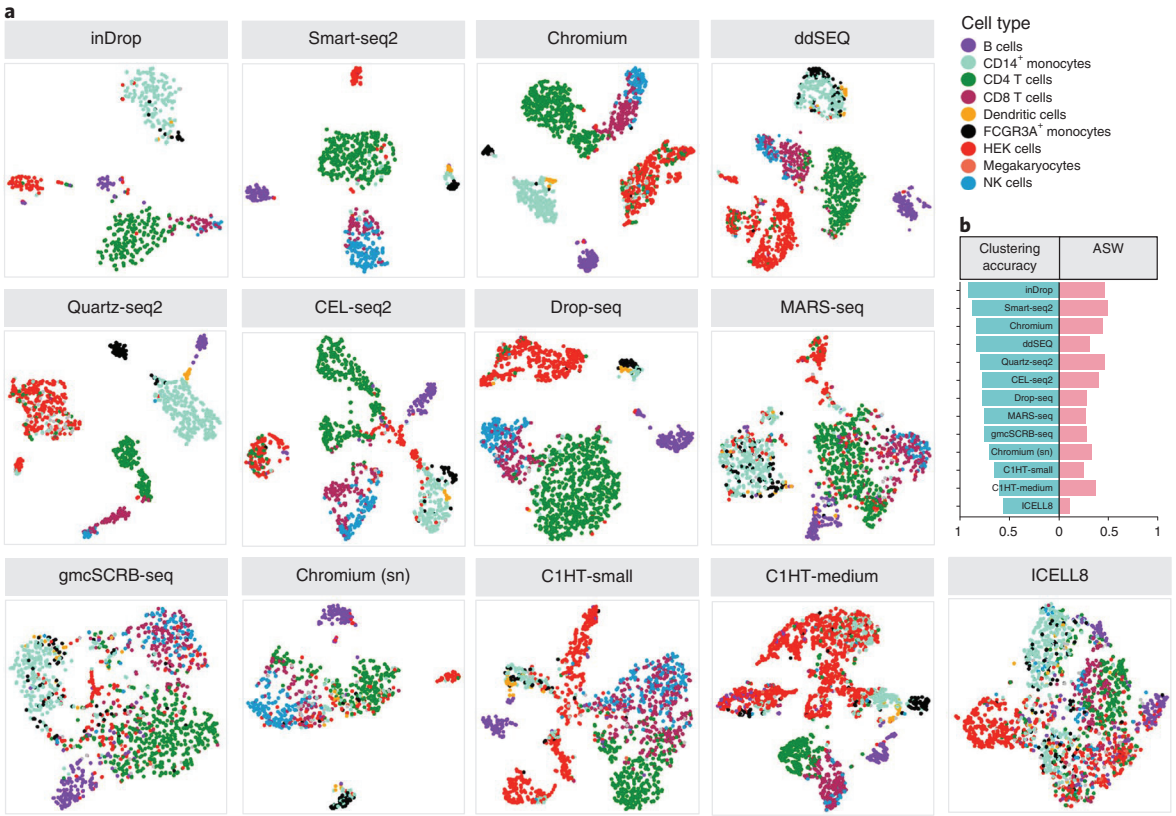


Fig. 4 | Clustering analysis of 13 sc/snRNA-seq methods on down-sampled datasets (20,000). a, The tSNE visualizations of unsupervised clustering in human samples from 13 different methods. Each dataset was analyzed separately after down-sampling to 20,000 reads per cell. Cells are colored by cell type inferred by matchSCore2 before down-sampling. Cells that did not achieve a probability score of 0.5 for any cell type were considered unclassified. **b**, Clustering accuracy and ASW for clusters in each protocol.

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

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

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

To further illustrate the power of the different protocols to chart the heterogeneity of complex samples, we clustered and plotted down-sampled datasets in two-dimensional space (Fig. 4a) and then calculated the cluster accuracy and average silhouette width (ASW[24], Fig. 4b), a commonly used measure for assessing the quality of data partitioning into communities. Consistent with the assumption that library complexity and sensitive marker detection provide greater power to describe complexity, methods that performed well for these two attributes showed better separation of subpopulations, and greater ASW and cluster accuracy. This is illustrated in the monocytes, for which accurate clustering protocols separated the major subpopulations ($CD14^+$ and $FCGR3A^+$), whereas methods with low ASW did not distinguish between them. Similarly, several methods were able to distinguish between $CD8^+$ and natural killer (NK) cells, whereas others were not.

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

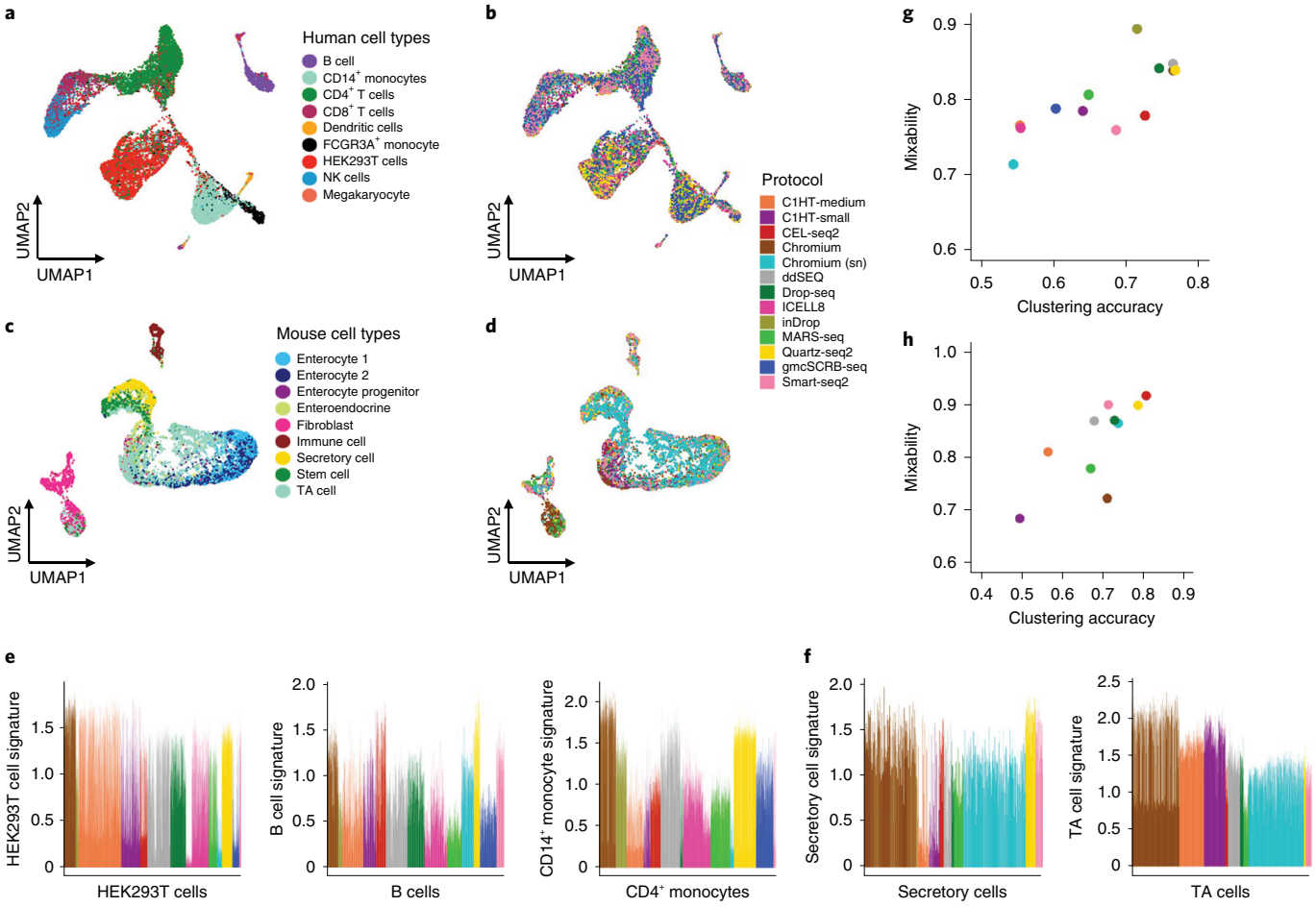


Fig. 5 | Integration of sc/snRNA-seq methods. a–d, UMAP visualization of cells after integrating technologies for 18,034 human (**a**,**b**) and 7,902 mouse (**c**,**d**) cells. Cells are colored by cell type (**a**,**c**) and sc/snRNA-seq protocol (**b**,**d**). **e**,**f**, Barplots showing normalized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

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

Cell atlas datasets will serve as a reference for annotating cell types and states in future experiments. Therefore, we assessed cells' ability to be projected on to our reference sample (Fig. 2b,c). We used the population signature model defined by matchSCore2 and evaluated the protocols based on their cell-by-cell mapping probability, which reflects the confidence of cell annotation (see Supplementary Fig. 25a–c). Although there were some differences in the projection probabilities of the protocols, and a potential bias due to the selection of the reference protocol, a confident annotation was observed for most cells with inDrop and ddSEQ reporting the highest probabilities. Notably, high probability scores were also observed in further down-sampled datasets (see Supplementary Fig. 25b). This has practical consequences, because data derived from less well-performing methods (from a cell atlas perspective), or from poorly sequenced experiments, could be identifiable and thus suitable for specific analysis types, such as tissue composition profiling.

Discussion

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

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

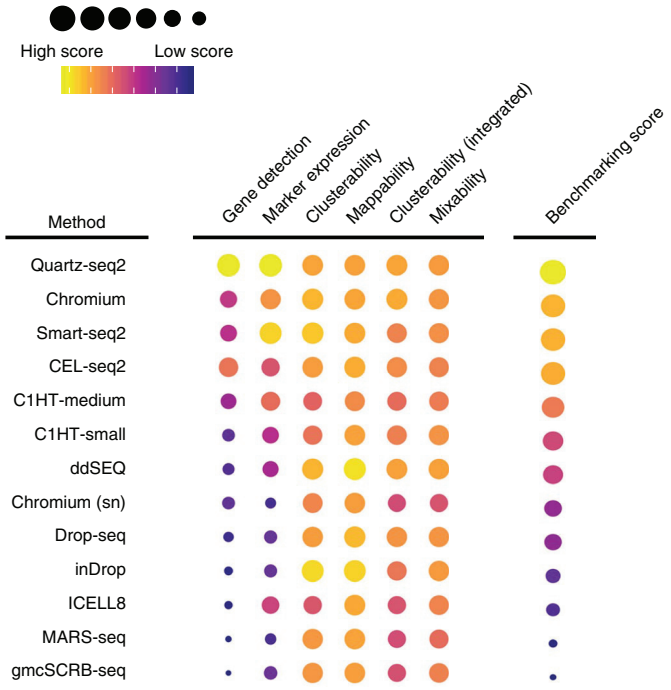


Fig. 6 | Benchmarking summary of 13 sc/snRNA-seq methods. Methods are scored by key analytical metrics, characterizing protocols according to their ability to recapitulate the original structure of complex tissues, and their suitability for cell atlas projects. The methods are ordered by their overall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

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

The use of PBMCs is ideal for multicenter benchmarking efforts; blood cells are easy to isolate and show a high recovery rate after freezing. We also included mouse colon, a solid tissue requiring dissociation before scRNA-seq. Tissue digestion and cryopreservation of colon cells present additional challenges (for example, increased rate of damaged cells), which we addressed by focusing on commonly detected cell types. Although we observed differences in the frequencies of cells from mice and humans, the composition of cell subtypes within tissues was conserved, reassuring the consistent capture of major cell types across all methods. Accordingly, subsequent analyses could be stratified by cell type, avoiding the need for a ground truth in sample composition. Furthermore, viability sorting with minimal mechanical forces (low speed and wide nozzle size) was applied to remove damaged cells and benchmark protocols with high-quality samples. This work standardized sample processing to limit technical variance in the library preparation steps, a crucial requisite for the multicenter benchmarking design. Nevertheless, on-site differences introduced during sample thawing or viability sorting could not be entirely excluded. However, our analysis also showed that viable cells selected by sorting or through thorough data quality control generate highly similar library complexity, suggesting that potential differences in sample processing have minor impacts on the data quality and supporting the robustness of our results. Processing time presents another variable related to sample and data quality. Although cells are directly sorted into their respective reaction volumes for plate-based methods, processing times can vary across microfluidic systems. However, this was considered to be an inherent feature of the library preparation workflow of the protocols that contributes to the overall performance.

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

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

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

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

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

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

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

Online content

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

Received: 7 May 2019; Accepted: 26 February 2020;
Published online: 06 April 2020

References

1. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* **13**, 2742–2757 (2018).
2. Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.* **18**, 345–361 (2017).
3. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
4. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
5. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
6. Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).
7. Karaiskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
8. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
9. Regev, A. et al. Science forum: the human cell atlas. *eLife* **6**, e27041 (2017).
10. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
11. Plass, M. et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018).
12. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl Acad. Sci. USA* **113**, 11046–11051 (2016).
13. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
14. Alioto, T. S. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).
15. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 (2017).
16. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
17. Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
18. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
19. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
20. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
21. Guillaumet-Adkins, A. et al. Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol.* **18**, 45 (2017).
22. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* **562**, 367–372 (2018).
23. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
24. Azuaje, F. A cluster validity framework for genome expression data. *Bioinforma* **18**, 319–320 (2002).
25. Lin, Y. et al. scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl Acad. Sci. USA* **116**, 9775–9784 (2019).
26. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
27. Stoeckius, M. et al. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
28. McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
29. Gaublomme, J. T. et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat. Commun.* **10**, 1–8 (2019).
30. Mora-Castilla, S. et al. Miniaturization technologies for efficient single-cell library preparation for next-generation sequencing. *J. Lab. Autom.* **21**, 557–567 (2016).
31. Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
32. Brink, S. Cvanden et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
33. Wohnhaas, C. T. et al. DMSO cryopreservation is the method of choice to preserve cells for droplet-based single-cell RNA sequencing. *Sci. Rep.* **9**, 1–14 (2019).
34. Tosti, L. et al. Single nucleus RNA sequencing maps acinar cell states in a human pancreas cell atlas. Preprint at *bioRxiv* https://doi.org/10.1101/733964 (2019).
35. Massoni-Badosa, R. et al. Sampling artifacts in single-cell genomics cohort studies. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.15.897066 (2020).
36. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
37. Sasagawa, Y. et al. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, 29 (2018).
38. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
39. Bagnoli, J. W. et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).
40. Sasagawa, Y. et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* **14**, 3097 (2013).
41. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 25533 (2016).
42. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
43. Saelens, W. et al. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
44. Holland, C. H. et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* **21**, 36 (2020).

Methods

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

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

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

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

Preparation of the reference sample. Cell counting was performed using an automated cell counter (TC20 Automated Cell Counter, Bio-Rad Laboratories). The reference sample was calculated to include human PBMCs (60%), mouse colon cells (30%), and HEK293T (6%, RFP-labeled human cell line), NIH3T3 (3%, GFP-labeled mouse cells) and MDCK (1%, TurboFP650-labeled dog cells) cells. To adjust for cell integrity loss during sample processing, we measured the viability during cell counting and accounted for an expected viability loss after cryopreservation (10% for cell lines and PBMCs; 50% for colon cells[21]). All single-cell solutions were combined in the proportions mentioned above and diluted to 250,000 viable cells per 0.5 ml. For cryopreservation, 0.5 ml of cell suspension was aliquoted into cryotubes and gently mixed with a freezing solution (final concentration 10% dimethylsulfoxide; 10% heat-inactivated FBS). Cells were then frozen by gradually decreasing the temperature (1 °C min^{-1}) to −80 °C (cryopreserved), and stored in liquid nitrogen. MARS-Seq and Smart-Seq2 experiments were performed to validate sample quality and composition before distributing aliquots to the partners.

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

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

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

MatchSCore2. To systematically assign cell identities to unannotated cells coming from different protocols, we used matchSCore2, a mathematical framework for classifying cell types based on reference data (https://github.com/elimereu/matchSCore2). The reference data consist of a matrix of gene expression counts in individual cells, the identity of which is known. The main steps of the matchSCore2 annotation are the following:

(1) Normalization of the reference data. Gene expression counts are log(normalized) for each cell using the natural logarithm of 1 + counts per 10,000. Genes are then scaled and centered using the ScaleData function in the Seurat package.
(2) Definition of signatures and their relative scores. For each of the cell types in the reference data, positive markers were computed using Wilcoxon's rank-sum test. The top 100 ranked markers in each cell type were used as the signature for that type. To each cell, we assigned a vector $\mathbf{x} = (\mathbf{x}_1, .., \mathbf{x}_n)$ of signature scores, where n is the number of cell types in the reference data. The ith signature score for the kth cell is computed as follows:

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

where J is the set of genes in signature i, and z_{jk} represents the z-score of gene j in the kth cell.
(3) Training of the probabilistic model on the reference data.

We proposed a supervised multinomial logistic regression model, which uses enrichment of the signature of each reference cell type in each cell to assign identity to that cell. In other words, for each cell k and signature i, we calculate the ith cell-type signature score $\mathbf{x}_i$ in the kth cell as described in point 2. The distribution of the signature scores is preserved, independent of which protocol is used (see Supplementary Figs. 28 and 29). More specifically, we defined the variables $\mathbf{x}_1$, …, $\mathbf{x}_n$, where $\mathbf{x}_i$ is the vector in which the scores for signature i of all cells are contained. Then we used $\mathbf{x}_i$ as the predictor of a multinomial logistic regression.

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

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

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

Model accuracy. To evaluate the fitted model using our reference datasets, we assessed the prediction accuracy in the test set, which was around 0.9 for human and 0.85 for mouse reference. We further assessed matchSCore2 classifications in datasets from other sequencing methods by looking at the agreement between clusters and classification. Notably, the resulting average agreement was 80% (range: from 58% in gmcSCRB-seq to 92% in Quartz-Seq2), whereas the rate for unclassified cells was <2%.

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

Estimation of dropout probabilities. We investigated the impact of dropout events in HEK293T cells, monocytes and B cells extracted for each technique on down-sampled data (20,000 reads per cell). For datasets with >50 cells from the selected populations, we randomly sampled 50 cells to eliminate the effect of differing cell number. The dropout probability was computed using the SCDE R package[45]. SCDE models the measurements of each cell as a mixture of a negative binomial process to account for the correlation between amplification and detection of a transcript and its abundance, and a Poisson process to account for the background signal. We then used estimated individual error models for each cell as a function of expression magnitude to compute dropout probabilities using

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

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

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

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

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

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

Dataset merging. Dataset integration across protocols is challenging and we applied different tools to assess the integratability of the sc/snRNA-seq methods, while conserving biological variability. To integrate datasets, we used Seurat[46], harmony[47] and scMerge[25], evaluated the results separately and averaged the integration capacity of the protocols into a joint score. We combined down-sampled count matrices using the sce_cbind function in scMerge, which includes the union of genes from different batches. Although both harmony and Seurat integration apply similar preprocessing steps (log(normalization), scaling and HVG identification), as implemented in the Seurat tool, scMerge uses a set of genes with stable expression levels across different cell types, and then creates pseudo-replicates across datasets, allowing the estimation and correction for undesired sources of variability. However, for all three alignment methods, Seurat was applied to perform clustering and Uniform Manifold Approximation and Projection (UMAP) after the protocol correction, to minimize the variability related to the downstream analysis. The clustering accuracy metric was used together with the mixability score to quantify the success of the integration. Omitting the cell integration step before visualizing the datasets together in a single tSNE/UMAP resulted in a protocol-specific distribution with cell types scattered to multiple clusters (see Supplementary Fig. 30).

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

Mixability. To account for the level of mixing of each technology, we used kBet[23] to quantify batch effects by measuring the rejection rate of Pearson's χ^2 test for random neighborhoods. To make a fair comparison, kBet was applied to the common cell types separately by subsampling batches to the minimum number of cells in each cell type. Due to the reduced number of cells, the option heuristic was set to 'False', and the testSize was increased to ensure a minimum number of cells.

Mixability was calculated by averaging cell-type-specific rejection rates.

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

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

Data availability

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

Code availability

All code for the analysis is provided as supplementary material. All code is also available under https://github.com/ati-lz/HCA_Benchmarking and https://github.com/elimereu/matchSCore2.

References

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

Acknowledgements

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

Author contributions

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

Competing interests

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

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

Additional information

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

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

nature biotechnology

SUPPLEMENTARY INFORMATION

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

In the format provided by the authors and unedited.

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

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

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

Supplementary Material

Supplementary Notes

Supplementary Figure legends 1-31.

Supplementary Figures 1-31.

Supplementary Tables 1-8.

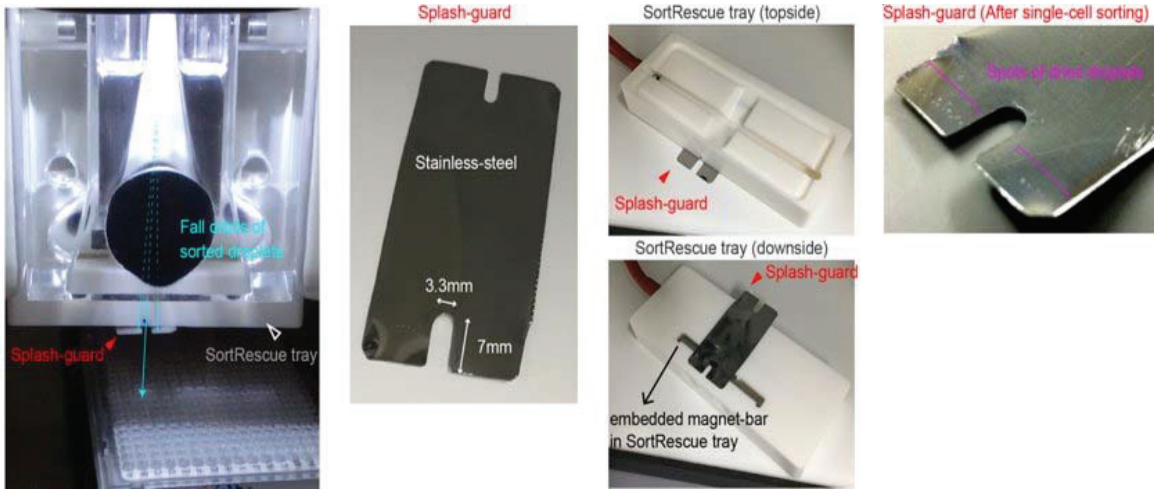
Supplementary Notes

Single-cell RNA sequencing library preparation

Quartz-Seq2[1]

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

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



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

inDrop System (1CellBio)[2]

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

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

ICELL8 SMARTer Single-Cell System (Takara Bio)[3]

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

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

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

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

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

Drop-Seq (Dolomite)[4]

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

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

Chromium V2 (10X Genomics): Single-cell RNA sequencing[5]

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

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

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

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

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

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

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

Smart-seq2[6]

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

final reagent concentrations of 1x Superscript first-strand buffer (Thermo Fisher), 5 mM DTT (Thermo Fisher), 1 M Betaine (Sigma-Aldrich), 9 mM MgCl2 (Sigma-Aldrich), 1 U/µl RNase inhibitor (Takara), 1 µM LNA template-switching oligo (5′-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3′; Exiqon), and 10 U/µl Superscript II RT enzyme. Next, pre-amplification PCR was performed for 22 cycles at final concentrations of 1x KAPA HiFi HotStart ReadyMix (Roche) and 0.08 µM ISPCR primer (5′-AAGCAGTGGTATCAACGCAGAGT-3′; IDT) in a total reaction volume of 11 µl. The cDNA was cleaned up by adding 10 µl of SPRI beads (bead stock composition: 19.5 % PEG, 1 M NaCl, 1 mM EDTA, 0.01 % IGEPAL CA-630), washing twice with 20 µl 80 % ethanol, and eluting in 10 µl H_2O. The cDNA concentration was measured for all wells using Picogreen dsDNA assay (Thermo Fisher), and diluted to 200 pg/µl using a Mantis liquid handler (Formulatrix). Next, 1 µl of cDNA was used as input for the Nextera XT library preparation kit (Illumina) at 1/5 volume, according to the manufacturer's instructions. During the 12 cycles library PCR, custom i7 and i5 indexing primers (IDT) were added at 0.5 µM each. Finally, 5 µl of library per well were pooled, cleaned and concentrated using SPRI beads (19.5 % PEG; see above). Final libraries were sequenced using HiSeq2500 V4 (Illumina). Sequence specification was as follows (Read1, 75 cycles; Index1, 8 cycles; Index2, 8 cycles; Read2, 75 cycles).

CEL-Seq2[7,8]

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

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

MARS-Seq[9]

To construct single-cell libraries from poly(A)-tailed RNA, we used massively parallel single-cell RNA sequencing (MARS-Seq). Briefly, single cells were FACS-isolated with a BD Aria III and a 100um nozzle (100 events per second) into 384-well plates containing lysis buffer (0.2% Triton X-100 (Sigma-Aldrich); RNase inhibitor (Invitrogen)) and reverse-transcription (RT) primers. Single-cell lysates were denatured and immediately placed on ice. The RT reaction mix, containing SuperScript III reverse transcriptase (Invitrogen), was added to each sample. After RT, the cDNA was pooled using an automated pipeline (epMotion, Eppendorf). Unbound primers were eliminated by incubating the cDNA with exonuclease I (NEB). A second stage of pooling was performed through cleanup with SPRI magnetic beads (Beckman Coulter). Subsequently, pooled cDNAs were converted into double-stranded DNA using the Second Strand Synthesis enzyme (NEB), followed by clean-up and linear amplification by T7 *in vitro* transcription overnight. The DNA template was then removed by Turbo DNase I (Ambion), and the RNA purified using SPRI beads. Amplified RNA was chemically fragmented using Zn2+ (Ambion), and then purified using SPRI beads. The fragmented RNA was ligated with ligation primers containing a pool barcode and partial Illumina Read1 sequencing adapter using T4 RNA ligase I (NEB). The ligated products were reverse-

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

C1 High-Throughput (HT-IFC)[10]

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

ddSEQ (Bio-Rad)

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

gmcSCRB-seq[11]

Cells were sorted and processed using the alternative lysis (Guanidine Hydrochloride) condition (gmcSCRB-seq) as described suitable for PBMCs in Bagnoli et al (2018). Briefly, single cells ("3 drops" purity mode) were sorted (Sample pressure: 5, 2-20 events per second) into 96-well DNA LoBind plates (Eppendorf) containing 5 µl lysis buffer using a Sony SH800 sorter (Sony Biotechnology #LE-SH800SZGCPL; Chip series: LE-C32, 100 µm). Lysis buffer consisted of 5 M guanidine hydrochloride (Sigma-Aldrich), 1% 2-mercaptoethanol (Sigma-Aldrich) and a 1:500 dilution of Phusion HF buffer (New England Biolabs). Samples were processed in six batches, with one batch of two plates and five batches of six plates. SPRI Beads (GE Healthcare) were prepared and diluted 50-fold (final concentration 1 mg/mL) in bead-binding buffer (22% PEG8000 (w/v), 1M NaCl, 10mM Tris-HCl pH 8.0, 1 mM EDTA, 0.01% IGEPAL, 0.05% Sodium Azide). Each well was cleaned up using a ratio of 2:1 of 1 µg/µL beads (10 µL beads and 5 µL lysate) and

resuspended in 4 µl H2O (Invitrogen) and a mix of 5 µl reverse transcription master mix, consisting of 20 units Maxima H- enzyme (Thermo Fisher), 2 × Maxima H- Buffer (Thermo Fisher), 2 mM each dNTPs (Thermo Fisher), 4 µM template-switching oligo (IDT), and 15% PEG 8000 (Sigma-Aldrich). For libraries containing ERCCs, 30,000 molecules of ERCC spike-in Mix 1 (Ambion) was used and the H2O (Invitrogen) was adjusted accordingly. After the addition of 1 µl 2 µM barcoded oligo-dT primer (E3V6NEXT, IDT), cDNA synthesis and template switching was performed for 90 min at 42 °C. Barcoded cDNA and remaining beads were then pooled in 2 ml DNA LoBind tubes (Eppendorf) and an equal volume of bead-binding buffer was added. Purified cDNA was eluted in 17 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) 1.66 × 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 19 cycles of 15 s at 98 °C, 30 s at 65 °C, 4 min at 68 °C. Final elongation was performed for 10 min at 72 °C. Batch 4 was erroneously denatured for 10 min due to a cycler error, but left in as we consider such errors as possible batch variation errors.

Following pre-amplification, all samples were purified using SPRI beads at a ratio of 1:0.8 of 1 µg/µL beads (40 µL beads and 50 µL sample) with a final elution in 10 µl of H2O (Invitrogen). The cDNA was then quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher). Size distributions were checked using high-sensitivity DNA Fragment Analyzer kits (AATI) and high-sensitivity DNA Bioanalyzer kits (Agilent). As the samples had large primer peaks, they were purified a second time using SPRI beads at a ratio of 1:0.8 and then pre-amplified for an additional 3 cycles, as above. The cDNA was then purified and reanalyzed as above. Samples passing the quantity and quality controls were used to construct Nextera XT libraries from 0.8 ng of pre-amplified cDNA. During library PCR, 3′ ends were enriched with a custom P5 primer (P5NEXTPT5, IDT). Libraries were pooled and size-selected using 2% E-Gel Agarose EX Gels (Life Technologies), cut out in the range of 300–800 bp, and extracted using the MinElute Kit

(Qiagen) according to manufacturer's recommendations. Libraries were sequenced on high output flow cells of an Illumina HiSeq 1500 instrument. Sequence specification was as follows (Read1, 16 cycles; Index1, 8 cycles; Read2, 50 cycles). 16 bases were sequenced with the first read to obtain cellular and molecular barcodes and 50 bases were sequenced in the second read into the cDNA fragment. An additional 8 base i7 barcode read was done to allow multiplexing.

Data analysis

Primary data preprocessing

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

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

For subsequent data analysis, we discarded cells with <10,000 total number of reads as well as the cells having <65% of the reads mapped to their reference genome. Cells in the 95th percentile of the number of genes/cell and those having <25% mitochondrial gene content were included in the downstream analyses. Genes that were expressed in less than five cells were removed.
For the comparative viability analysis, we used EmptyDrops[17]to determine the inflection point on the ranked barcodes vs number of detected UMIs for each library separately. We assigned all barcodes before the inflection point as cells and the remaining as empty drops.

Clustering

Filtering, normalization, selection of highly variable genes (HVG), and clustering of cells were performed using the Seurat[18] package (version 2.3.4). We normalized the gene expression measurements for each cell by the total expression, multiplied by a scale factor (10e4), and log-transformed the result. We used 10e4 (instead of 10e6 more commonly used in bulk RNA-seq) due to the reduced number of transcripts present in single-cell data. To avoid spurious correlations, the library sizes were regressed out, and the genes were scaled and centered. The scaled Z-score values were then used as normalized gene measurement input for clustering and for visualizing differences in expression between cell clusters. We selected HVGs by evaluating the relationship between gene dispersion (y.cutoff = 0.5) and the log mean expression. The clustering procedure projects cells onto a reduced dimensional space, and then groups them into subpopulations by computing a shared-nearest-neighbour (SNN) based on the Euclidean distance (finding highly interconnected communities). The algorithm is a variant of the Louvain method, which uses a resolution parameter to determine the number of clusters.
In this step, the dimension of the subspace was set to the number of significant principal components (PC) based on the distribution of the PC standard deviations and by inspecting the ElbowPlot graph. For downsampled data, the number of PCs was set to 8 after inspecting all ElbowPlot separately. The number of clusters was aligned to the expected biological variability, and

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

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

Sample deconvolution and annotation

To identify and annotate cell types and states, we analyzed the individual single-cell experiments separately, taking advantage of the original sequencing depth. Gene expression counts were log-normalized to identify HVGs, as input to compute cell-to-cell distances and graph-based clustering (see Clustering). Cell clusters were visualized in two-dimensional space using t-SNE and UMAP, and then annotated by examining previously described cell population marker genes[5,19] (**Supplementary Fig. 8** and **9**). All methods were able to recapitulate most cell types in both human and mouse samples, although in different proportions and resolutions.

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

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

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

Reference datasets

To compare the efficiency of scRNA-seq protocols in describing the structure of a mixed population, we produced a reference dataset with 30,807 human and 19,749 mouse cells. Cells were clustered and annotated as described above. Due to the high number of cells, major cell types were clustered and clearly identifiable using population marker genes (**Supplementary Fig. 4a-b**)[5,19]. However, to improve cell-to-cell annotations, we combined clustering with additional analyses. To annotate human blood cells, we used *matchSCore2* (see Methods) using an annotated set of 2700 PBMCs[5] as reference (**Supplementary Fig. 4c-d**). We used cluster-specific markers of annotated populations as input to create a multinomial logistic model according to the *matchSCore2* algorithm. For each unknown cell, we assigned probability values for any possible cell identity, and the most likely identity was used for the classification (where this probability was >0.5; otherwise the cell was considered unclassified). Cell identities inferred by *matchSCore2* were highly consistent with clusters, with agreement ranging from 96% for CD4 T-cells to 100% for B-cells. Cell-by-cell prediction helped to identify smaller cell subsets, such as FCGR3A monocytes,

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

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

Supplementary Notes References

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

Supplementary Figure legends

Supplementary Fig. 1. The effect of viability sorting on data quality (human cells).
a,b. Quality control displaying the number of detected genes and the relative proportion of reads mapped to mitochondrial transcripts (**a**; indicating cell damage) or total number of mapped reads (**b**). Cells with a mitochondrial proportion >25% and <1,778 (log10=3.25) sequencing reads were considered low-quality cells. **c.** T-SNE visualizations of unsupervised clustering in human samples with (left, 4,941 cells) or without (right, 4,094 cells) viability selection. Each dataset was analyzed separately and cells are colored by cell types inferred by *matchSCore2*. Cells that did not reach a probability score of 0.5 for any cell type were considered unclassified. **d.** Cell type composition of samples with or without viability selection with annotations from the reference dataset.

Supplementary Fig. 2. The effect of viability sorting on data quality (mouse cells).
a,b. Quality control displaying the number of detected genes and the relative proportion of reads mapped to mitochondrial transcripts (**a**; indicating cell damage) or total number of mapped reads (**b**). Cells with a mitochondrial proportion >25% and <1,778 (log10=3.25) sequencing reads were considered low-quality cells. **c.** Relationship between the number of mapped reads and detected genes for high-quality cells color-coded by cell type. **d.** T-SNE visualizations of unsupervised clustering in mouse samples with (left, 1,159 cells) or without (right, 4,245 cells) viability selection. Each dataset is analyzed separately and cells are colored by cell types inferred by *matchSCore2*. **d.** Cell type composition of samples with or without viability selection with annotations from the reference dataset.

Supplementary Fig. 3. Gene expression levels of selected marker genes.
UMAP visualization of normalized expression levels for selected marker genes of the most common PBMC (**a**, 30,807 cells) and colon (**b**, 19,749 cells) populations. Maps are shown for CD4+ T-cell markers IL7R and CD4 (expressed also in monocytes), the CD8+ T-cell marker CD8A, the B-cell marker CD79A, NK cell markers GNLY and NKG7, and monocyte-specific markers LYZ, CD14 and FCGR3A. In (**b**) maps are shown for markers of Intestinal Stem cell and proliferation (Smoc2, Miki67 and Top2a), secretory markers (Muc2, Agr2 and Tff3), enteroendocrine cell markers (Chga and Chgb), and enterocyte markers (Slc26a3, Car1 and Fabp2).

Supplementary Fig. 4. Identifying PBMC cell types using unsupervised clustering and classification.
a. UMAP visualization of 38,195 human PBMC and HEK293T human cells colored according to their assignment to clusters. Cluster labels are defined by examining the expression levels of known markers. **b.** Heatmap indicating the relative expression and gene detection rates for most common PBMC marker genes. **c.**UMAP visualization of 38,195 PBMC and HEK293T cells color coded by cell classification inferred by matchSCore2. 17% of cells were unclassified and were removed from the analysis. **d.** UMAP visualization of 38,195 PBMC and HEK293T cells showing the number of genes per cell, and scores for transcriptional signatures obtained by computing cell-type-specific markers (*lightgray*: low-score, *blue*: high score).

Supplementary Fig. 5. Identifying colon cell types by unsupervised clustering and trajectory analysis.
a. UMAP visualization of 17,558 mouse colon cells. Cells are colored by their assignment to clusters. Annotations are defined by examining the expression of known markers and differentially expressed genes (DEG). **b.** Heatmap of top DEG per cluster. Key markers of common colon cell populations are shown. **c.** Trajectory and pseudotime analysis of 8,716 immature enterocytes (IE) showing the transition from intestinal stem cells (ISC) to enterocytes. Trajectories with the relative expression of known markers are shown (yellow: low, gray: mid, blue: high). **d.** (Top) Ordered 17,558 colon cells are grouped into four different states according to their differentiation stage: intestinal stem cell (ISC), transit amplifying (TA), enterocyte progenitor (Epr), Enterocytes (E). (Bottom) UMAP visualization of IE cells colored according to the four resulting states.

Supplementary Fig. 6. Comparison of PBMC human reference with PBMC data from Zheng et al., (Nature Communications 2017).
a. UMAP visualization of 2,700 PBMCs from the Zheng et al. Chromium PBMC-3k dataset (left) and our human reference dataset (right). The colors indicate the cell types based on the annotation of the PBMC-3k dataset. Cell labels are transferred from the PBMC-3k data using the *matchSCore2* classification. **b.** Jaccard Indexes (JI) of cell type-specific markers from the two datasets (30,807 vs 2,700). For each annotated cluster,

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

Supplementary Fig. 7. Comparison of our mouse colon reference with the Tabula Muris (TM) colon dataset (Nature 2019).
a. UMAP visualization of 3,938 colon cells from the Smart-seq2 TM dataset (left) and our mouse reference dataset (right). Colors indicate the cell type based on the annotation provided by the TM Consortium. Cell labels of the mouse reference are transferred from the TM using the *matchSCore2* classification. **b**. Jaccard Indexes (JI) of cell type-specific markers from the two datasets (19,749 vs 3,938). For each annotated cluster, the top 100 ranked markers were considered. **c.** Cell type composition of our mouse reference clusters with annotations from the TM dataset.

Supplementary Fig. 8. Clustering analysis of 13 sc/snRNA-seq methods.
T-SNE visualizations of unsupervised clustering in human samples from 13 different methods. Each dataset is analyzed separately by taking advantage of its original sequencing depth. Cells are colored by cell type inferred by *matchSCore2*. Cells that did not reach a probability score of 0.5 for any cell type were considered unclassified.

Supplementary Fig. 9. Clustering analysis of 11 sc/snRNA-seq methods.
T-SNE visualizations of the unsupervised clustering in mouse samples from 11 different methods. Each dataset is analyzed separately by taking advantage of its original depth. Cells are colored according to cell type inferred by *matchSCore2*. Cells that did not reach a probability score of 0.5 for any cell types were considered unclassified.

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

Supplementary Fig. 11. Performance comparison of 13 scRNA sequencing methods.
a. Boxplots comparing the number of detected genes across protocols on downsampled data (20K), in mouse secretory and transit amplifying cells. Cell identities were defined by cell projection onto the reference**. b.** Number of genes detected at step-wise downsampled sequencing depths. Points represent the average number of genes detected for all cells of the corresponding cell type at the corresponding sequencing depth. **c,d.** Boxplots comparing the number of detected genes from countification of reads mapping to only exonic regions (c) and UMI (d, from exonic and intronic counts) across protocols on downsampled data (20K) of human HEK293T cells, monocytes and B-cells. All the boxplots display the minimum, 1st, 2nd, 3rd quantiles and maximum values.

Supplementary Fig. 12. Performance across Chromium versions and application types (sc/snRNA-seq).
a,b. Boxplots comparing the number of molecules (a) and genes (b), in downsampled (10K) HEK293T cells, monocytes and B-cells. The results are displayed for gene quantification including (open boxes) or excluding (filled boxes) intronically mapping reads. **c.** Cumulative gene counts per protocol as the average of 50 randomly sampled HEK293T cells, monocytes and B-cells on downsampled data (10K). **d.** Overlap of detected genes using cumulative gene counts from the maximum of consistently detected cells numbers (HEK293T: 46, Monocytes: 50, B-cells: 13) on downsampled (10K) data from different cells types. All the boxplots display the minimum, 1st, 2nd, 3rd quantiles and maximum values.

Supplementary Fig. 13. Technical reproducibility within sc/snRNA-seq protocols. a,b. Boxplots comparing the number of genes detected across processing units (e.g. plates, droplet lanes and IFCs), in downsampled (20K) HEK293T (a) and B-cells (b). Each protocol was stratified into processing units and only replicates with >5 cells were included. All the boxplots display the minimum, 1st, 2nd, 3rd quantiles and maximum values. **c,d.** Pearson correlation plots across replicates using the expression of all genes and cells per replicate for HEK293T (c) and B-cells (d). Protocols are ordered by Ward agglomerative hierarchical clustering. **e.** R-squared measures of the PC regression model using KBET to quantify variation in the total human dataset introduced by processing units (Online Methods).

Supplementary Fig. 14. T-SNE representation of human cell types using highly variable genes.
a,b. T-SNE representation (calculated on first 8 principle components) on downsampled data (20K) using highly variable genes across protocols, separated by HEK293T cells, monocytes and B-cells and color coded by protocols (**a**) or the number of detected genes per cell (**b**).

Supplementary Fig. 15. PCA representation of human cell types using cell type markers.
a,b. PCA analysis on downsampled data (20K) for HEK293T cells, monocytes and B-cells separately using the corresponding cell type's reference markers and color coded by protocols (**a**) or number of detected genes per cell (**b**).

Supplementary Fig. 16. Gene expression correlations across 13 sc/snRNA-seq methods.
Pearson correlation plots between protocols using gene expression of cell-type-specific signatures for HEK293T cells (**a**), monocytes (**b**) and B-cells (**c**). For a fair comparison, cells were downsampled to the same number for each method (B cells=32, Monocytes = 57, HEK293T= 55). Cells are ordered by agglomerative hierarchical clustering.

Supplementary Fig. 17. Comparison of cell type-specific markers across protocols.
a. Jaccard Indexes (JI) of B-cells, monocytes and HEK293T cell markers comparison across protocols. For each protocol, the top 100 ranked markers were considered for the JI computation. **b.** Evaluation of human marker accuracy. Protocols are compared in their ability to identify cell type-specific markers (as defined from the human reference). Jaccard Indexes are shown per cell type for each protocol (left) and their averages are displayed in relation with the clustering accuracy (right). **c.** Evaluation of mouse marker accuracy. Protocols are compared in their ability to identify cell type-specific markers (as defined from the mouse reference).

Supplementary Fig. 18. Marker overlap across protocols.
Overlap percentages of B-cells, monocytes and HEK293T markers across protocols considering the top 100 ranked markers.

Supplementary Fig. 19. Data integration using Seurat.
a,b. UMAP visualization of clusters after the integration of technologies for 18,034 human (**a**) and 7,902 mouse (**b**) cells. Cluster annotations are assigned on the basis of the most frequent cell type. **c,d.** Barplots showing normalized and method-corrected (integrated) expression scores in cell type specific signatures for CD4+ and CD8+ T-cells (**c**) and enterocytes 1, enterocytes 2 and intestinal stem cells (**d**). Bars are colored by method. **e.** Evaluation of dataset mixability after integration. Protocols are compared in their ability to mix with other technologies within same cell types. Barplots correspond to the mixability scores and colors are indicating the level of sequencing depths (10K and 20K), highlighting the drop of integratability at lower depth.

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

Supplementary Fig. 21. Integration of human sc/snRNA-seq.
a,b. UMAP visualization of 18,034 cells after harmony (**a**) and scMerge (**b**) integrations for human sc/snRNA-seq datasets (downsampled to 20K). Cells are colored by cell type (**left**) and protocol (**right**). **c,d.** Evaluation of protocol integratability in harmony (**c**) and scMerge (**d**). Protocols are compared according to their ability to group cell types into clusters (after integration) and mix with other technologies within the same clusters. Points are colored by sc/snRNA-seq protocol.

Supplementary Fig. 22. Integration of mouse sc/snRNA-seq downsampled datasets.
a,b. UMAP visualization of 7,902 cells after harmony (**a**) and scMerge (**b**) integrations for mouse sc/snRNA-seq datasets (downsampled to 20K). Cells are colored by cell type (**left**) and protocol (**right**). **c,d**. Evaluation of protocol integratability in harmony (**c**) and scMerge (**d**). Protocols are compared according to their ability to group cell types into clusters (after integration) and mix with other technologies within the same clusters. Points are colored by sc/snRNA-seq protocol.

Supplementary Fig. 23. Integration of human Chromium (V2) sc/snRNA-seq datasets.
a,b. UMAP visualization of cells after data integration with scMerge, Seurat and harmony for human Chromium scRNA-seq (1,599 cells) and snRNA-seq (856 cells) datasets (downsampled to 20K). Cells are colored by cell type (**a**) and protocol (**b**). **c.** Evaluation of protocol integratability based on the clustering accuracy after merging (separately for the three integration tools). The boxplots display the minimum, 1st, 2nd, 3rd quantiles and maximum clustering accuracies obtained for the different cell types. For all three alignment methods, Seurat was applied to perform clustering and UMAP after the protocol correction, in order to minimize the variability related to the downstream analysis. Results were consistent across tools with Chromium (single-cell) showing the highest clustering accuracy and Chromium (single-nuclei) displaying higher variability. While B-cells, monocytes and T-cells were robustly clustered, NK cells were grouped with CD8+ T-cells in Chromium scRNA-seq. CD14+ monocytes, CD8+ T-cells and HEK293T cells were poorly clustered in Chromium snRNA-seq.

Supplementary Fig. 24. Integration of mouse Chromium (V2) sc/snRNA-seq datasets.
a,b. UMAP visualization of 7,902 cells after data integration with scMerge, Seurat and harmony for mouse Chromium scRNA-seq and snRNA-seq datasets (downsampling to 20K). Cells are colored by cell type (**a**) and protocol (**b**). **c.** Evaluation of protocol integratability based on the clustering accuracy after merging for the three integration tools. Boxplots displaying the minimum, 1st, 2nd, 3rd quantiles and maximum clustering accuracies obtained for the different cell types. For all three alignment methods, Seurat was applied to perform clustering and UMAP after the protocol correction, in order to minimize the variability related to the downstream analysis. After integration, the clustering accuracy was largely conserved. Of note, transit amplifying cells were divided into two main cluster pointing to a heterogeneity between the protocols and potentially due to the decreased frequency of highly abundant ribosomal genes when sampling from the nucleus.

Supplementary Fig. 25. Comparison of mappability scores across technologies.
Boxplots displaying minimum, 1st, 2nd, 3rd quantiles and maximum probabilities values (scores) obtained by *matchSCore2* in classifying most common cell types in human (**a,b**) and mouse (**c**) samples. B-cells, HEK293T cells and CD14+ monocytes are shown with data downsampled to 20K (**a**) and 10K (**b**) sequencing reads.

Supplementary Fig. 26. Comparing column and bead purification in Quartz-seq2.
a. Sequential processing steps from poly-A tailed RNA to sequencing-ready libraries common to most sc/snRNA-seq protocols. **b.** Experimental design to systematically compare the yield of amplified cDNA using column or bead cDNA purifications, at different bead concentrations. **c.** Relative amount of amplified cDNA using different concentrations of beads. **d.** Comparing the yield of amplified cDNA using column and bead purification.

Supplementary Fig. 27. FACS sample processing strategy.
Representative FACS plot (BD Aria III) displaying sample composition and viability statistics for the HCA reference sample.

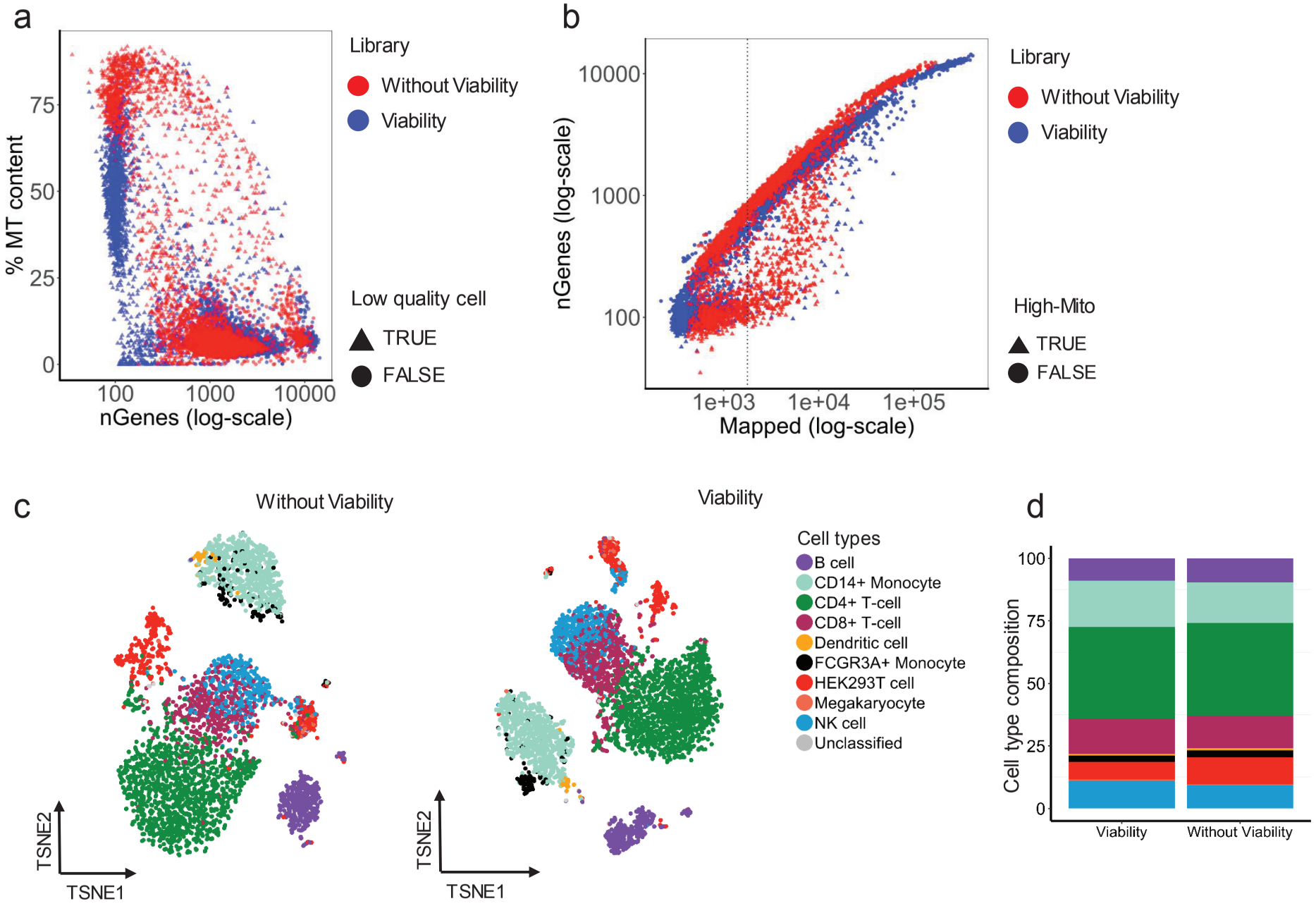
Supplementary Fig. 28. Human reference signature scores for plate-based protocols.
Boxplots comparing the distribution of B cell, monocyte and HEK293T signature scores across the different human cell types. For each cell, a score is computed by combining z-scores of genes in each signatures.

Supplementary Fig. 29. Human reference signature scores for microfluidic-based protocols.
Boxplots comparing the distribution of B cell, monocyte and HEK293T signature scores across the different human cell types. For each cell, a score is computed by combining z-scores of genes in each signatures.

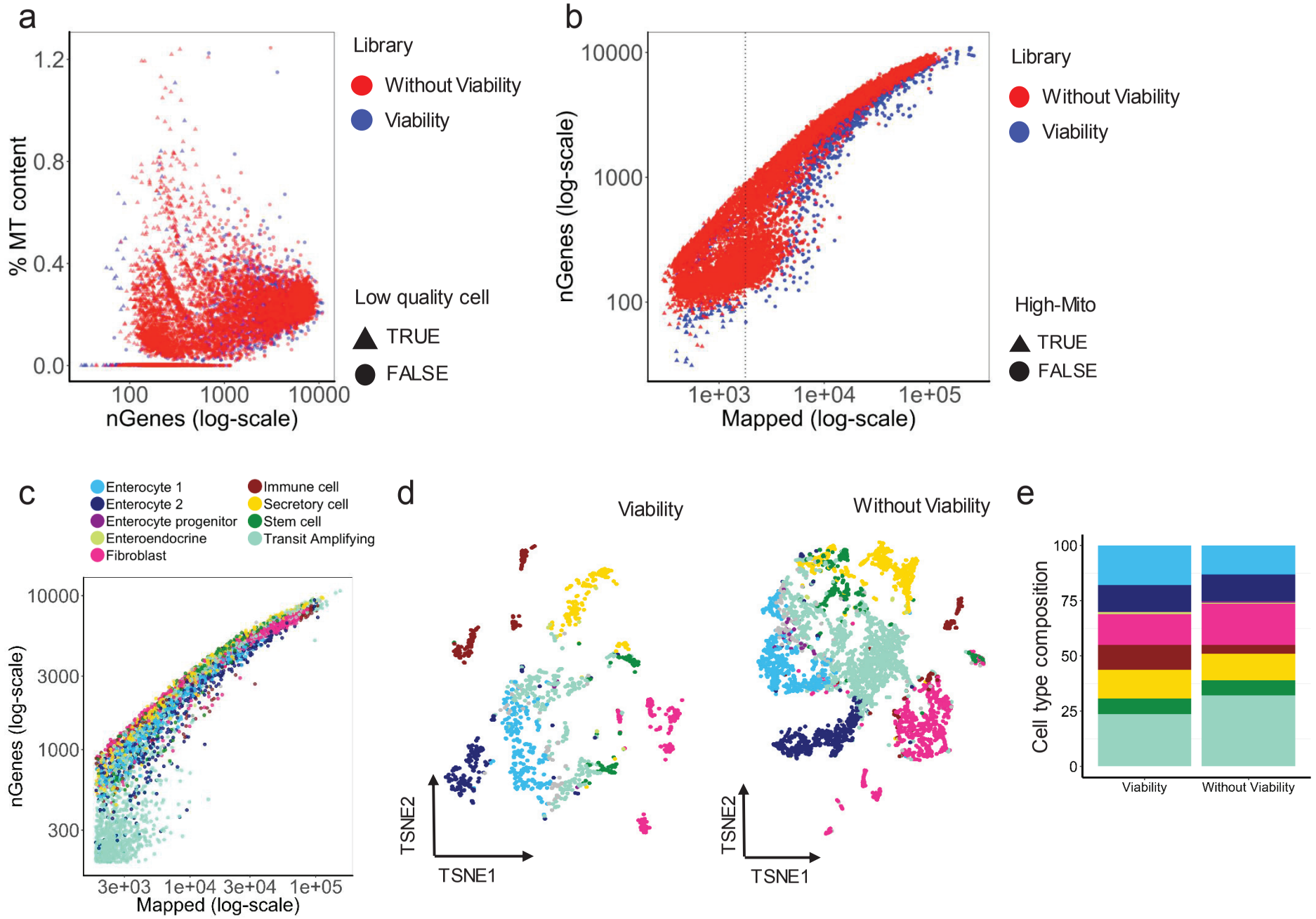
Supplementary Fig. 30. Merging of human and mouse sc/snRNA-seq datasets.
a,b. T-SNE (left) and UMAP (right) visualization of 18,034 cells after the datasets were combined and normalized by library size. Cells are colored by cell type (**a**) and protocol (**b**), showing a strong protocol-specific distribution.

Supplementary Fig. 31. Protocol performance with Chromium or inDrops as reference dataset.
a. Mappability comparison assigning the human Chromium or inDrop datasets as reference. High similarity in the ranking of mappability for B-cells, monocytes and HEK293T cells. **b.** Similar overall performance despite the reduced dataset size of the inDrop reference. **c.** Comparison of the protocol ranking to detect cell type-specific marker expression levels (using Chromium or inDrop as reference datasets). Scaled values of the averaged expression levels (data downsampled to 20K) between B-cells, monocytes and HEK293T cells are displayed.

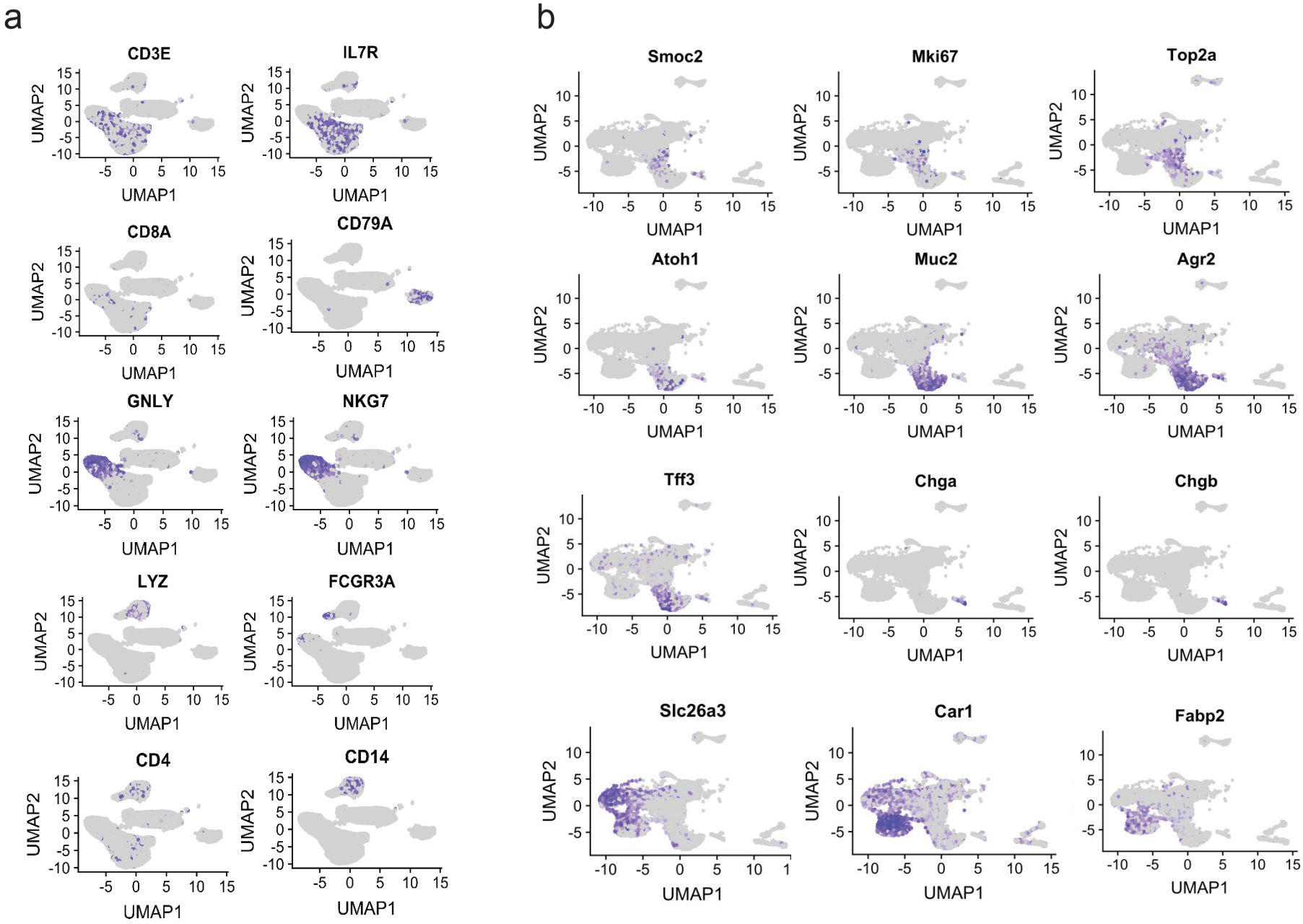
Supplementary Figure 1

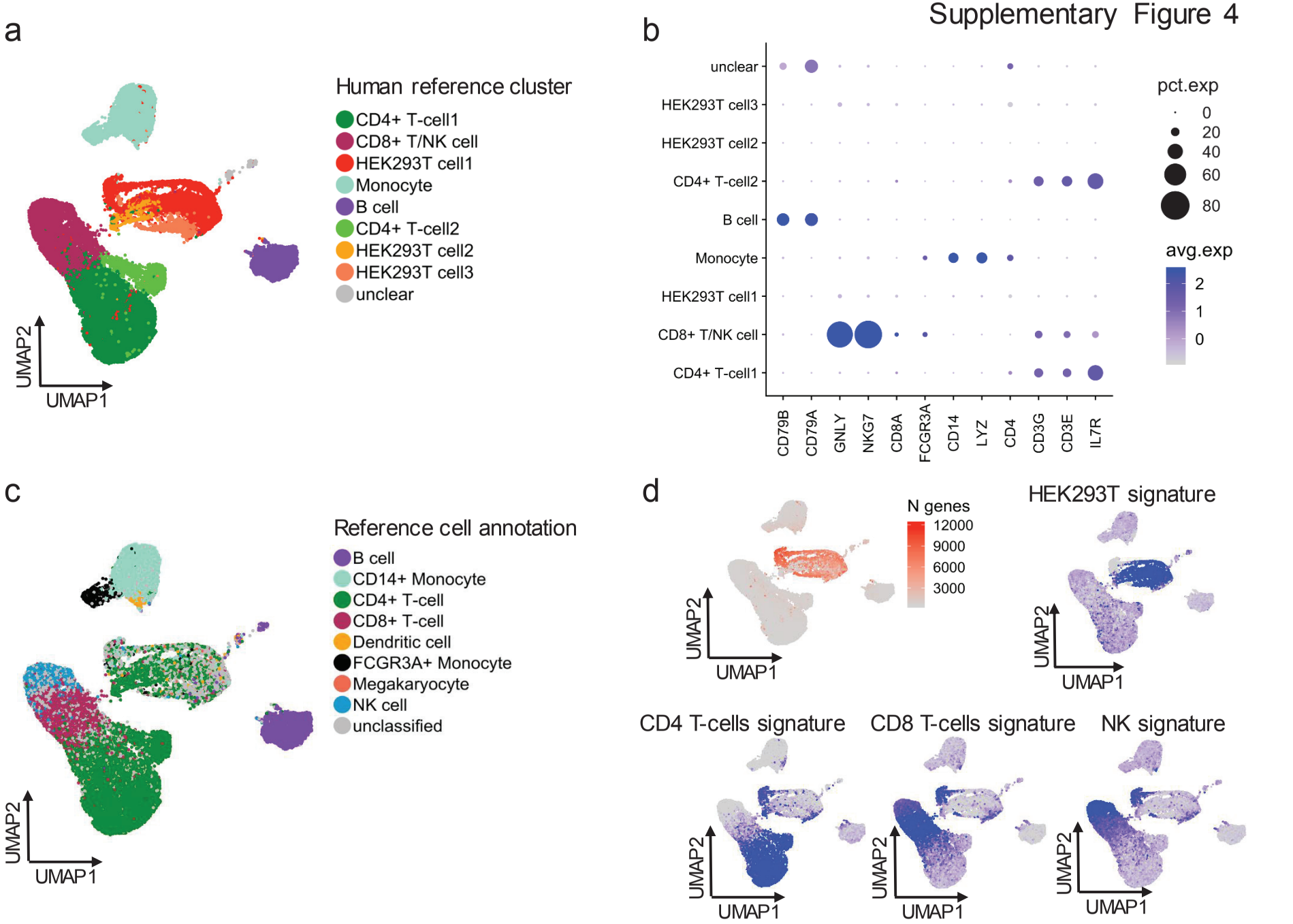


Supplementary Figure 2

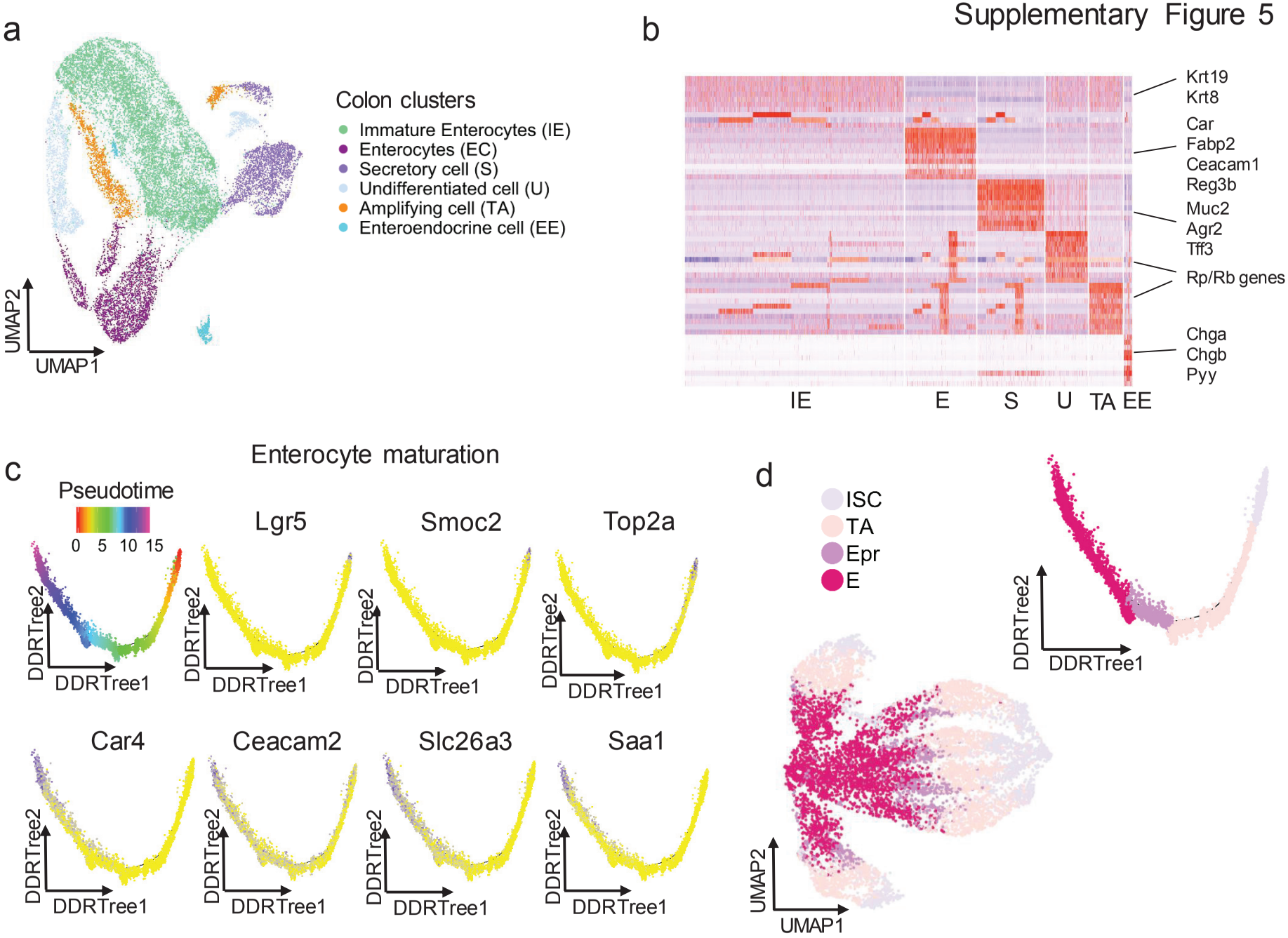


Supplementary Figure 3

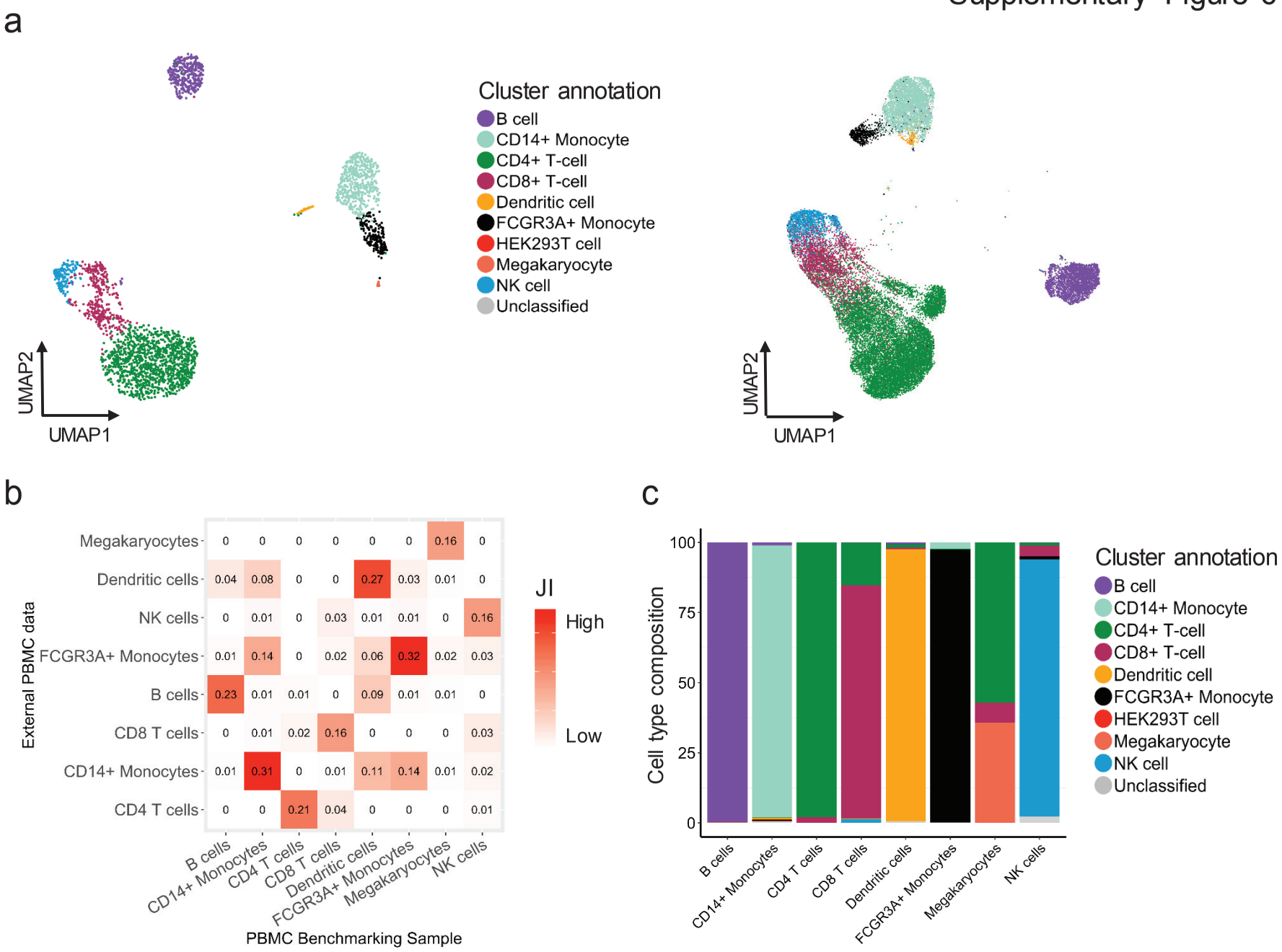


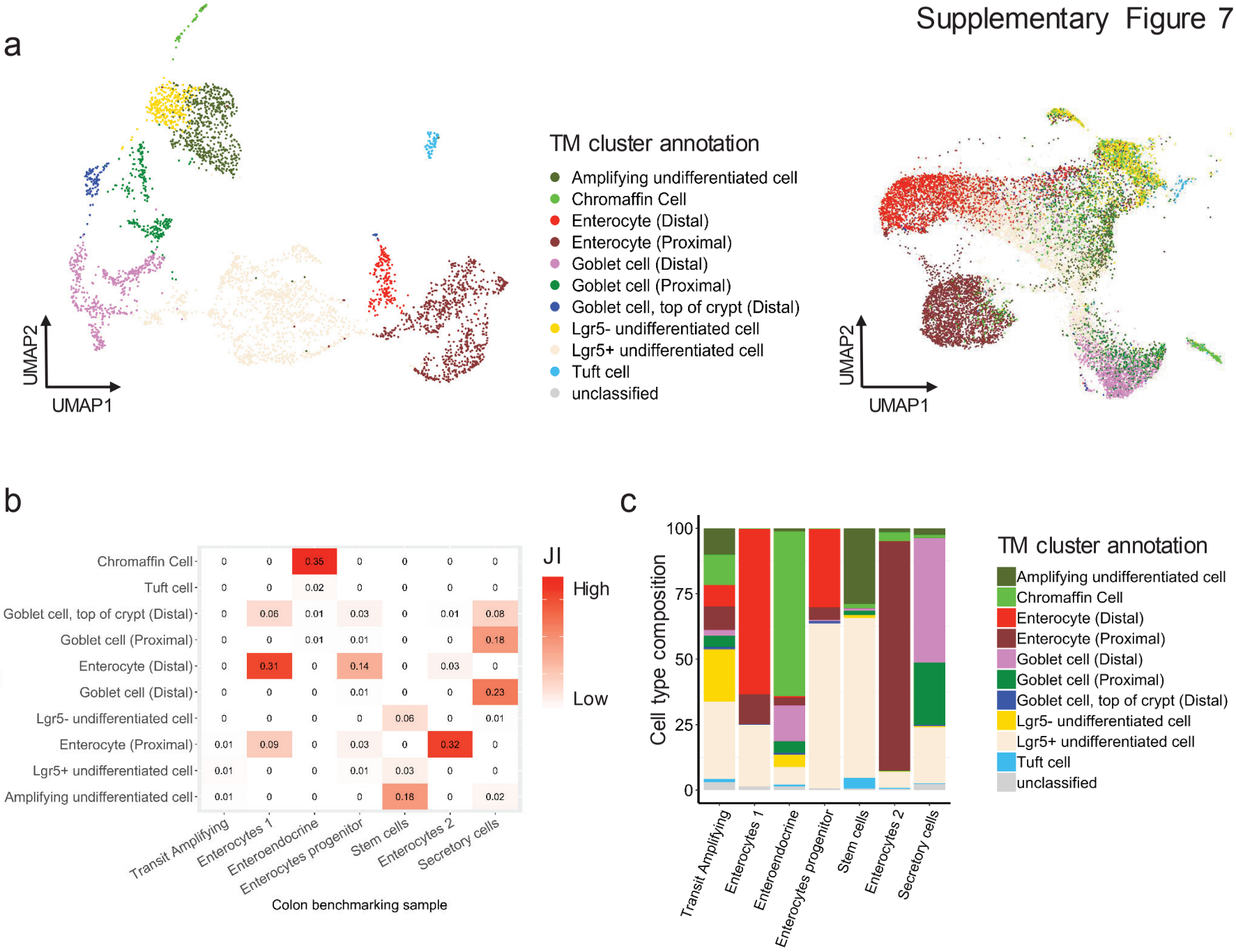

a
Human reference cluster
CD4+ T-cell1
CD8+ T/NK cell
HEK293T cell1
Monocyte
B cell
CD4+ T-cell2
HEK293T cell2
HEK293T cell3
unclear
UMAP2
UMAP1
b
unclear
HEK293T cell3
HEK293T cell2
CD4+ T-cell2
B cell
Monocyte
HEK293T cell1
CD8+ T/NK cell
CD4+ T-cell1
CD79B
CD79A
GNLY
NKG7
CD8A
FCGR3A
CD14
LYZ
CD4
CD3G
CD3E
IL7R
pct.exp
0
20
40
60
80
avg.exp
2
1
0
c
Reference cell annotation
B cell
CD14+ Monocyte
CD4+ T-cell
CD8+ T-cell
Dendritic cell
FCGR3A+ Monocyte
Megakaryocyte
NK cell
unclassified
UMAP2
UMAP1
d
N genes
12000
9000
6000
3000
HEK293T signature
CD4 T-cells signature
CD8 T-cells signature
NK signature
UMAP2
UMAP1

Supplementary Figure 5

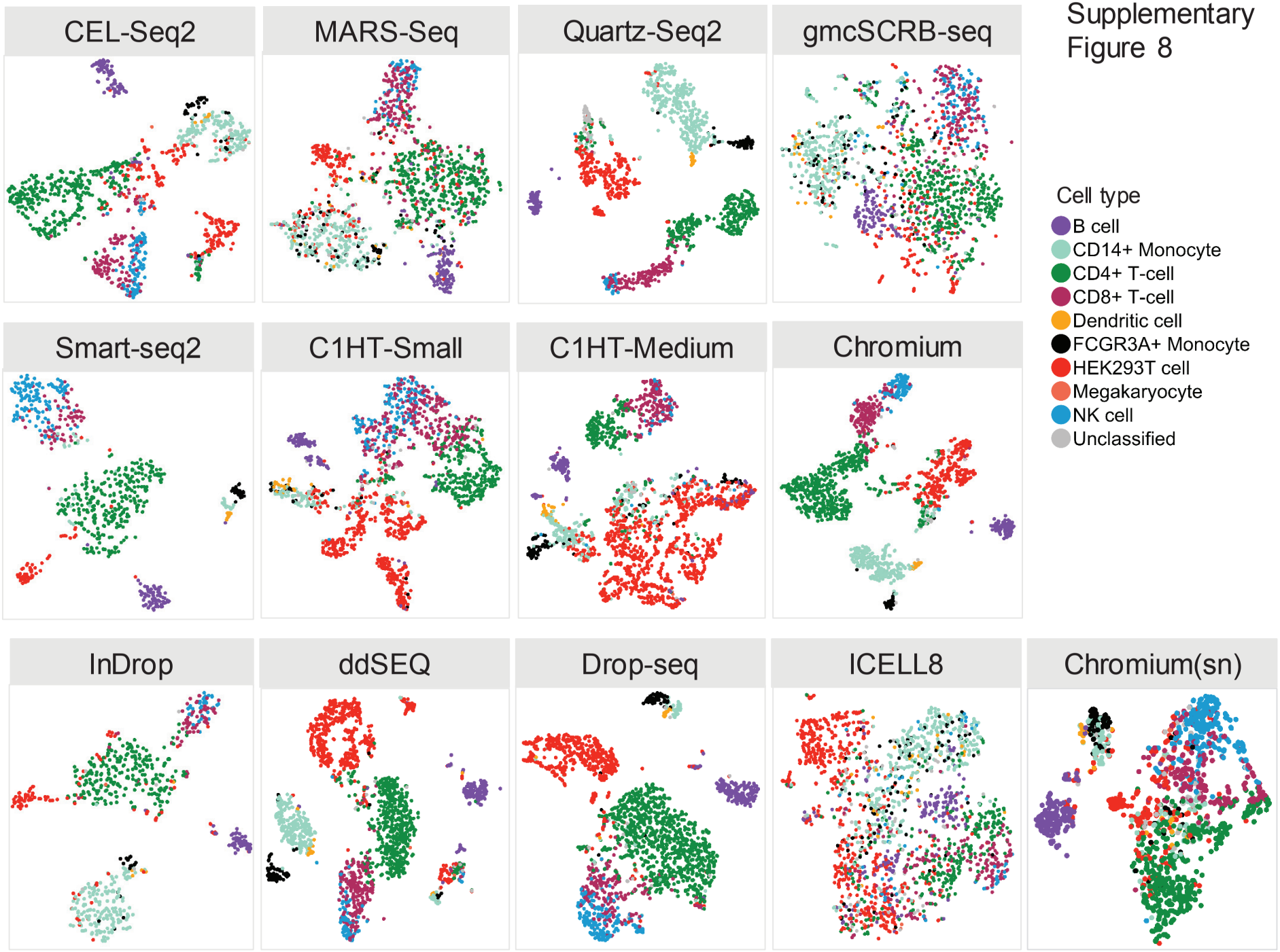


Supplementary Figure 6

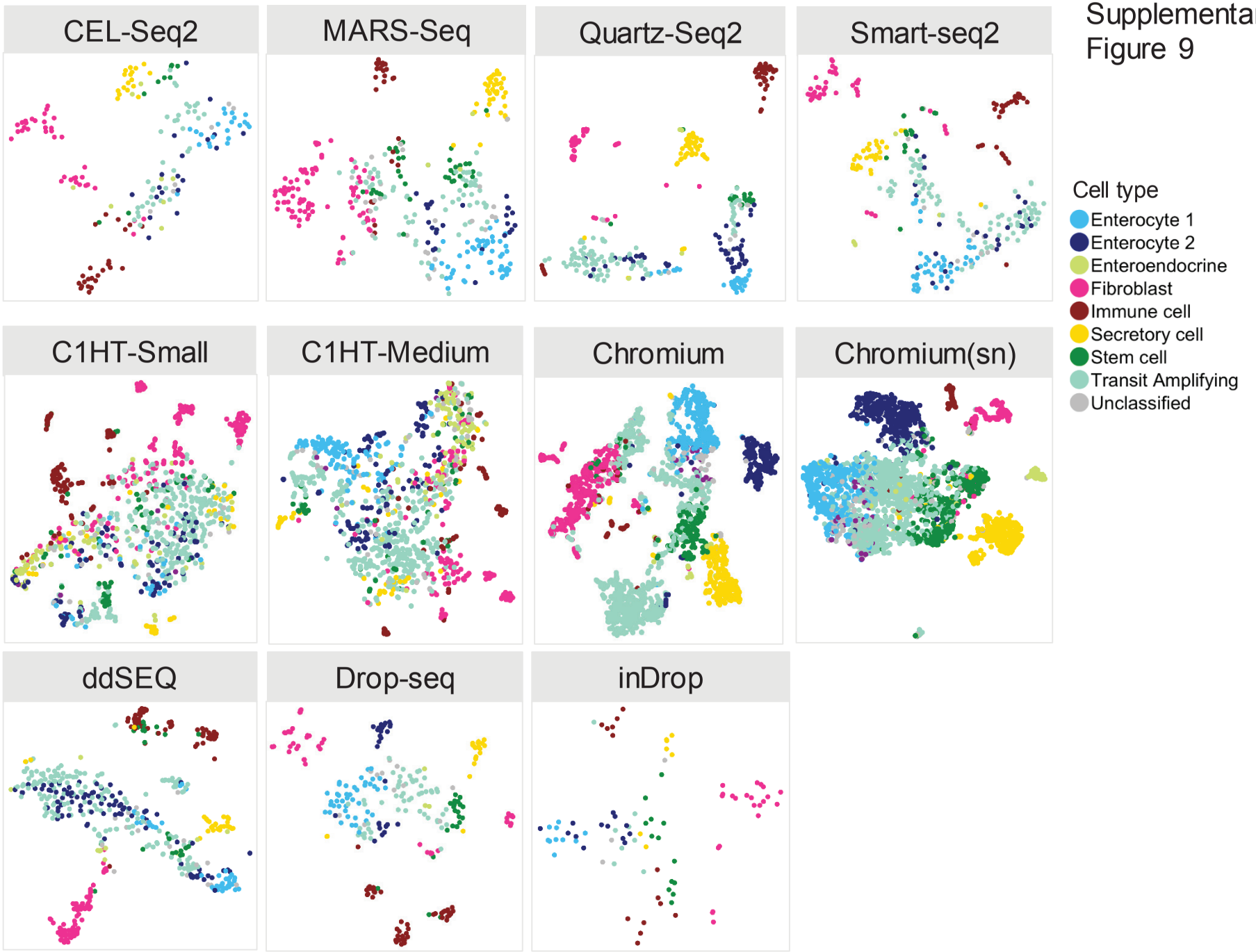



a
UMAP2
UMAP1
TM cluster annotation
Amplifying undifferentiated cell
Chromaffin Cell
Enterocyte (Distal)
Enterocyte (Proximal)
Goblet cell (Distal)
Goblet cell (Proximal)
Goblet cell, top of crypt (Distal)
Lgr5- undifferentiated cell
Lgr5+ undifferentiated cell
Tuft cell
unclassified
UMAP2
UMAP1
b
TM
Chromaffin Cell 0 0 0.35 0 0 0 0
Tuft cell 0 0 0.02 0 0 0 0
Goblet cell, top of crypt (Distal) 0 0.06 0.01 0.03 0 0.01 0.08
Goblet cell (Proximal) 0 0 0.01 0.01 0 0 0.18
Enterocyte (Distal) 0 0.31 0 0.14 0 0.03 0
Goblet cell (Distal) 0 0 0 0.01 0 0 0.23
Lgr5- undifferentiated cell 0 0 0 0 0.06 0 0.01
Enterocyte (Proximal) 0.01 0.09 0 0.03 0 0.32 0
Lgr5+ undifferentiated cell 0.01 0 0 0.01 0.03 0 0
Amplifying undifferentiated cell 0.01 0 0 0 0.18 0 0.02
Transit Amplifying
Enterocytes 1
Enteroendocrine
Enterocytes progenitor
Stem cells
Enterocytes 2
Secretory cells
Colon benchmarking sample
JI
High
Low
c
Cell type composition
100
75
50
25
0
Transit Amplifying
Enterocytes 1
Enteroendocrine
Enterocytes progenitor
Stem cells
Enterocytes 2
Secretory cells
TM cluster annotation
Amplifying undifferentiated cell
Chromaffin Cell
Enterocyte (Distal)
Enterocyte (Proximal)
Goblet cell (Distal)
Goblet cell (Proximal)
Goblet cell, top of crypt (Distal)
Lgr5- undifferentiated cell
Lgr5+ undifferentiated cell
Tuft cell
unclassified

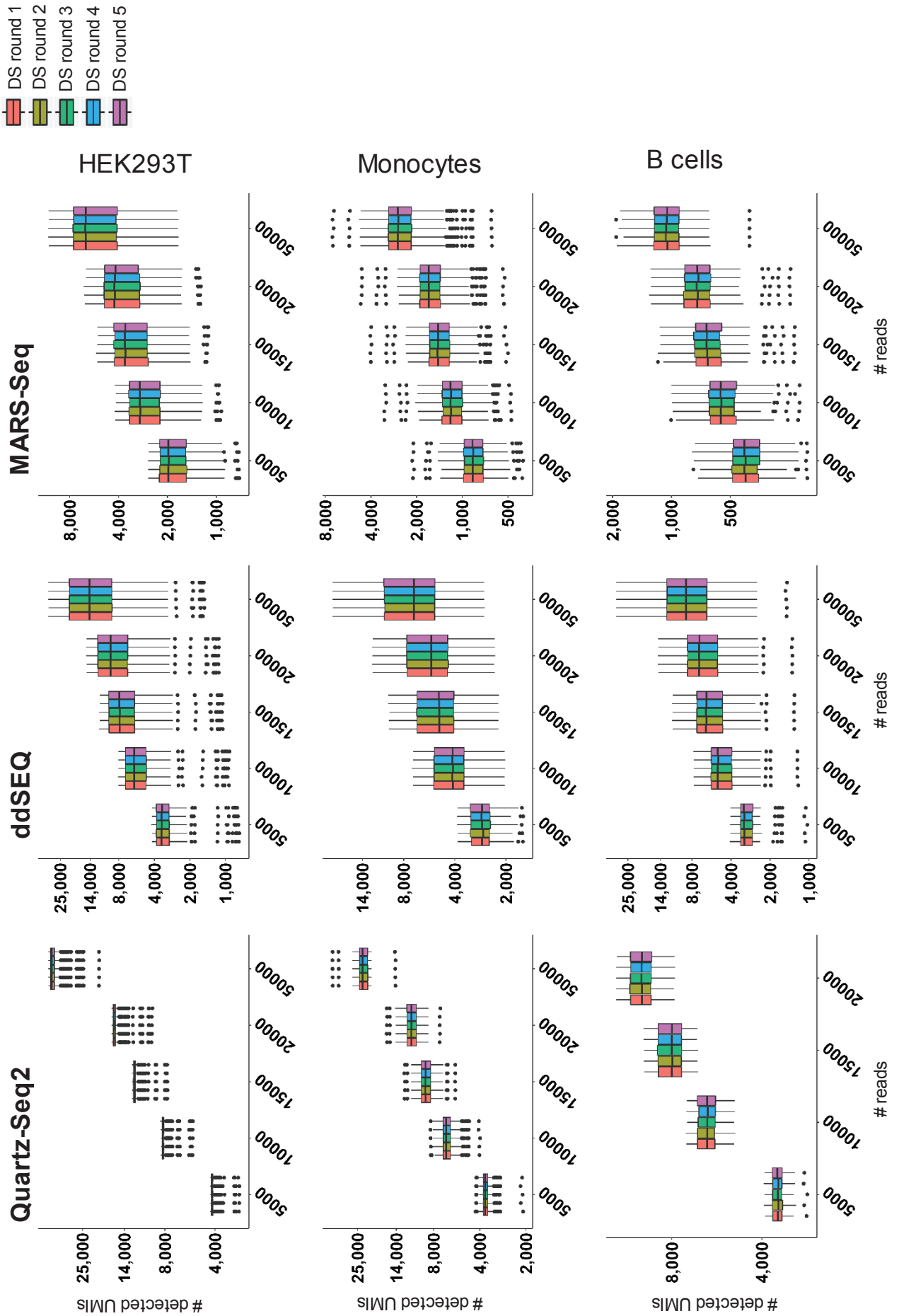
Supplementary Figure 8



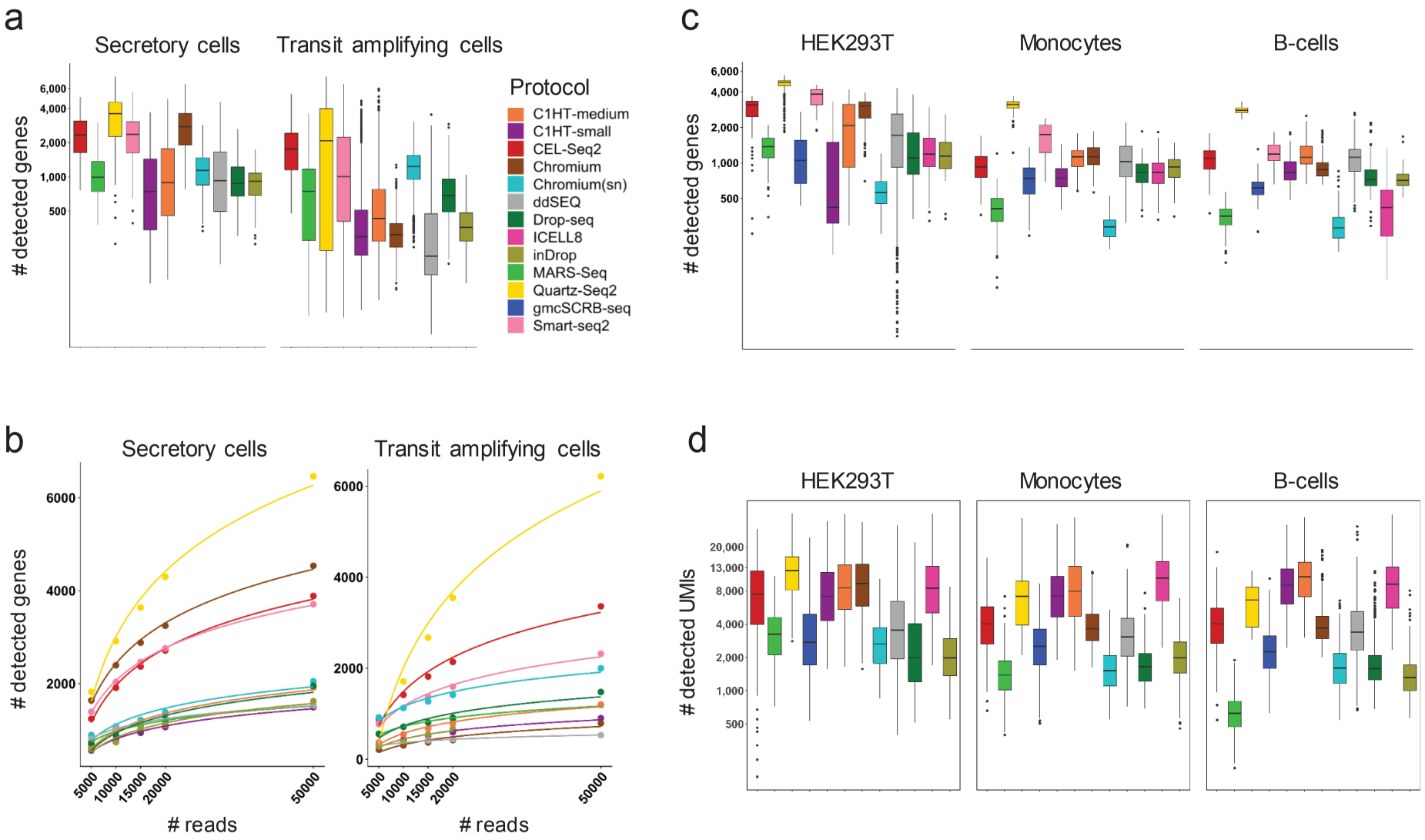
Supplementary Figure 9



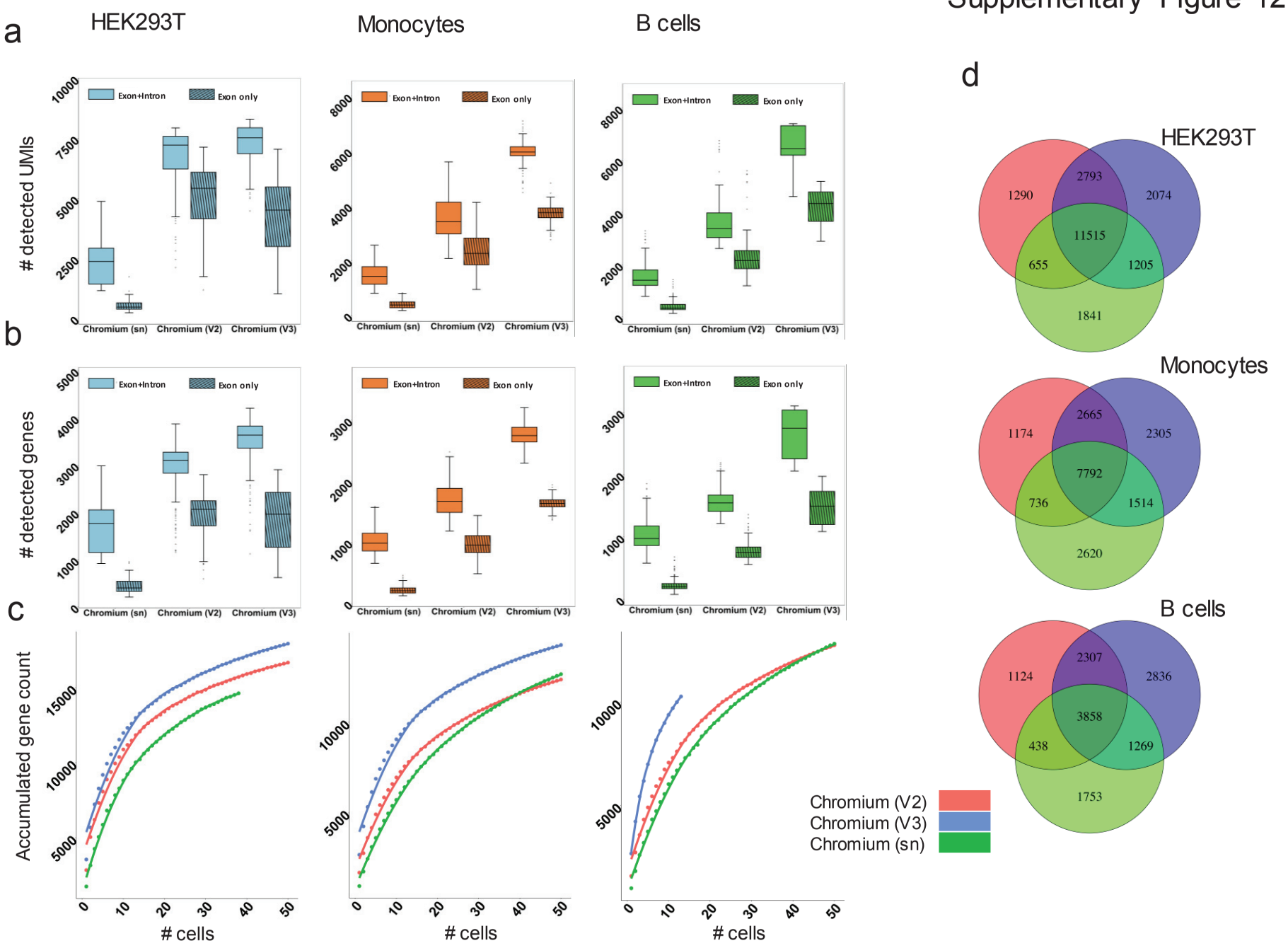
Supplementary Figure 10



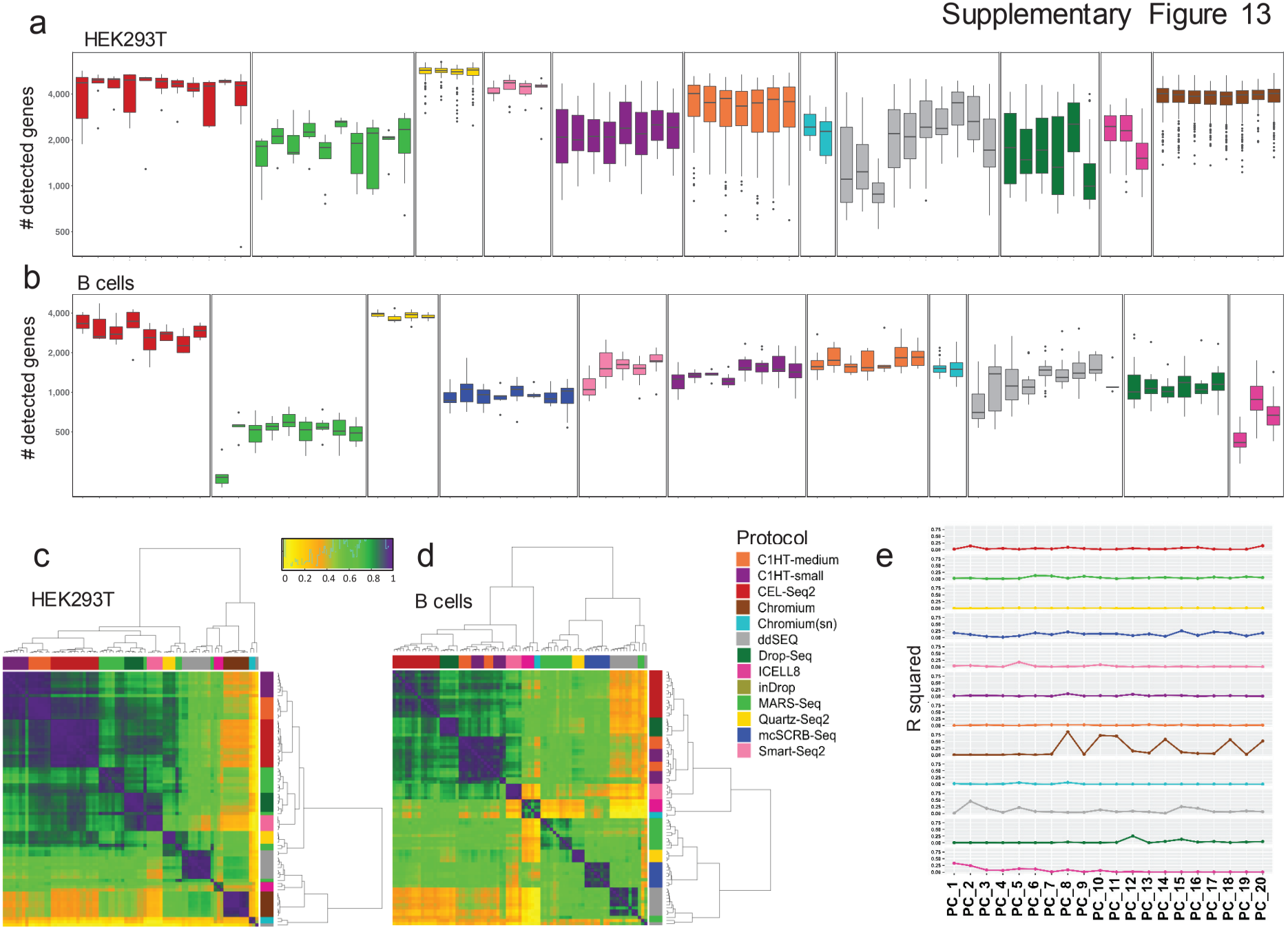
Supplementary Figure 11



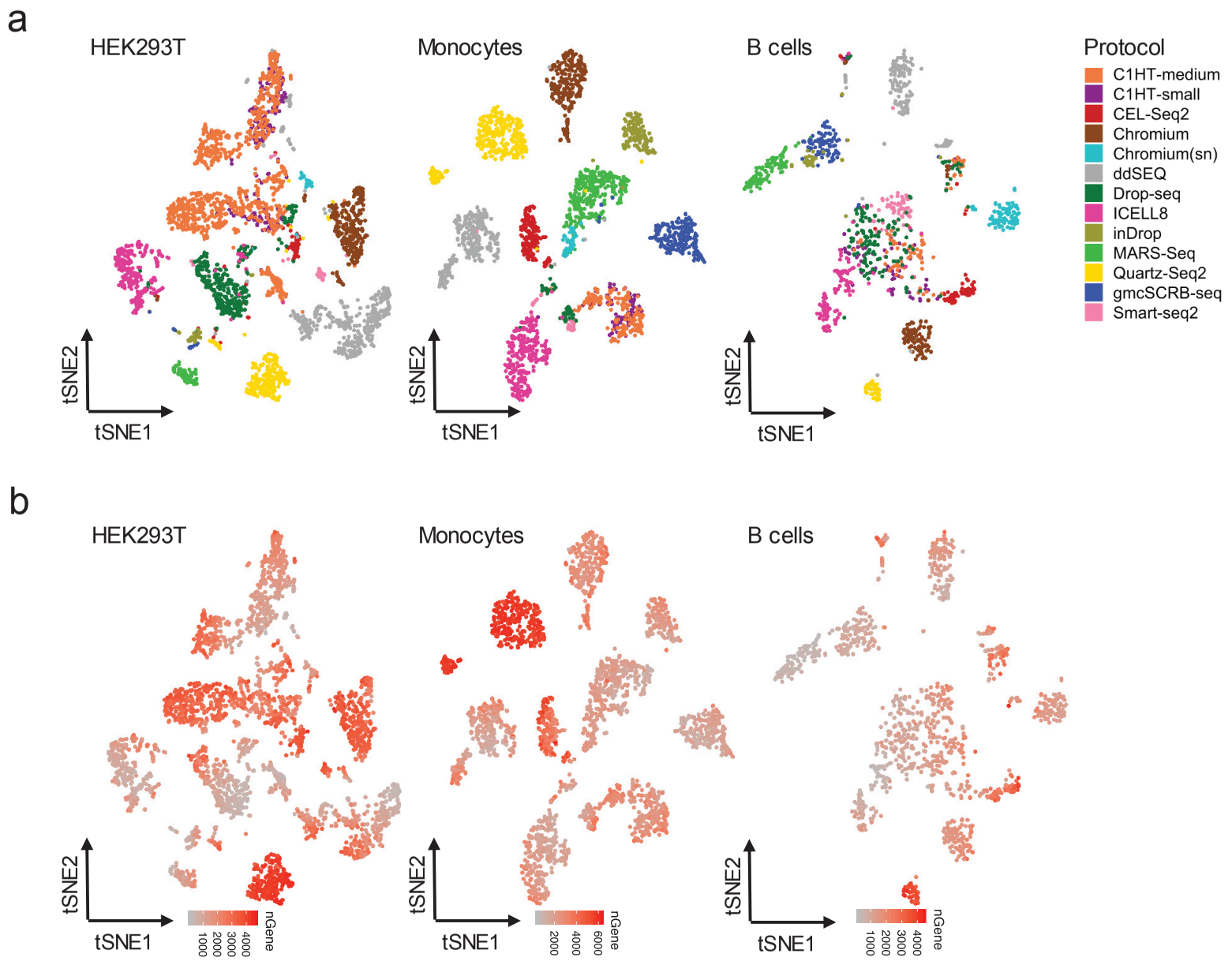
Supplementary Figure 12



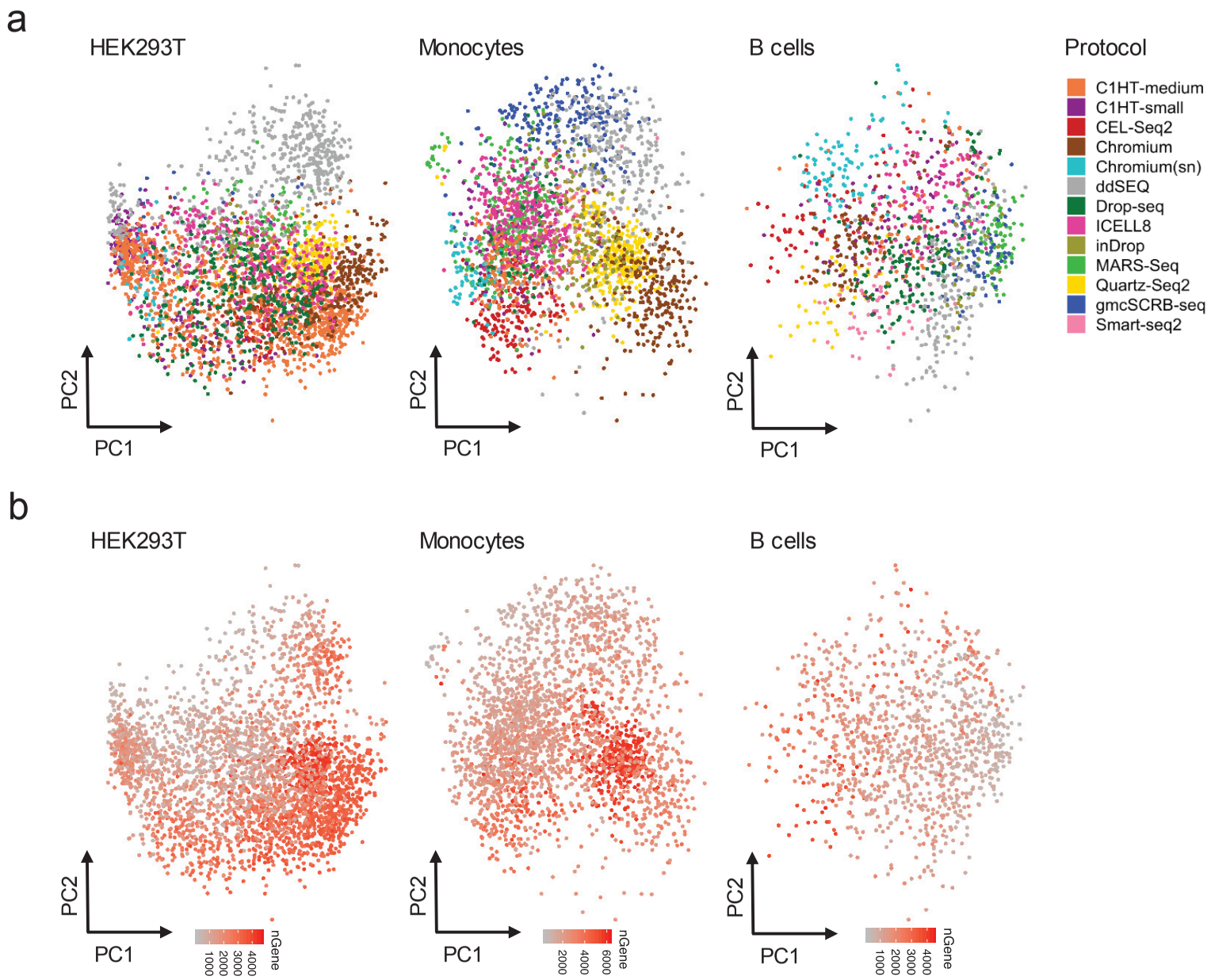
Supplementary Figure 13



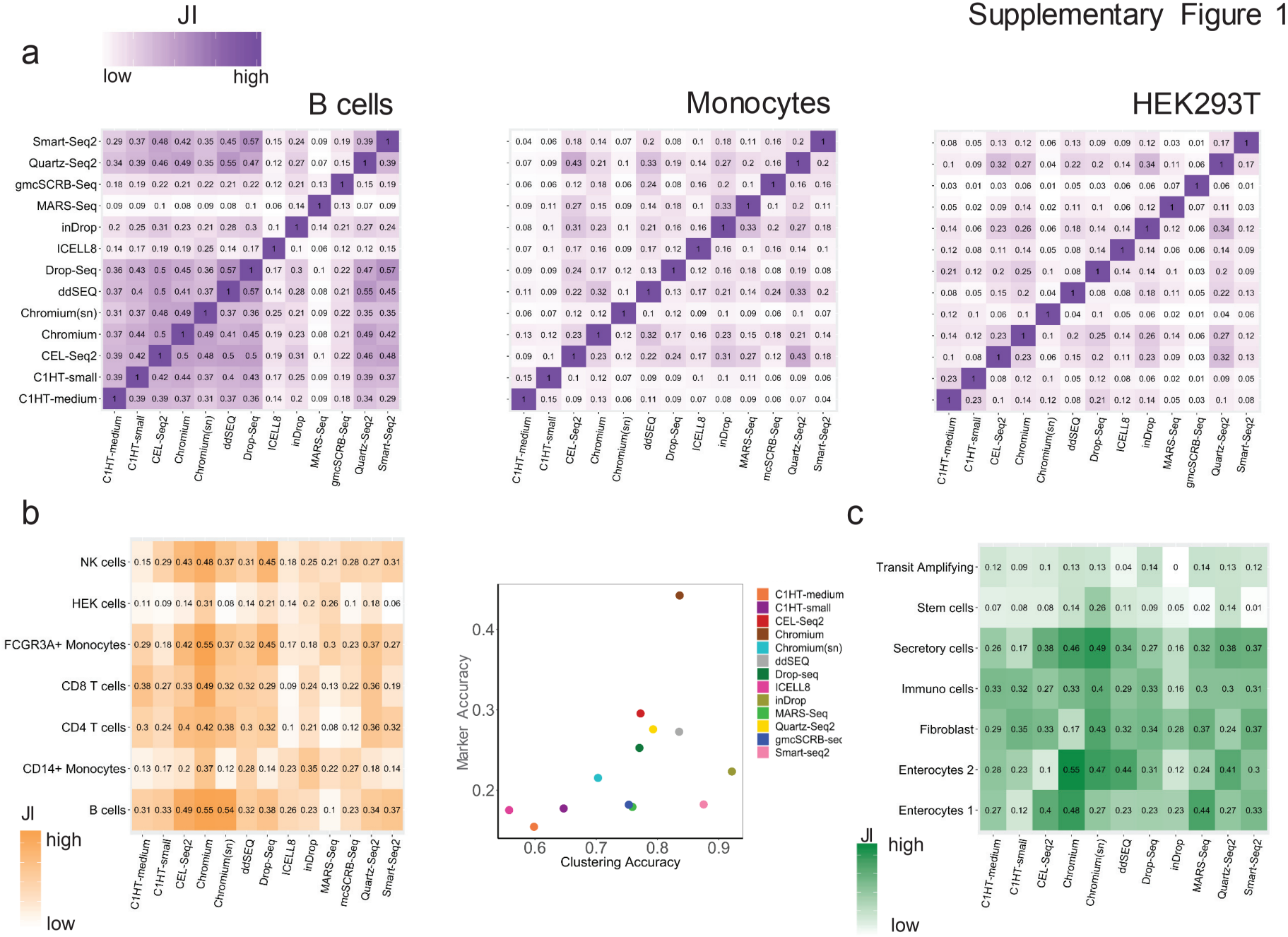
Supplementary Figure 14



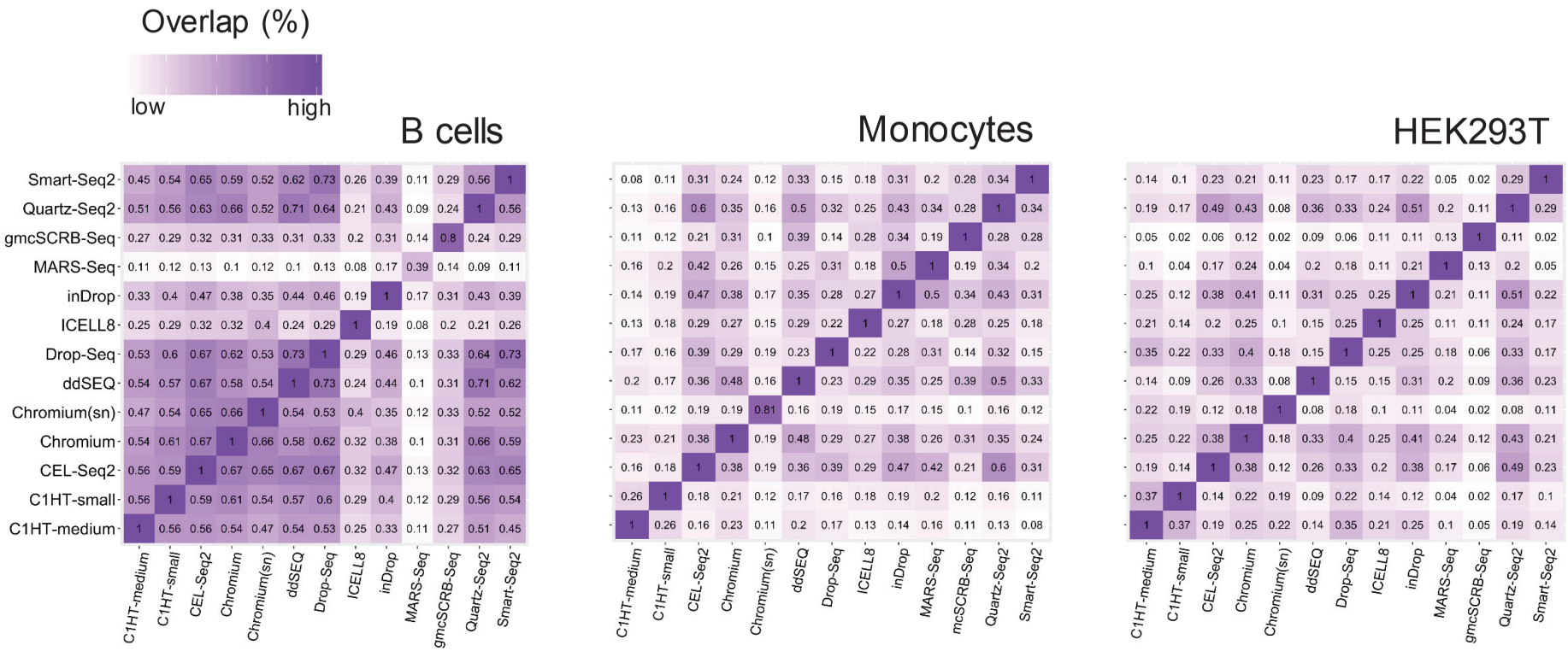
Supplementary Figure 15



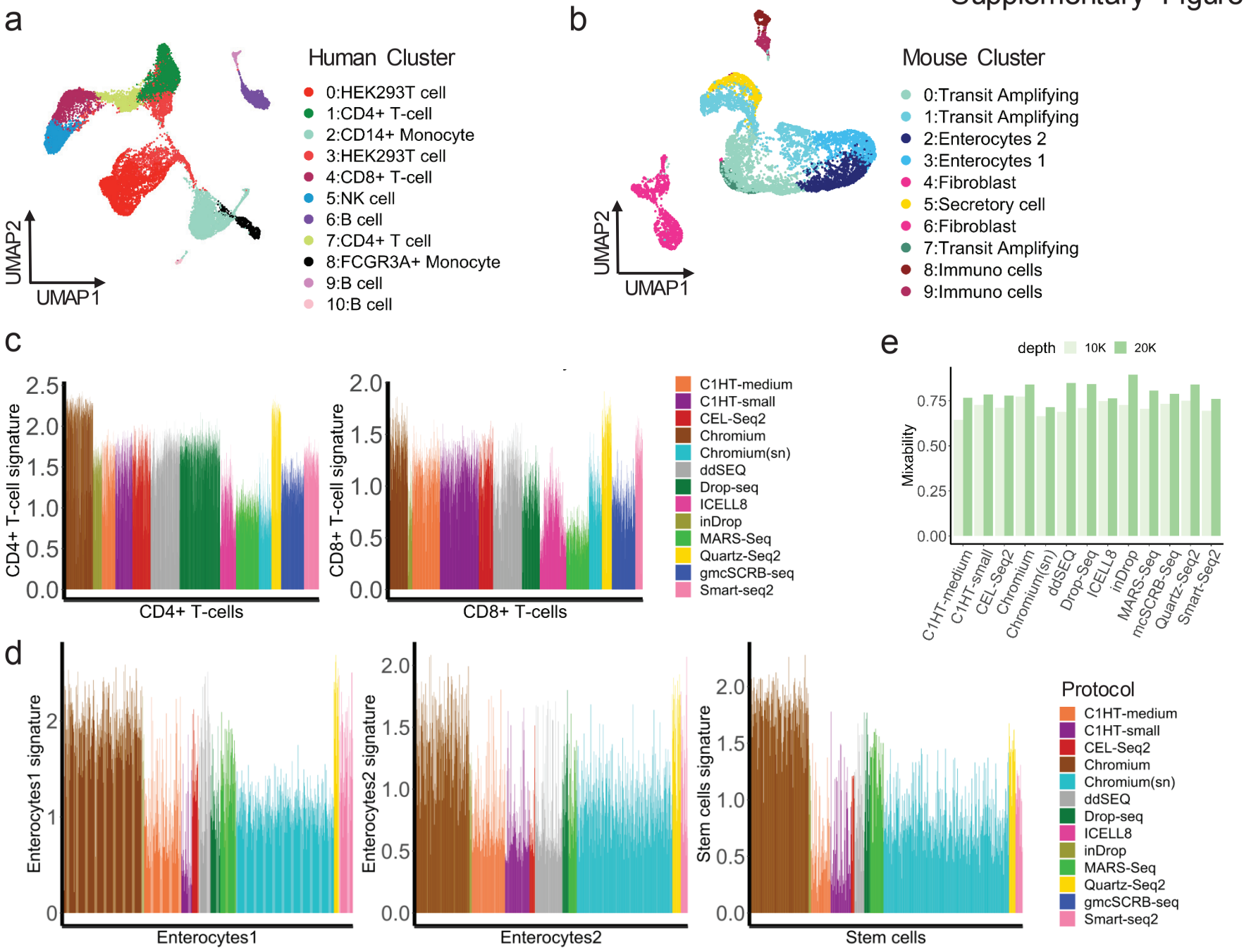
Supplementary Figure 16


a
JI
low
high
B cells
Monocytes
HEK293T
Smart-Seq2
Quartz-Seq2
gmcSCRB-Seq
MARS-Seq
inDrop
ICELL8
Drop-Seq
ddSEQ
Chromium(sn)
Chromium
CEL-Seq2
C1HT-small
C1HT-medium
b
NK cells
HEK cells
FCGR3A+ Monocytes
CD8 T cells
CD4 T cells
CD14+ Monocytes
B cells
JI
high
low
Marker Accuracy
Clustering Accuracy
C1HT-medium
C1HT-small
CEL-Seq2
Chromium
Chromium(sn)
ddSEQ
Drop-seq
ICELL8
inDrop
MARS-Seq
Quartz-Seq2
gmcSCRB-sec
Smart-seq2
c
Transit Amplifying
Stem cells
Secretory cells
Immuno cells
Fibroblast
Enterocytes 2
Enterocytes 1
JI
high
low

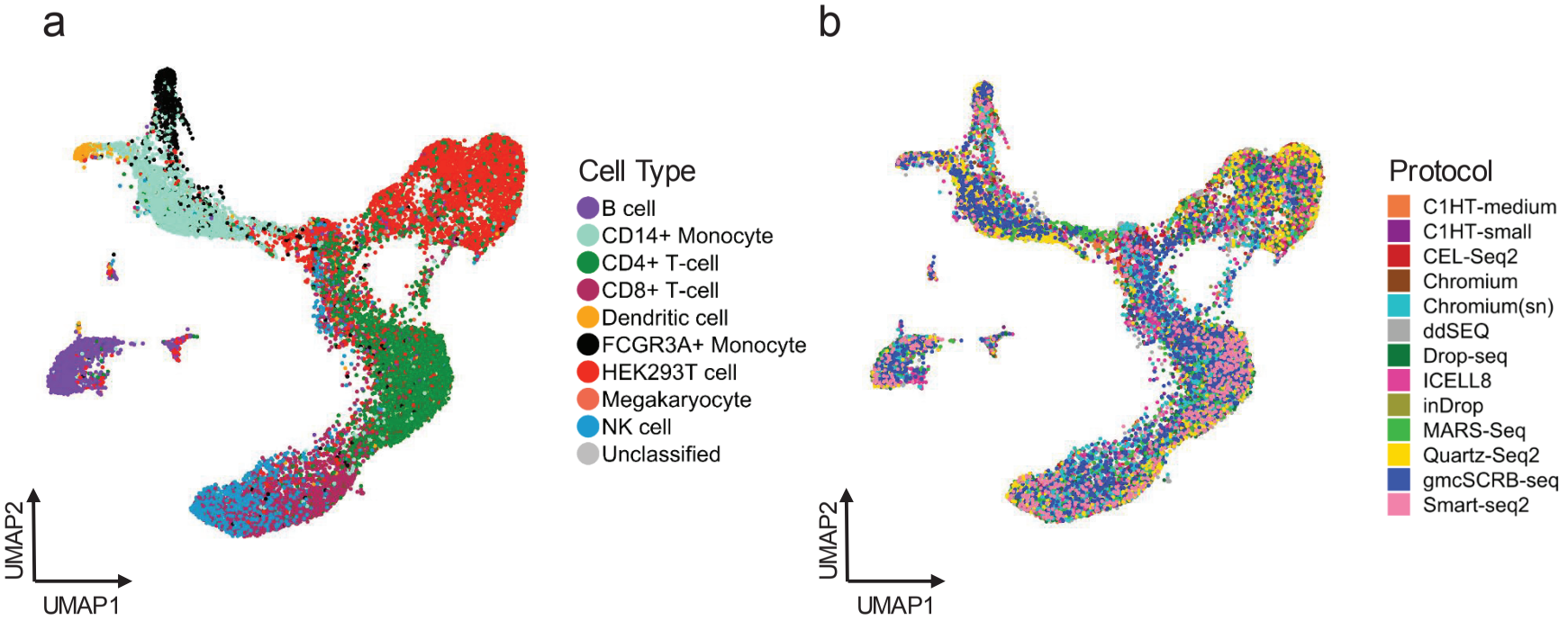
Supplementary Figure 18



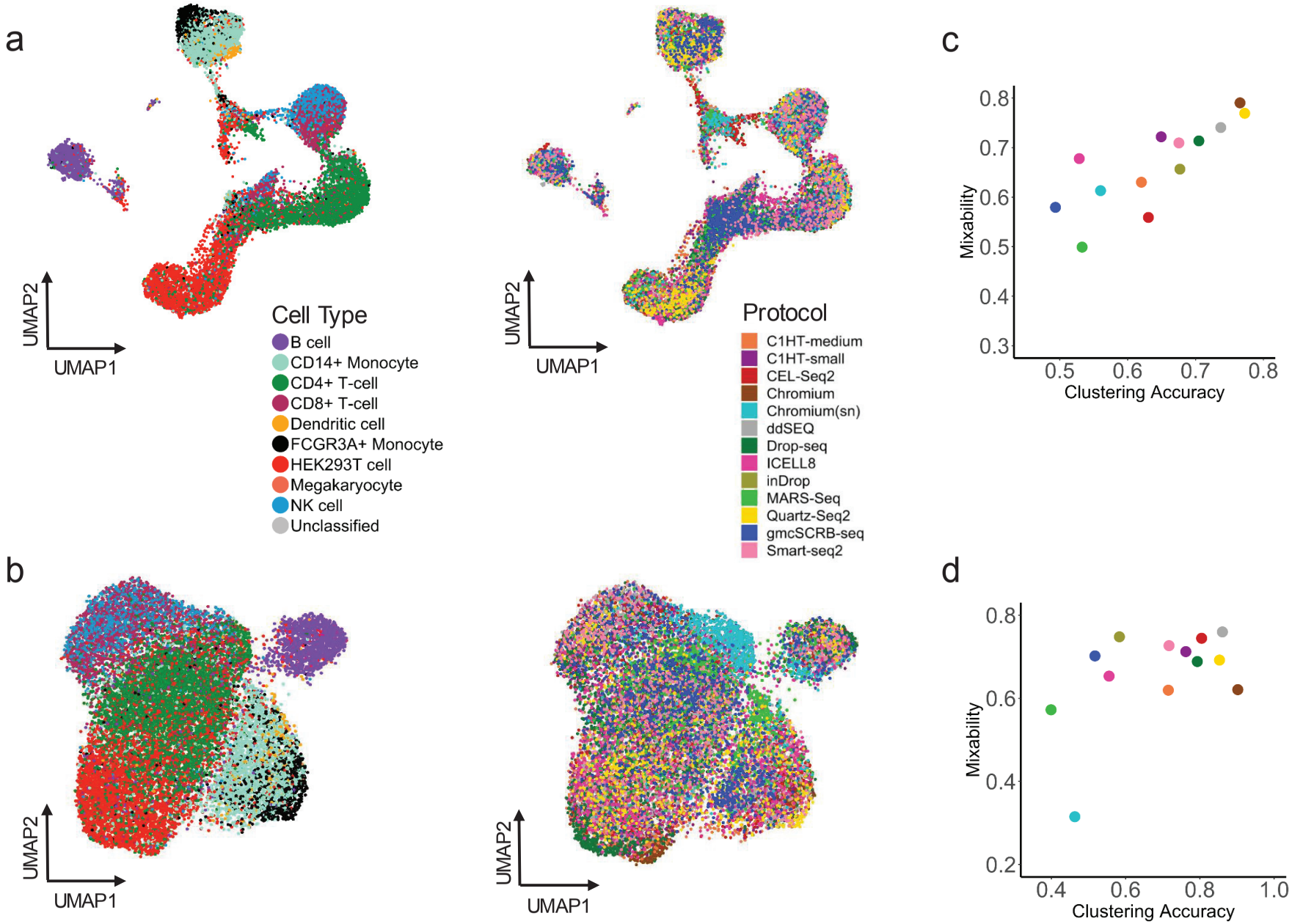
Supplementary Figure 19



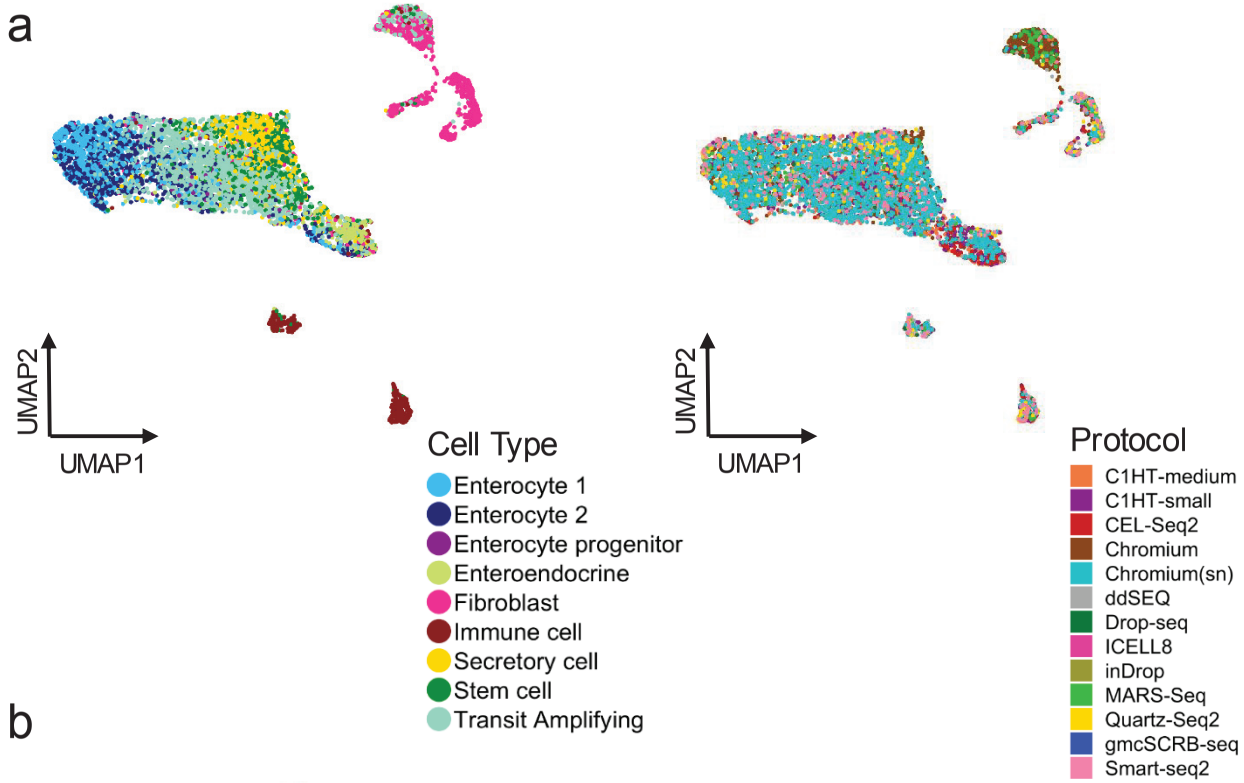
Supplementary Figure 20

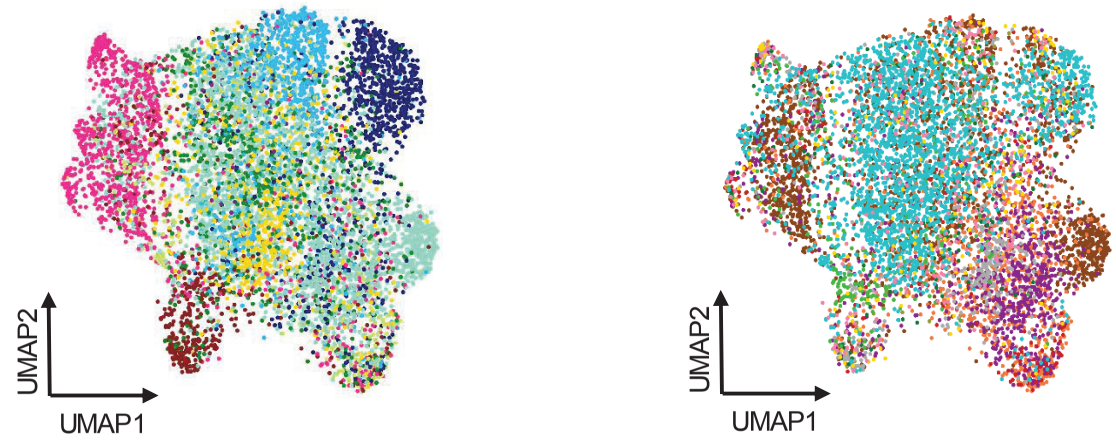


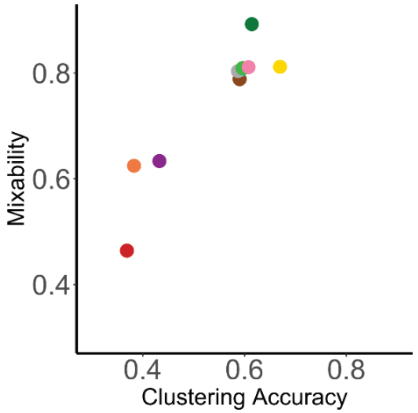
Supplementary Figure 21

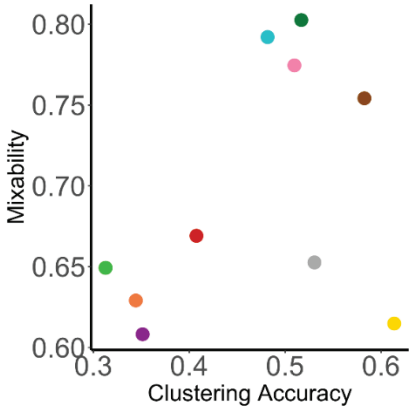


Supplementary Figure 22

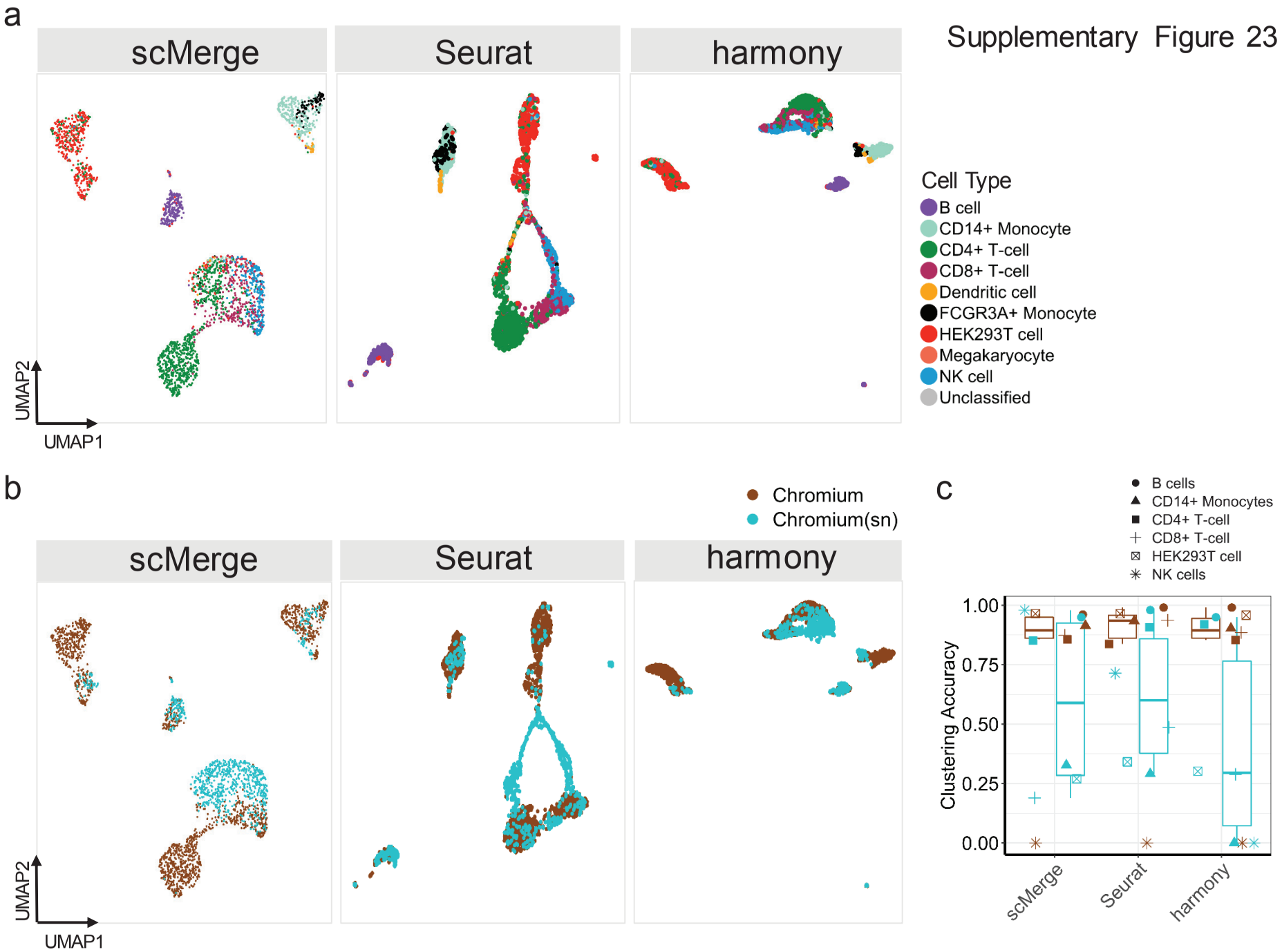




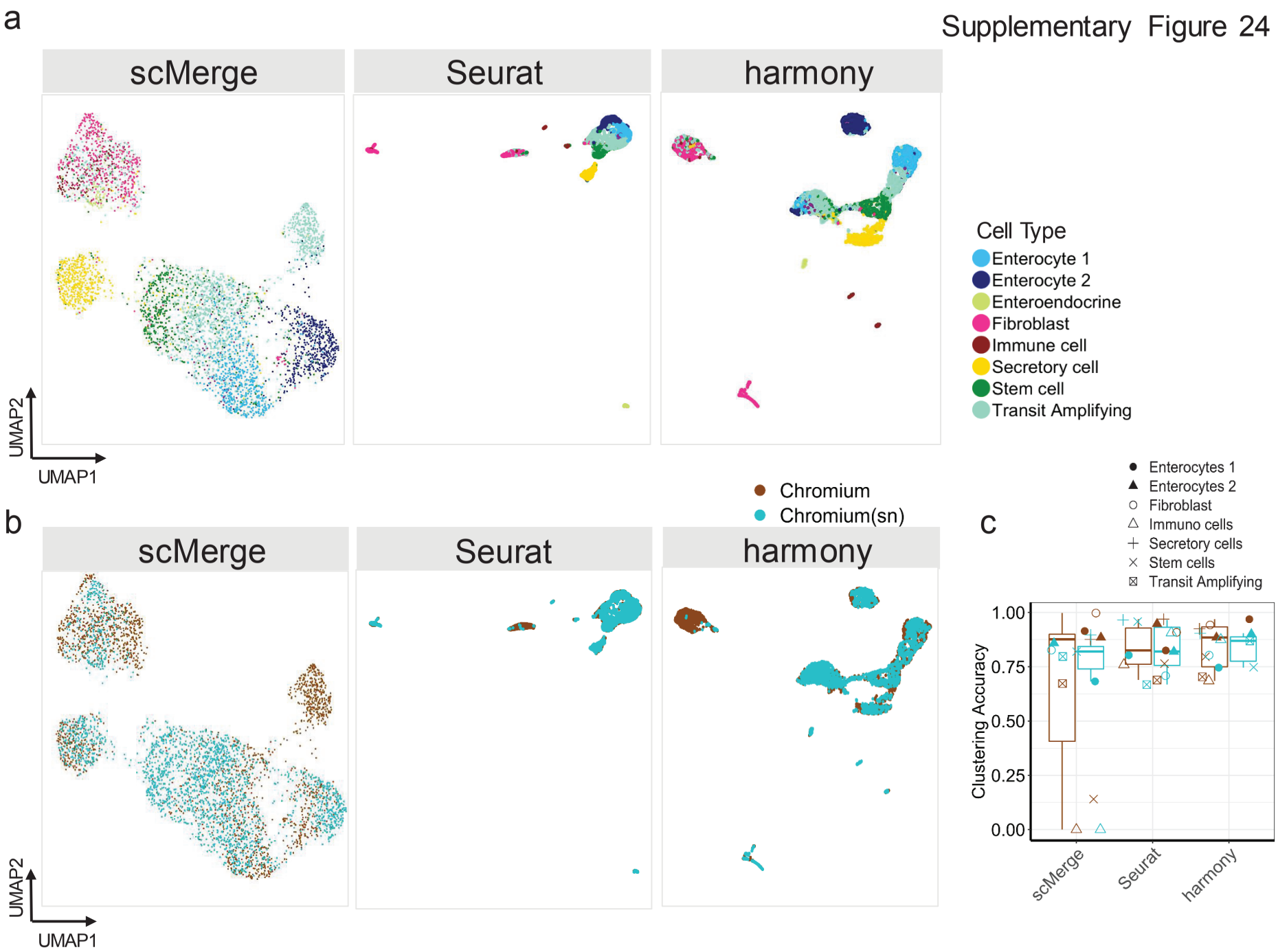




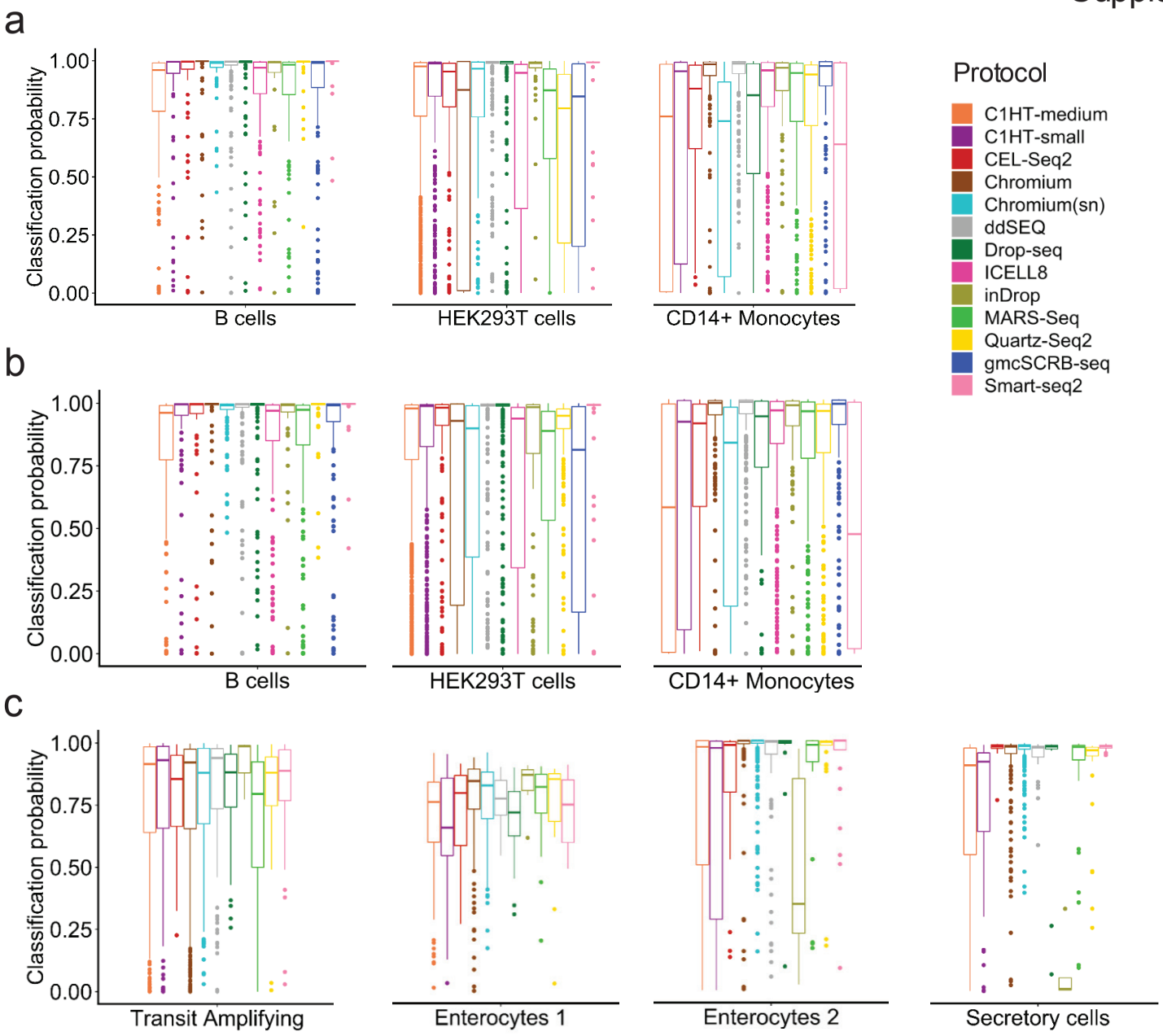
Supplementary Figure 23



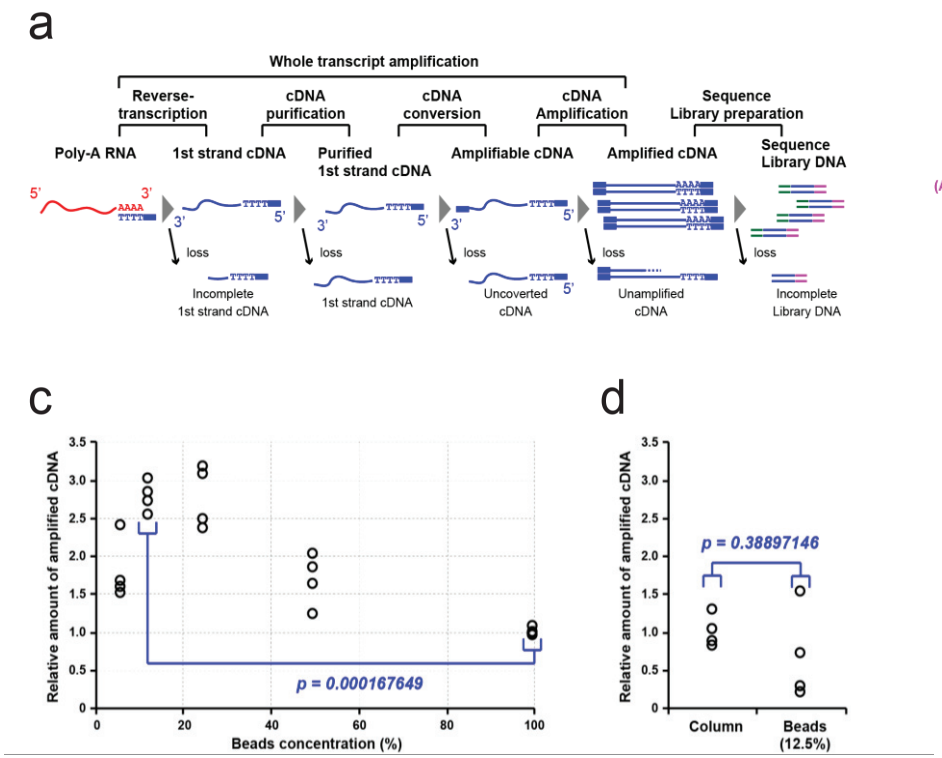
Supplementary Figure 24



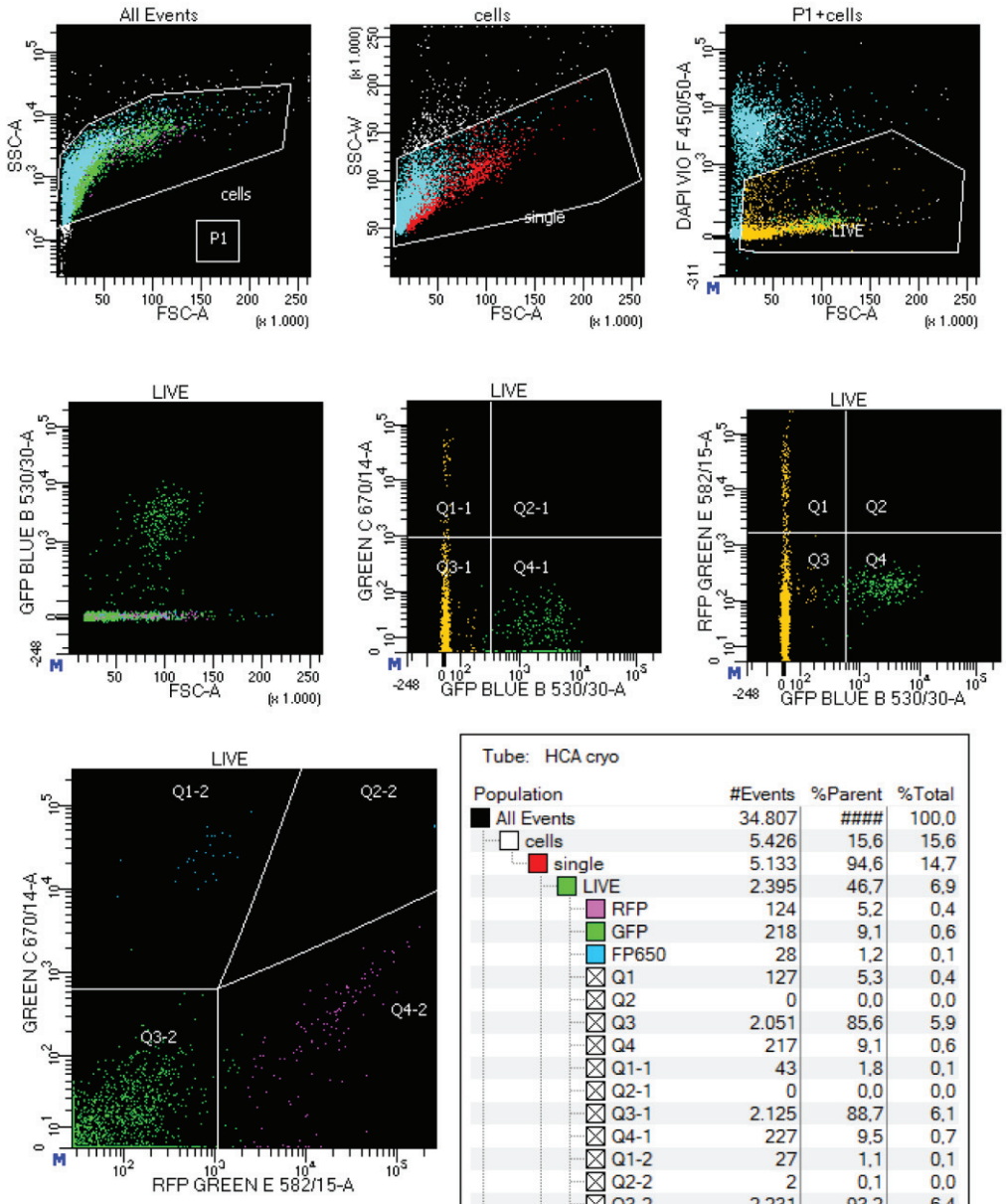
Supplementary Figure 25



Supplementary Figure 26



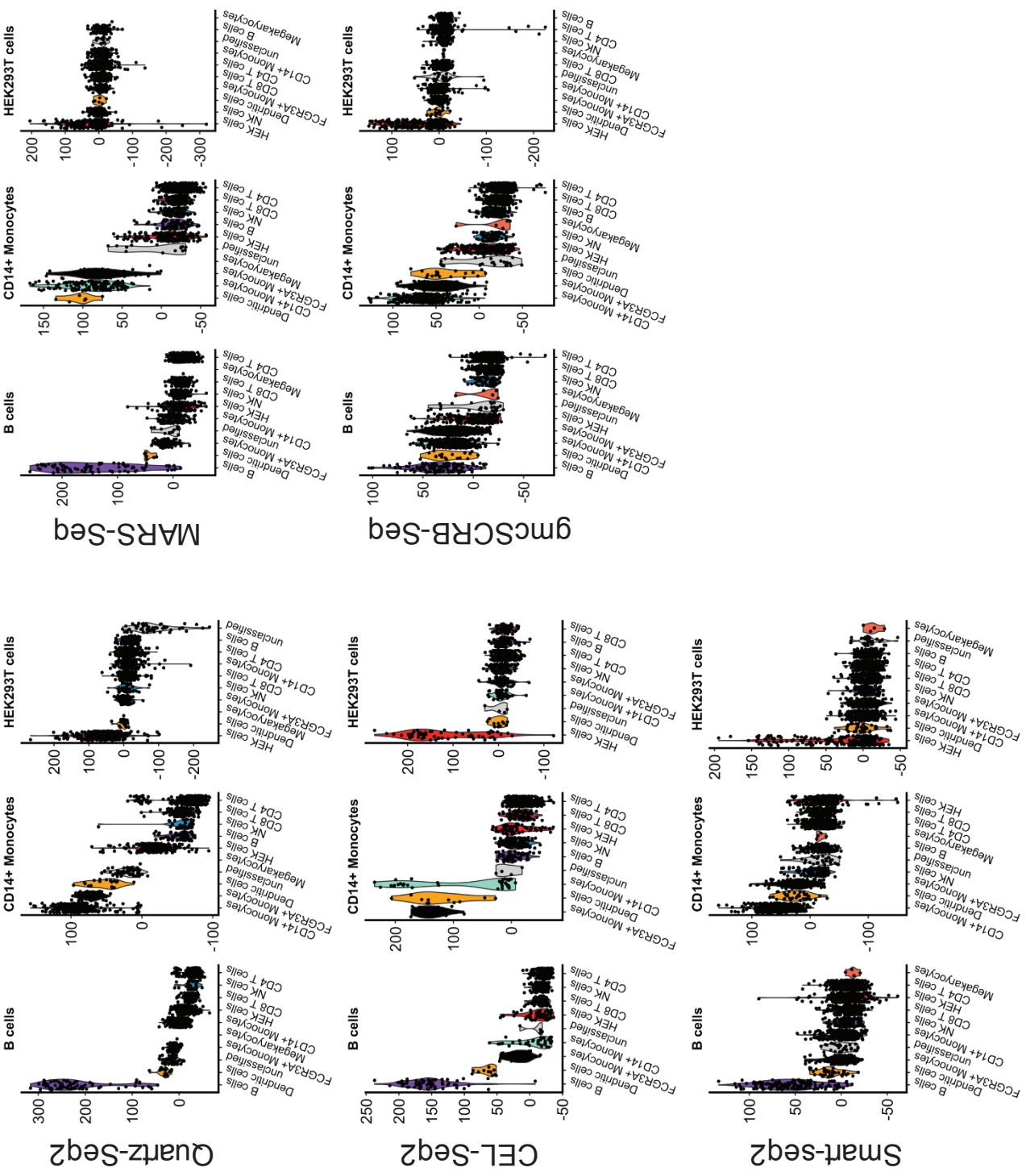
Supplementary Figure 27



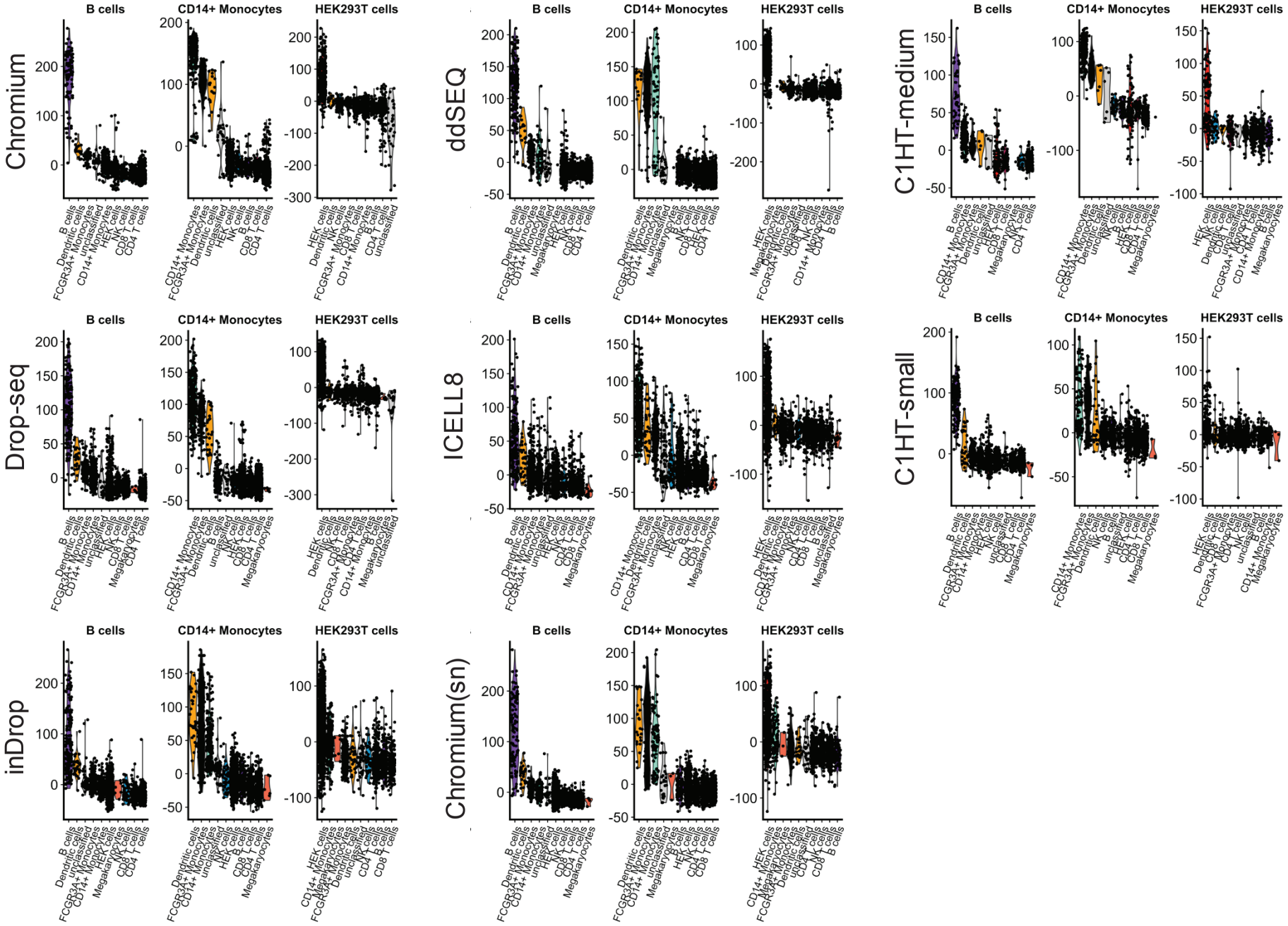
Tube: HCA cryo

Population	#Events	%Parent	%Total
All Events	34.807	####	100,0
cells	5.426	15,6	15,6
single	5.133	94,6	14,7
LIVE	2.395	46,7	6,9
RFP	124	5,2	0,4
GFP	218	9,1	0,6
FP650	28	1,2	0,1
Q1	127	5,3	0,4
Q2	0	0,0	0,0
Q3	2.051	85,6	5,9
Q4	217	9,1	0,6
Q1-1	43	1,8	0,1
Q2-1	0	0,0	0,0
Q3-1	2.125	88,7	6,1
Q4-1	227	9,5	0,7
Q1-2	27	1,1	0,1
Q2-2	2	0,1	0,0
Q3-2	2.231	93,2	6,4
Q4-2	135	5,6	0,4
GFP neg	2.160	90,2	6,2
NOT(LIVE)	2.738	53,3	7,9
FLOWCHECK	25	0,1	0,1
P1	0	0,0	0,0

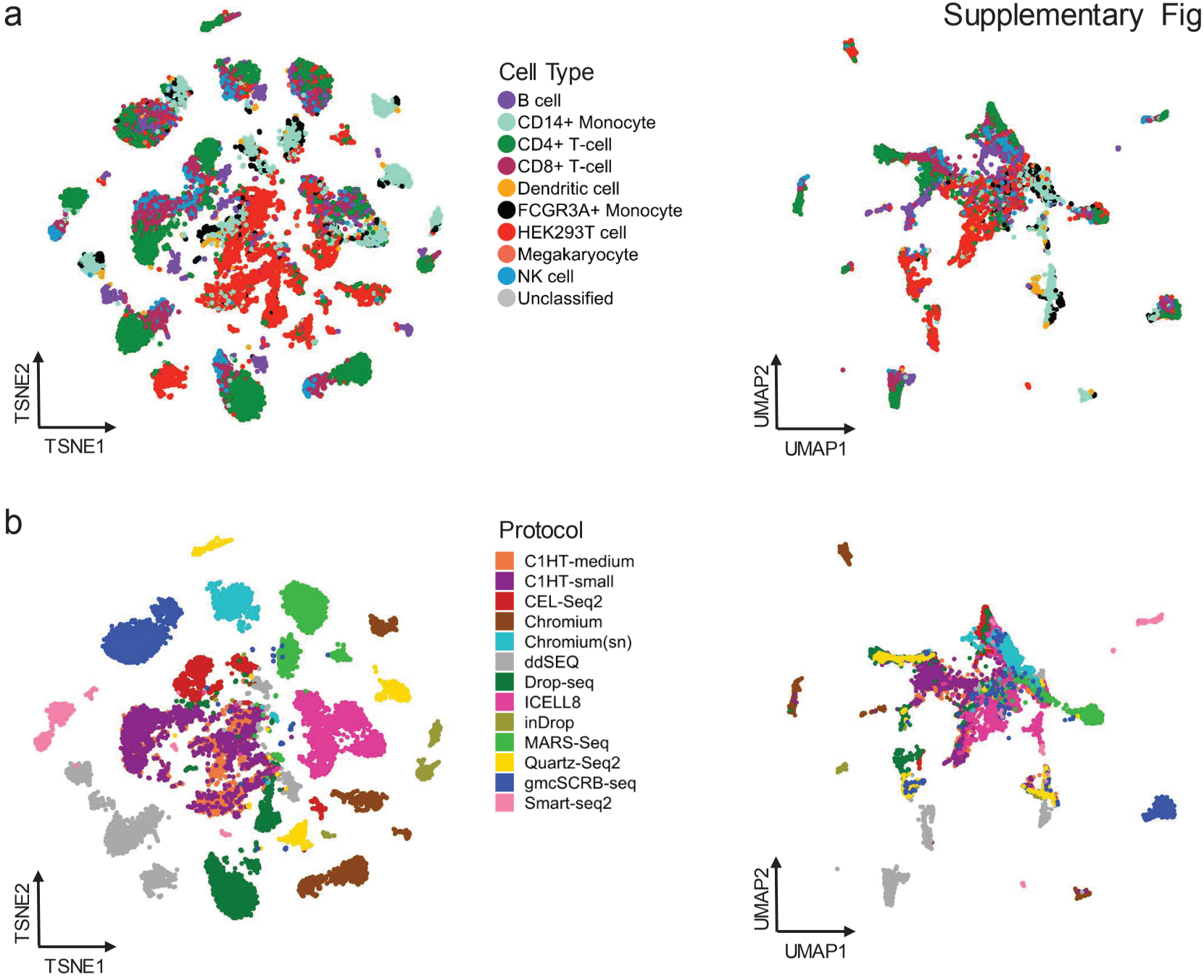
Supplementary Figure 28



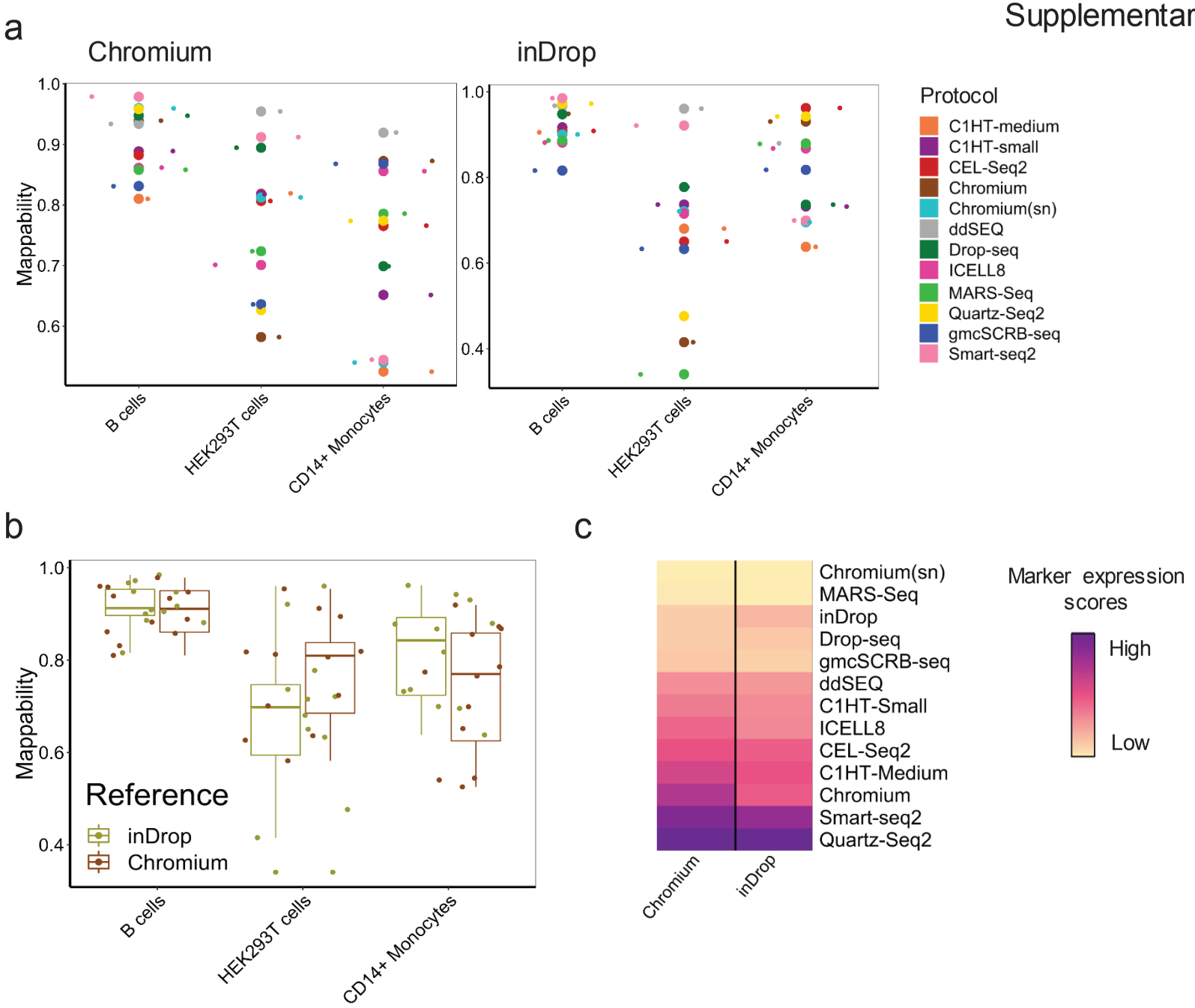
Supplementary Figure 29



Supplementary Figure 30



Supplementary Figure 31



Discussion

AML research profits from optimized bulk RNA-seq methods

Acute myeloid leukemia is a very heterogeneous disease and understanding the underlying forces resulting in the transformation of healthy hematopoietic stem cells, the progression and evolution of the cancerous cell population as well as the underlying causes of treatment resistance and relapse formation, remains a challenging task. The incorporation of expression analysis to understand the molecular pathways involved in these biological processes holds great promises but suffers from several technological complications.

We therefore set out to create a bulk RNA-sequencing protocol, which based on the advances made by single cell methods, overcomes these restrictions. We used the established scRNA-seq method SCRB-seq (Single cell RNA and barcoding sequencing) (Soumillon et al., 2014) as the basis and enhanced cell lysis and RNA-extraction prior to reverse transcription to fit the requirements. In short, cells of any source are lysed in a lysis buffer containing RLT plus buffer supplemented with 1% 2-mercaptoethanol. Afterwards, a proteinase K digestion ensures that the following RNA extraction using solid phase reversible immobilisation (SPRI) beads is not inhibited by excess protein (DeAngelis et al., 1995). After binding nucleic acids to the beads, genomic DNA is digested by DNAseI and the purified RNA can be reverse transcribed as described in the initial protocol.

Compared to existing methods this advanced bulk RNA-seq protocol is able to tackle the three major methodological challenges in combining AML and RNA-seq and simplifies the integration of RNA-seq to AML research. First, one crucial benefit is that it does not require

pre-isolated RNA as input, but can be applied to several types of samples including cell lysates from sorted cells or crude lysates from frozen cell stocks. The isolation of RNA from samples is mostly performed using commercially available kits utilizing silica membrane columns. While these are suited to extract large amounts of high quality RNA from cells, they are time consuming and cost intensive. Second, due to utilizing early barcoding it is possible to integrate thousands of samples which is furthermore facilitated by the low costs of around 3 € per sample (excluding sequencing). This allows one to incorporate more samples into each experiment in order to increase the power to identify differential expression between groups and additionally makes the method highly suited to investigate large patient cohorts. Lastly, the demand for large amounts of input material for previous methods restricts their usage to characterize rare subtypes, yielding only a few hundred cells per biological replicate (Alpern et al., 2019). Being capable of generating high quality RNA-seq data from as little as 200 cells per sample, is a strong benefit of our approach, as due to the subclonal structure of AML, the further analysis of these subclones could promote a deeper understanding in treatment resistance and relapse emergence. However, the integration of those features also comes with some caveats. For example, due to the integration of UMIs the method enriches for the 3' end of transcripts. Hence, the identification of transcript isoforms and mutations as well as allelic expression within samples based on the sequencing data is strongly limited. Both of these could be beneficial to understand the molecular pathways involved in leukemia (Batcha et al., 2019; Li et al., 2014; Petti et al., 2019).

In combination, the key advantages provide a solid ground for future applications, such as refinements of risk stratifications, characterization of rare subpopulations and treatment response. For example, we could successfully apply our method to investigate the transcriptomic phenotype of rare dormant AML cells in PDX. We could confirm a reduced

proliferation rate in AML cells which retained carboxyfluorescein succinimidyl ester (CFSE) staining (Label Retaining Cells, LRC), a dye which is not metabolized by the cell but decreases over time by cell divisions and can be measured via flow cytometry. Interestingly, label retaining cells of patients with a very different genetic background showed a very similar expression profile and were more similar to each other than to their matching non-LRC counterparts, indicating a highly conserved function of this subpopulation of cells. In addition, we found striking similarities between LRC in AML and the previous defined LRC in ALL including upregulation of cell adhesion molecules indicating localization in the hematopoietic niche. Interestingly, while both LRC and non-LRC cells were able to form tumors upon retransplantation and thereby demonstrating LIC capability in both, LRC showed increased resistance against conventional "7+3" induction chemotherapy *in vivo*. Furthermore, retransplantation experiments also showed that the label retaining feature can be reversed and vice versa suggesting it to be a temporary cell state rather than a defined subpopulation.

In conclusion, we could identify a flexible cell state in AML, characterized by dormancy and stemness which could play a major role in relapse formation. Understanding the underlying factors which drive these cell states could therefore help to further increase patients' prognosis. In addition, due to the low number of LRC cells which can be isolated per sample, we could show the usability and need for low input bulk RNA-seq methods in AML research. Furthermore, we could successfully apply this method to various sample types, including but not limited to AML relevant projects (Ebinger et al., 2016, 2020; Garz et al., 2017; Redondo Monte et al., 2020).

Improving the technical performance of scRNA-seq methods remains challenging

Molecular crowding increases sensitivity during reverse transcription

Single cell RNA sequencing (scRNA-seq) has become a wide spread tool to analyze global expression levels of single cells in various biological and medical fields (Wagner et al., 2016; Ziegenhain et al., 2018). Being a relatively new tool, scRNA-seq remains an emerging technology with it's own unique challenges and limitations. Over the last years, numerous modifications have been published focusing on increasing throughput and sensitivity and decreasing noise and costs (Ziegenhain et al., 2017, 2018). Based on a previous benchmark of several scRNA-sequencing methods (Ziegenhain et al., 2017) we therefore set out to systematically improve SCRB-seq, an efficient plate based 3` enrichment method.

Reverse transcription is considered to be the major limiting reaction step determining the sensitivity. Strikingly, it is estimated that only 10-49% of mRNA molecules present in a cell are reverse transcribed (Bagnoli et al., 2018; Grün et al., 2014; Islam et al., 2014). Unsurprisingly, we found that different MMLV derived RT enzymes showed a high variance in performance concerning sensitivity, with Maxima H- being the most sensitive. In addition to the enzyme, certain reaction enhancers have been shown to boost the efficiency of RT reactions. Interestingly, we could not replicate the positive effect of enhancers like Betaine, MgCl2 or trehalose, which had been previously reported to enhance RT efficiency in Smart-seq2 (Picelli et al., 2013). However, we found that the molecular crowding agent PEG8000 (polyethylene glycol 8000) increases cDNA yield and sensitivity. It is generally thought that molecular crowding agents increase enzymatic reaction rates by reducing the

effective reaction volume and thereby mimicking (macro)molecular crowding (Rivas and Minton, 2016). Indeed, reducing reaction volumes has been previously reported to increase the sensitivity of scRNA-seq protocols (Hashimshony et al., 2016; Svensson et al., 2017). However, very small reaction volumes also come with technical challenges and often require special equipment and are therefore less versatile than molecular crowding. In conclusion, the systematic evaluation of the reverse transcription reaction suggested that the underlying interactions between the enzyme, buffer composition, reaction volume and additives are very complex and not well understood. Accordingly, optimizations cannot easily be transferred from method to method, even though they might follow the same reaction principles.

Moreover, we saw an additional increase in sensitivity concerning the number of detected UMIs and genes when using Terra direct polymerase during cDNA amplification. Due to the very low starting material, scRNA-seq protocols require a considerable amount of amplification to acquire a sufficient amount of material for sequencing. Hence, polymerase biases such as preferential amplification, e.g. due to GC content or initial abundance, can lead to major problems. Although UMIs can generally remove such amplification biases computationally, it nevertheless can influence the sensitivity of a protocol as sequencing duplicated reads (having the same UMI) decreases detection efficiency at a given sequencing depth.

Together with other minor changes, including but not limited to enhanced pooling strategies and lysis conditions, we established molecular crowding SCRB-seq (mcSCRB-seq) (Bagnoli et al., 2018). In order to verify the relative increase in sensitivity of mcSCRB-seq compared to SCRB-seq we generated a side by side comparison using mouse embryonic stem cells (mESCs). Indeed, we saw a 2.5x increase in detected UMIs, and therefore detected unique RNA molecules. To further quantify the sensitivity in a more quantitative manner we used

ERCC spike in molecules to first calculate the capture sensitivity and second to compare mcSCRB-seq to other published scRNA-seq protocols (Baker et al., 2005; Jiang et al., 2011). At a sequencing depth of two million reads mcSCRB was able to detect 48.9% of spiked-in ERCC molecules. As proposed by others (Svensson et al., 2017), we used binomial logistic regression to model the detection of ERCC transcripts in relation to their initial abundance. This enabled us to compare mcSCRB-seq to various other scRNA-seq protocols by integrating our newly generated data with a previous published benchmarkings and methods (Sasagawa et al., 2017; Svensson et al., 2017; Zheng et al., 2017; Ziegenhain et al., 2017). Utilizing this comparison approach we could show that mcSCRB-seq is one of the most sensitive methods, requiring only 2 RNA molecules present to achieve a 50% detection efficiency.

Although ERCC spike ins are often the only possibility to compare the sensitivity of several methods efficiently, there are several factors which could possibly lead to incorrect conclusions from such analysis. First of all, to calculate absolute abundances of RNA molecules and detection efficiencies it must be assumed that ERCC molecules are equally likely to be converted to cDNA as endogenous mRNA molecules. Several design attributes of these spike ins could contribute to rule this assumption invalid. ERCC molecules, for example, do not contain 5`cap structures. However, recent studies have shown that these structures lead to the efficient stalling of the RT enzyme at the 5'end leading to increased template switching efficiencies (Wulf et al., 2019). Moreover, other physical properties such as length (250-2000 nucleotides), GC content (5-51%) and the short polyA-tails do not perfectly model endogenous mRNA transcripts (Grün et al., 2014; Stegle et al., 2015; Svensson et al., 2017). Lastly, we also saw that other, non protocol derived, factors could contribute to the capture efficiencies. This was evident as several Smart-seq2 and CEL-seq2

datasets showed very different capture efficiencies, depending on the study. In addition, particularly the Smart-seq2 dataset used in *Svensson et al.* showed a very high cell to cell variation spanning over a 1000 fold range, subdivided in three distinct populations. Both of these observations could possibly show that capture efficiency could be highly driven by batch to batch variation but also by distinct external factors which differ from lab to lab or experiment to experiment.

All in all, we could highly improve SCRB-seq to be one of the most sensitive methods, while still retaining its advantages concerning costs (initial and running) and scalability over other highly sensitive methods, such as CEL-seq2 and Smart-seq2. In addition, the advantageous effect of molecular crowding during reverse transcription was also recently proven to enhance Smart-seq3 as well as Seq-Well (Hagemann-Jensen et al., 2020; Hughes et al., 2019).

Systematic comparison of mcSCRB-seq shows potential for additional improvements

As mentioned above, the systematic comparison of methods using ERCC spike in molecules can suffer from bias. Hence, we took part in a benchmarking study for the HCA (Human Cell Atlas) Consortium driven to overcome such limitations (Mereu et al., 2020). In contrast to previous studies, and to prevent user related errors or influences, all participating scRNA-seq methods were performed within their corresponding development lab or if not possible, in groups with expertise in using these. In order to validate the performance of each method accordingly, each group was supplied with aliquots of a frozen cell suspension prepared in one batch. The complex composition of this reference sample ensured that several key factors that are particularly important in creating expression atlases can be addressed. For example, both highly different as well as closely related cell types (PBMCs) were included, and all cell types are well characterised with distinct expression markers. In addition, the sample included a wide range of cell sizes, as expected in heterogeneous tissues. Last but not least, different species were included in order to estimate cross contamination between cells.

We generated and sequenced mcSCRB-seq libraries from 3008 single cells of this reference sample which were further analysed and compared by the responsible center lab. Interestingly, we found that using the same lysis buffer as previously used for mESCs, lead to degraded cDNA, implying RNA degradation during single cell isolation, possibly via endogenous RNase enzymes of the cells. We therefore decided to use a more stringent lysis condition, which we previously used for PBMCs (Bagnoli et al., 2018) as well as AML and ALL cells (Bagnoli et al., 2019). However, since the stringent lysis condition requires an

additional clean up step before reverse transcription, which is not present in the original mcSCRB-seq protocol, we named this method gmcSCRB-seq (Guanidine mcSCRB-seq).

Surprisingly, when comparing the sensitivity of the participating methods by the number of detected genes using exonic and intronic mapping reads, gmcSCRB-seq performed poorly across all cell types. For example, methods, which showed a similar sensitivity using ERCC spike ins as described above, detected around 2.5x more genes at the same sequencing depth of 25,000 reads per cell. Furthermore, while the major sensitivity trends were consistent across the different cell types, some methods performed disproportionately better with cells having a high RNA content. Hence, different methods might be more or less suitable for different cell types.

It should be noted that a high variability between the methods could be seen when comparing mapping feature distribution. For example, gmcSCRB-seq, Quartz-seq2 and Smart-seq2 showed a very high exonic mapping fraction, while CEL-seq2, MARS-seq and C1HT showed a higher intronic and intergenic fraction. To what extent intronic and intergenic mapping reads should be addressed as further viable information, for example by using intronic reads to predict future cell fate decisions (La Manno et al., 2018), is still under discussion (Parekh et al., 2018). Especially intergenic mapping reads could hint towards off target binding of primers on gDNA, which subsequently would dilute the expression signal.

Moreover, the initial complexity of the reference sample could not be retained in gmcSCRB-seq. For example, canine cells, which were present at 1% could be detected in all methods (1-9%) except for gmcSCRB-seq. In addition, it was found that mouse colon cells were more present when fluorescence activated cell sorting with viability staining was not performed, suggesting a higher sensitivity of these cells towards the sample preparation. These observations clearly expose that sample processing as well as single cell isolation

strategies can have a big impact on the performance of each method and should be taken into account. This further demonstrated that using ERCC spike ins to assess the performance cannot be fully representative.

To further estimate the accuracy of the methods, marker gene expression and cluster accuracy was taken into account. In addition to the low sensitivity performance, gmcSCRB-seq performed poorly for marker gene detection as well as clustering. Interestingly, marker expression and clusterability did not always correlate. For example, while the ICELL8 SMARTer Single-Cell System performed better than gmcSCRB-seq and MARS-seq concerning marker expression, clusterability of the latter two was better. While a low sensitivity can subsequently cause missing clusterability, several other reasons which are partially difficult or even impossible to prove are within the realm of possibilities. For example, within methods utilizing early barcoding and pooled amplifications, leftover barcoded primers from the RT can introduce noise during amplification (Macosko et al., 2015). In addition, chimeric PCR fragments are common in multi template PCRs and are known to be a major source for cross contamination (Dixit, 2016; Kalle et al., 2014).

Taken together, the combined data of this extensive benchmark not only provides guidance for researchers and consortia to select an appropriate method but in addition demonstrates key factors for further improvements of methods and reveals missing interpretability of single cell expression data.

For gmcSCRB-seq and mcSCRB-seq in particular, the benchmarking data in combination with the previous development data, pinpoint to several possible enhancements to further optimize the methods, such as cell isolation, cell lysis and possible cross contamination.

Conclusion and Outlook

Whole transcriptome analysis is currently undergoing a transition phase between a niche method requiring expertise and large budgets towards a commonly used tool to investigate biomedical processes, especially in cancer. Both, single cell and bulk RNA-seq can possibly contribute to further understanding the cause, development and resistance in AML and even promise to be powerful diagnostic tools. However, especially single cell RNA-seq is still a rapidly evolving technique and methods still need to be optimized

In this work, I developed method improvements for single-cell and bulk RNA sequencing and applied it to a relevant biomedical question investigating acute myeloid leukemia. Modifying a previous scRNA-seq protocol towards small input bulk RNA-sequencing, enabled us and our collaborators to further understand a rare subgroup of AML cells and refine the knowledge of the cancer stem cell theory and its role in relapse. Additionally, we set the basis for large scale RNA-seq experiments using large patient cohorts unrestricted to sample quantities and qualities.

Furthermore, we have contributed methodological optimizations to the field of scRNA-seq tackling low sensitivity, high costs and complicated workflows. For example, the concept of molecular crowding to enhance reverse transcription has been applied several times by now. However, further improvements and research is still required to be able to fully unleash and utilize the potential and the power of scRNA-sequencing, especially for cancer research and diagnostics.

References

Adamia, S., Haibe-Kains, B., Pilarski, P.M., Bar-Natan, M., Pevzner, S., Avet-Loiseau, H., Lode, L., Verselis, S., Fox, E.A., Burke, J., et al. (2014). A genome-wide aberrant RNA splicing in patients with acute myeloid leukemia identifies novel potential disease markers and therapeutic targets. Clin. Cancer Res. *20*, 1135–1145.

Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., et al. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. *11*, R119.

Albrecht, T.A. (2014). Physiologic and psychological symptoms experienced by adults with acute leukemia: an integrative literature review. Oncol. Nurs. Forum *41*, 286–295.

Alpern, D., Gardeux, V., Russeil, J., Mangeat, B., Meireles-Filho, A.C.A., Breysse, R., Hacker, D., and Deplancke, B. (2019). BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. Genome Biol. *20*, 71.

Alwine, J.C., Kemp, D.J., and Stark, G.R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. U. S. A. *74*, 5350–5354.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Arber, D.A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M.J., Le Beau, M.M., Bloomfield, C.D., Cazzola, M., and Vardiman, J.W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood *127*, 2391–2405.

Avery, O.T., Macleod, C.M., and 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. J. Exp. Med. *79*, 137–158.

Bagnoli, J.W., Ziegenhain, C., Janjic, A., Wange, L.E., Vieth, B., Parekh, S., Geuder, J., Hellmann, I., and Enard, W. (2018). Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. Nat. Commun. *9*, 2937.

Bagnoli, J.W., Wange, L.E., Janjic, A., and Enard, W. (2019). Studying Cancer Heterogeneity by Single-Cell RNA Sequencing. In Lymphoma: Methods and Protocols, R. Küppers, ed. (New York, NY: Springer New York), pp. 305–319.

Bainbridge, M.N., Warren, R.L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., et al. (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics *7*, 246.

Baker, S.C., Bauer, S.R., Beyer, R.P., Brenton, J.D., Bromley, B., Burrill, J., Causton, H., Conley, M.P., Elespuru, R., Fero, M., et al. (2005). The External RNA Controls Consortium: a progress report. Nat. Methods *2*, 731–734.

Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A., and Grant, G.R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat. Methods *14*, 135–139.

Batcha, A.M.N., Bamopoulos, S.A., Kerbs, P., Kumar, A., Jurinovic, V., Rothenberg-Thurley, M., Ksienzyk, B., Philippou-Massier, J., Krebs, S., Blum, H., et al. (2019). Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia. Sci. Rep. *9*, 11796.

Bcop, D.M.P., Marjoncu, D., Phar, BCOP, Andrick, B., Phar, and BCOP (2020). Gilteritinib: A Novel FLT3 Inhibitor for Relapsed/Refractory Acute Myeloid Leukemia. Journal of the Advanced Practitioner in Oncology *11*.

Becker-André, M., and 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 Res. *17*, 9437–9446.

Bennett, J.M., Catovsky, D., Daniel, M.T., Flandrin, G., Galton, D.A., Gralnick, H.R., and Sultan, C. (1976). Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. Br. J. Haematol. *33*, 451–458.

van den Brink, S.C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C., and van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat. Methods *14*, 935–936.

Cancer Genome Atlas Research Network, Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A.G., Hoadley, K., Triche, T.J., Jr, Laird, P.W., et al. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. *368*, 2059–2074.

Cheung, F., Haas, B.J., Goldberg, S.M.D., May, G.D., Xiao, Y., and Town, C.D. (2006). Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. BMC Genomics *7*, 272.

Choy, J.Y.H., Boon, P.L.S., Bertin, N., and Fullwood, M.J. (2015). A resource of ribosomal RNA-depleted RNA-Seq data from different normal adult and fetal human tissues. Sci Data *2*, 150063.

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol. *4*, 265–270.

Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods *5*, 613–619.

Cocciardi, S., Dolnik, A., Kapp-Schwoerer, S., Rücker, F.G., Lux, S., Blätte, T.J., Skambraks, S., Krönke, J., Heidel, F.H., Schnöder, T.M., et al. (2019). Clonal evolution patterns in acute myeloid leukemia with NPM1 mutation. Nat. Commun. *10*, 2031.

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

DeAngelis, M.M., Wang, D.G., and Hawkins, T.L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res. *23*, 4742–4743.

De Kouchkovsky, I., and Abdul-Hay, M. (2016). Acute myeloid leukemia: a comprehensive review and 2016 update. Blood Cancer J. *6*, e441.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510.

Dixit, A. (2016). Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments.

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.

Döhner, H., Estey, E.H., Amadori, S., Appelbaum, F.R., Büchner, T., Burnett, A.K., Dombret, H., Fenaux, P., Grimwade, D., Larson, R.A., et al. (2010). Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. Blood *115*, 453–474.

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F.R., Büchner, T., Dombret, H., Ebert, B.L., Fenaux, P., Larson, R.A., et al. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood *129*, 424–447.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78–81.

Ebinger, S., Özdemir, E.Z., Ziegenhain, C., Tiedt, S., Castro Alves, C., Grunert, M., Dworzak, M., Lutz, C., Turati, V.A., Enver, T., et al. (2016). Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia. Cancer Cell *30*, 849–862.

Ebinger, S., Zeller, C., Carlet, M., Senft, D., Bagnoli, J.W., Liu, W.-H., Rothenberg-Thurley, M., Enard, W., Metzeler, K.H., Herold, T., et al. (2020). Plasticity in growth behavior of patients' acute myeloid leukemia stem cells growing in mice. Haematologica.

Edfors, F., Danielsson, F., Hallström, B.M., Käll, L., Lundberg, E., Pontén, F., Forsström, B., and Uhlén, M. (2016). Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol. Syst. Biol. *12*, 883.

Emrich, S.J., Barbazuk, W.B., Li, L., and Schnable, P.S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res. *17*, 69–73.

Engström, P.G., The RGASP Consortium, Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Rätsch, G., Goldman, N., Hubbard, T.J., Harrow, J., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. Nature Methods *10*, 1185–1191.

Estey, E.H. (2014). Acute myeloid leukemia: 2014 update on risk-stratification and management. Am. J. Hematol. *89*, 1063–1081.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. *34*, e22.

Gale, R.E., Green, C., Allen, C., Mead, A.J., Burnett, A.K., Hills, R.K., Linch, D.C., and Medical Research Council Adult Leukaemia Working Party (2008). The impact of FLT3 internal tandem duplication mutant level, number, size, and interaction with NPM1 mutations in a large cohort of young adult patients with acute myeloid leukemia. Blood *111*, 2776–2784.

van Galen, P., Hovestadt, V., Wadsworth, M.H., Ii, Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Lombardi Story, J., et al. (2019). Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell *176*, 1265–1281.e24.

Garz, A.-K., Wolf, S., Grath, S., Gaidzik, V., Habringer, S., Vick, B., Rudelius, M., Ziegenhain, C., Herold, S., Weickert, M.-T., et al. (2017). Azacitidine combined with the selective FLT3 kinase inhibitor crenolanib disrupts stromal protection and inhibits expansion of residual leukemia-initiating cells in FLT3-ITD AML with concurrent epigenetic mutations. Oncotarget *8*, 108738–108759.

Gilliland, D.G., and Griffin, J.D. (2002). The roles of FLT3 in hematopoiesis and leukemia. Blood *100*, 1532–1542.

Grimwade, D., Hills, R.K., Moorman, A.V., Walker, H., Chatters, S., Goldstone, A.H., Wheatley, K., Harrison, C.J., Burnett, A.K., and on behalf of the National Cancer Research Institute Adult Leukaemia Working Group (2010). Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. Blood *116*, 354–365.

Grossmann, V., Schnittger, S., Kohlmann, A., Eder, C., Roller, A., Dicker, F., Schmid, C., Wendtner, C.-M., Staib, P., Serve, H., et al. (2012). A novel hierarchical prognostic model of AML solely based on molecular mutations. Blood *120*, 2963–2972.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640.

Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A.J.M., Faridani, O.R., and Sandberg, R. (2020). Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nat. Biotechnol.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

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 Biology *17*.

Hayer, K.E., Pizarro, A., Lahens, N.F., Hogenesch, J.B., and Grant, G.R. (2015). Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. Bioinformatics *31*, 3938–3945.

Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R., and Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. Biotechniques *56*, 61–64, 66, 68, passim.

Herold, T., Jurinovic, V., Batcha, A.M.N., Bamopoulos, S.A., Rothenberg-Thurley, M., Ksienzyk, B., Hartmann, L., Greif, P.A., Phillippou-Massier, J., Krebs, S., et al. (2018). A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. Haematologica *103*, 456–465.

Hrdlickova, R., Toloue, M., and Tian, B. (2017). RNA-Seq methods for transcriptome analysis. Wiley Interdisciplinary Reviews: RNA *8*, e1364.

Hughes, T.K., Wadsworth, M.H., Gierahn, T.M., Do, T., Weiss, D., Andrade, P.R., Ma, F., de Andrade Silva, B.J., Shao, S., Tsoi, L.C., et al. (2019). Highly Efficient, Massively-Parallel Single-Cell RNA-Seq Reveals Cellular States and Molecular Features of Human Skin Pathology.

Iacobucci, I., Lonetti, A., Candoni, A., Sazzini, M., Papayannidis, C., Formica, S., Ottaviani, E., Ferrari, A., Michelutti, A., Simeone, E., et al. (2013). Profiling of drug-metabolizing enzymes/transporters in CD33+ acute myeloid leukemia patients treated with Gemtuzumab-Ozogamicin and Fludarabine, Cytarabine and Idarubicin. Pharmacogenomics J. *13*, 335–341.

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., et al. (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.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

Kalle, E., Kubista, M., and Rensing, C. (2014). Multi-template polymerase chain reaction. Biomol Detect Quantif *2*, 11–29.

Kałużna, M., Kuras, A., Mikiciński, A., and Puławska, J. (2016). Evaluation of different RNA extraction methods for high-quality total RNA and mRNA from Erwinia amylovora in planta. Eur. J. Plant Pathol. *146*, 893–899.

Kantarjian, H. (2016). Acute myeloid leukemia-Major progress over four decades and glimpses into the future. American Journal of Hematology *91*, 131–145.

Kircher, M., and Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. Bioessays *32*, 524–536.

Kircher, M., Stenzel, U., and Kelso, J. (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol. *10*, R83.

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. *40*, e3.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. Nature 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.

Knorre, D.G., Kudryashova, N.V., and Godovikova, T.S. (2009). Chemical and functional aspects of posttranslational modification of proteins. Acta Naturae *1*, 29–51.

Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nat. Commun. *10*, 5416.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The technology and biology of single-cell RNA sequencing. Mol. Cell *58*, 610–620.

Krug, U., Röllig, C., Koschmieder, A., Heinecke, A., Sauerland, M.C., Schaich, M., Thiede, C., Kramer, M., Braess, J., Spiekermann, K., et al. (2010). Complete remission and early death after intensive chemotherapy in patients aged 60 years or older with acute myeloid leukaemia: a web-based application for prediction of outcomes. Lancet *376*, 2000–2008.

Kurimoto, K. (2006). An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. Nucleic Acids Research *34*, e42–e42.

Kurimoto, K., Yabuta, Y., Ohinata, Y., and Saitou, M. (2007). Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. Nature Protocols *2*, 739–752.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods *7*, 709–715.

Li, X.-Y., Yao, X., Li, S.-N., Suo, A.-L., Ruan, Z.-P., Liang, X., Kong, Y., Zhang, W.-G., and Yao, Y. (2014). RNA-Seq profiling reveals aberrant RNA splicing in patient with adult acute myeloid leukemia during treatment. Eur. Rev. Med. Pharmacol. Sci. *18*, 1426–1433.

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

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. PLoS Comput. Biol. *13*, e1005457.

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.

Marcucci, G., Maharry, K.S., Metzeler, K.H., Volinia, S., Wu, Y.-Z., Mrózek, K., Nicolet, D., Kohlschmidt, J., Whitman, S.P., Mendler, J.H., et al. (2013). Clinical role of microRNAs in cytogenetically normal acute myeloid leukemia: miR-155 upregulation independently identifies high-risk patients. J. Clin. Oncol. *31*, 2086–2093.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research *18*, 1509–1517.

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Álvarez-Varela, A., Batlle, E., Sagar, Grün, D., Lau, J.K., et al. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat. Biotechnol.

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. *2010*, db.prot5448.

Meyers, J., Yu, Y., Kaye, J.A., and Davis, K.L. (2013). Medicare fee-for-service enrollees with primary acute myeloid leukemia: an analysis of treatment patterns, survival, and healthcare resource utilization and costs. Appl. Health Econ. Health Policy *11*, 275–286.

Miller, M.B., and Tang, Y.-W. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. Clin. Microbiol. Rev. *22*, 611–633.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Okoniewski, M.J., and Miller, C.J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics *7*, 276.

Ong, C.-T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. *12*, 283–293.

Pachmann, K. (1987). In situ hybridization with fluorochrome-labeled cloned DNA for quantitative determination of the homologous mRNA in individual cells. J. Mol. Cell. Immunol. *3*, 13–19.

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.

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

Petti, A.A., Williams, S.R., Miller, C.A., Fiddes, I.T., Srivatsan, S.N., Chen, D.Y., Fronick, C.C., Fulton, R.S., Church, D.M., and Ley, T.J. (2019). A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. Nat. Commun. *10*, 3660.

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. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. *24*, 2033–2040.

Pollyea, D.A., and Jordan, C.T. (2017). Therapeutic targeting of acute myeloid leukemia stem cells. Blood *129*, 1627–1635.

Pozhitkov, A.E., Tautz, D., and Noble, P.A. (2007). Oligonucleotide microarrays: widely applied—poorly understood. Brief. Funct. Genomics.

Rana, T.M. (2007). Illuminating the silence: understanding the structure and function of small RNAs. Nat. Rev. Mol. Cell Biol. *8*, 23–36.

Ravandi, F., Alattar, M.L., Grunwald, M.R., Rudek, M.A., Rajkhowa, T., Richie, M.A., Pierce, S., Daver, N., Garcia-Manero, G., Faderl, S., et al. (2013). Phase 2 study of azacytidine plus sorafenib in patients with acute myeloid leukemia and FLT-3 internal tandem duplication mutation. Blood *121*, 4655–4662.

Redondo Monte, E., Wilding, A., Leubolt, G., Kerbs, P., Bagnoli, J.W., Hartmann, L., Hiddemann, W., Chen-Wichmann, L., Krebs, S., Blum, H., et al. (2020). ZBTB7A prevents RUNX1-RUNX1T1-dependent clonal expansion of human hematopoietic stem and progenitor cells. Oncogene.

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.

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 Research *43*, e47.

Rivas, G., and Minton, A.P. (2016). Macromolecular Crowding In Vitro, In Vivo, and In Between. Trends Biochem. Sci. *41*, 970–981.

Röllig, C., Serve, H., Hüttmann, A., Noppeney, R., Müller-Tidow, C., Krug, U., Baldus, C.D., Brandts, C.H., Kunzmann, V., Einsele, H., et al. (2015). Addition of sorafenib versus placebo to standard therapy in patients aged 60 years or younger with newly diagnosed acute myeloid leukaemia (SORAML): a multicentre, phase 2, randomised controlled trial. Lancet Oncol. *16*, 1691–1699.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. Nature *265*, 687–695.

Sasagawa, Y., Danno, H., Takada, H., Ebisawa, M., Hayashi, T., Kurisaki, A., and Nikaido, I. (2017). Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467–470.

Shah, A., Andersson, T.M.-L., Rachet, B., Björkholm, M., and Lambert, P.C. (2013). Survival and cure of acute myeloid leukaemia in England, 1971-2006: a population-based study. Br. J. Haematol. *162*, 509–516.

Shu, C., Sun, S., Chen, J., Chen, J., and Zhou, E. (2014). Comparison of different methods for total RNA extraction from sclerotia of Rhizoctonia solani. Electron. J. Biotechnol. *17*, 9–9.

Sim, G.K., Kafatos, F.C., Jones, C.W., Koehler, M.D., Efstratiadis, A., and Maniatis, T. (1979). Use of a cDNA library for studies on evolution and developmental expression of the chorion multigene families. Cell *18*, 1303–1316.

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.

Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. *16*, 133–145.

Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R.A., and Skotheim, R.I. (2016). Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene *35*, 2413–2427.

Svensson, V., Natarajan, K.N., Ly, L.-H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., and Teichmann, S.A. (2017). Power analysis of single-cell RNA-sequencing experiments. Nat. Methods *14*, 381–387.

Swaminathan, M., Kantarjian, H.M., Daver, N., Borthakur, G., Ohanian, M., Kadia, T., DiNardo, C.D., Jain, N., Estrov, Z., Ferrajoli, A., et al. (2017). The combination of quizartinib with azacitidine or low dose cytarabine is highly active in patients (Pts) with FLT3-ITD mutated myeloid leukemias: interim report of a phase I/II trial. Blood *130*, 723–723.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.

Tarca, A.L., Bhatti, G., and Romero, R. (2013). A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLoS One *8*, e79217.

Utterback, J. (2020). Illumina Remains the Clear Leader of the Genomic Sequencing Market (Morningstar). Available at: http://analysisreport.morningstar.com/stock/research/c-report?&t=XNAS:ILMN [Accessed March 10th,2020]

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., et al. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. *18*, 1051–1063.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. *10*, 252–263.

Vardiman, J.W., Thiele, J., Arber, D.A., Brunning, R.D., Borowitz, M.J., Porwit, A., Harris, N.L., Le Beau, M.M., Hellström-Lindberg, E., Tefferi, A., et al. (2009). The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. Blood *114*, 937–951.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484–487.

Vera, J.C., Wheat, C.W., Fescemyer, H.W., Frilander, M.J., Crawford, D.L., Hanski, I., and Marden, J.H. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol. Ecol. *17*, 1636–1647.

Veuger, M.J., Honders, M.W., Landegent, J.E., Willemze, R., and Barge, R.M. (2000). High incidence of alternatively spliced forms of deoxycytidine kinase in patients with resistant acute myeloid leukemia. Blood *96*, 1517–1524.

Veuger, M.J.T., Heemskerk, M.H.M., Honders, M.W., Willemze, R., and Barge, R.M.Y. (2002). Functional role of alternatively spliced deoxycytidine kinase in sensitivity to cytarabine of acute myeloid leukemic cells. Blood *99*, 1373–1380.

Vick, B., Rothenberg, M., Sandhöfer, 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.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

Voss, T.C., and Hager, G.L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat. Rev. Genet. *15*, 69–81.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nat. Biotechnol. *34*, 1145–1160.

Walter, R.B., Othus, M., Burnett, A.K., Löwenberg, B., Kantarjian, H.M., Ossenkoppele, G.J., Hills, R.K., van Montfort, K.G.M., Ravandi, F., Evans, A., et al. (2013). Significance of FAB subclassification of "acute myeloid leukemia, NOS" in the 2008 WHO classification: analysis of 5848 newly diagnosed patients. Blood *121*, 2424–2431.

Walter, R.B., Othus, M., Burnett, A.K., Löwenberg, B., Kantarjian, H.M., Ossenkoppele, G.J., Hills, R.K., Ravandi, F., Pabst, T., Evans, A., et al. (2015a). Resistance prediction in AML: analysis of 4601 patients from MRC/NCRI, HOVON/SAKK, SWOG and MD Anderson Cancer Center. Leukemia *29*, 312–320.

Walter, R.B., Othus, M., Paietta, E.M., Racevskis, J., Fernandez, H.F., Lee, J.-W., Sun, Z., Tallman, M.S., Patel, J., Gönen, M., et al. (2015b). Effect of genetic profiling on prediction of therapeutic resistance and survival in adult acute myeloid leukemia. Leukemia *29*, 2104–2107.

Wang, E.S., Stone, R.M., Tallman, M.S., Walter, R.B., Eckardt, J.R., and Collins, R. (2016). Crenolanib, a Type I FLT3 TKI, Can be Safely Combined with Cytarabine and Anthracycline Induction Chemotherapy and Results in High Response Rates in Patients with Newly Diagnosed FLT3 Mutant Acute Myeloid Leukemia (AML). Blood *128*, 1071–1071.

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184–2185.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Warner, J.R. (1999). The economics of ribosome biosynthesis in yeast. Trends Biochem. Sci. *24*, 437–440.

Weber, A.P.M., Weber, K.L., Carr, K., Wilkerson, C., and Ohlrogge, J.B. (2007). Sampling the Arabidopsis Transcriptome with Massively Parallel Pyrosequencing. Plant Physiology *144*, 32–42.

Weis, J.H., Tan, S.S., Martin, B.K., and Wittwer, C.T. (1992). Detection of rare mRNAs via quantitative RT-PCR. Trends Genet. *8*, 263–264.

Wiernik, P.H., Banks, P.L., Case, D.C., Jr, Arlin, Z.A., Periman, P.O., Todd, M.B., Ritch, P.S., Enck, R.E., and Weitberg, A.B. (1992). Cytarabine plus idarubicin or daunorubicin as induction and consolidation therapy for previously untreated adult patients with acute myeloid leukemia. Blood *79*, 313–319.

Wiggers, C.R.M., Baak, M.L., Sonneveld, E., Nieuwenhuis, E.E.S., Bartels, M., and Creyghton, M.P. (2019). AML Subtype Is a Major Determinant of the Association between Prognostic Gene Expression Signatures and Their Clinical Significance. Cell Rep. *28*, 2866–2877.e5.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature *453*, 1239–1243.

World Health Organization (2001). Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues (IARC).

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

Wulf, M.G., Maguire, S., Humbert, P., Dai, N., Bei, Y., Nichols, N.M., Corrêa, I.R., and Guan, S. (2019). Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. Journal of Biological Chemistry *294*, 18220–18231.

Yeung, K.Y., and Ruzzo, W.L. (2001). Principal component analysis for clustering gene expression data. Bioinformatics *17*, 763–774.

Zajac, P., Islam, S., Hochgerner, H., Lönnerberg, P., and Linnarsson, S. (2013). Base Preferences in Non-Templated Nucleotide Incorporation by MMLV-Derived Reverse Transcriptases. PLoS ONE *8*, e85270.

Zhao, B.S., Roundtree, I.A., and He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. Nat. Rev. Mol. Cell Biol. *18*, 31–42.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

Zhou, J., and Chng, W.-J. (2017). Aberrant RNA splicing and mutations in spliceosome complex in acute myeloid leukemia. Stem Cell Investig *4*, 6.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell *65*, 631–643.e4.

Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I., and Enard, W. (2018). Quantitative single-cell transcriptomics. Brief. Funct. Genomics.

Abbreviations

allo-HSCT	allogeneic hematopoietic stem cell transplantation
AML/ALL	acute myeloid/lymphocytic leukemia
CFSE	carboxyfluorescein succinimidyl ester
DNA	deoxyribonucleic acid
mESC	mouse embryonic stem cell
FACS	fluorescence-activated cell sorting
FISH	fluorescence in-situ hybridization
HCA	human cell atlas
HSC	hematopoietic stem cell
IVT	in-vitro transcription
miRNA	micro RNA
MMLV	Moloney Murine Leukemia Virus
MRD	minimal residual disease
mRNA	messenger RNA
NGS	next generation sequencing
PCR	polymerase chain reaction
QC	quality control
qPCR	quantitative polymerase chain reaction
RNA	ribonucleic acid
rRNA	ribosomal RNA
RT	reverse transcription
SAGE	serial analysis of gene expression
SBS	sequencing by synthesis
scRNA-seq	single-cell RNA sequencing
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
TRM	treatment related mortality
UMI	unique molecular identifier

List of Figures

List of Publications

I. **Bagnoli, J.W.,** Ziegenhain, C., Janjic, A., Wange, L.E., Vieth, B., Parekh, S., Geuder, J., Hellmann, I., and Enard, W. (2018). Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. Nat. Commun. 9, 2937.

II. Ebinger, S., Zeller, C., Carlet, M., Senft, D., **Bagnoli, J.W.**, Liu, W.-H., Rothenberg-Thurley, M., Enard, W., Metzeler, K.H., Herold, T., et al. (2020). Plasticity in growth behavior of patients' acute myeloid leukemia stem cells growing in mice. Haematologica.

III. Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Álvarez-Varela, A., Batlle, E., Sagar, Grün, D., Lau, J.K., …, **Bagnoli, J.W.,** …, et al. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat. Biotechnol.

Other publications (not included in the thesis):

IV. Redondo Monte, E., Wilding, A., Leubolt, G., Kerbs, P., **Bagnoli, J.W.**, Hartmann, L., Hiddemann, W., Chen-Wichmann, L., Krebs, S., Blum, H., et al. (2020). ZBTB7A prevents RUNX1-RUNX1T1-dependent clonal expansion of human hematopoietic stem and progenitor cells. Oncogene.

V. **Bagnoli, J.W.**, Wange, L.E., Janjic, A., and Enard, W. (2019). Studying Cancer Heterogeneity by Single-Cell RNA Sequencing. In Lymphoma: Methods and Protocols, R. Küppers, ed. (New York, NY: Springer New York), pp. 305–319.

VI. Alterauge, D., **Bagnoli, J.W.**, Dahlström, F., Bradford, B.M., Mabbott, N.A., Buch, T., Enard, W., and Baumjohann, D. Continued Bcl6 expression is required for the maintenance of T follicular helper cell identity. Submitted Manuscript (in revision).

VII. Carlet, M., Voelse, K., Vergalli, J., Becker, M., Herold, T., Arner, A., Liu, W.-H., Dill, V., Fehse, B., Baldus, C.D., …, **Bagnoli, J.W.**, …, et al. (2020). In vivo inducible reverse genetics in patients' tumors to identify individual therapeutic targets. Submitted Manuscript. https://doi.org/10.1101/2020.05.02.073577

Declaration of Contribution as a co-author

Plasticity in Growth Behavior of Patients' Acute Myeloid Leukemia Stem Cells Growing in Mice.

This study was conceived and supervised by Irmela Jeremias. Patient cells were provided by Karsten Spiekermann. Financing and infrastructure was provided by Irmela Jeremias, Karsten Spiekermann and Wolfgang Enard. Patient derived xenografts were established by Sarah Ebinger and Binje Vick. Lentiviral transduction, CFSE labelling, Fluorescence activated cell sorting, and analysis of growth kinetics were performed by Sarah Ebinger with help from Christina Zeller. Michela Carlet and Wen-Hsin Liu provided material. Maja Rothenberg-Thurley and Klaus H. Metzeler performed and analyzed next generation targeted resequencing. I performed bulk RNA sequencing library preparations from cell lysates and primary sequencing data analysis. RNA-seq data was analysed by Tobias Herold and me .The manuscript was written by Daniela Senft, Sarah Ebinger, Binje Vick and Irmela Jeremias.
Figures were drawn by Daniela Senft and Sarah Ebinger.

Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas Projects.

This study was conceived and supervised by Holger Heyn. Catia Moutinho, Adrian Alvarez and Eduard Batlle prepared the reference sample. Elisabetta Mereu and Atefeh Lafzi performed all data analyses. Aleksandar Janjic, Lucas E. Wange and me performed scRNA sequencing library preparations primary sequencing data analysis for gmcSCRB-seq.
Holger Heyn, Elisabetta Mereu and Atefeh Lafzi wrote the manuscript with contributions from all co-authors.

According to the regulations for the Cumulative Doctoral Thesis at the Faculty of Biology, LMU München, I confirm the above contributions of Johannes Walter Bagnoli to these publications.

Date/Datum

Prof. Dr. Wolfgang Enard

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

Wolfgang Enard and Christoph Ziegenhain 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 Christoph Ziegenhain, Aleksandar Janjic, Lucas Wange and me. Sequencing data was processed by Swati Parekh and Christoph Ziegenhain. Christoph Ziegenhain, Aleksandar Janjic, Beate Vieth and I analyzed the data. Christoph Ziegenhain, 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

_____ _____

Date/Datum Prof. Dr. Wolfgang Enard

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 11.05.2020 Johannes Bagnoli

(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 11.05.2020 Johannes Bagnoli

(Unterschrift)

Acknowledgements

During my PhD, I was blessed to work with many people that shaped me and my research to a great extent. I would like to thank all of you and if I forgot anyone, I am sorry!

First and foremost, I want to thank my PhD advisor Wolfgang Enard. Wolfi, you are an exceptional person and an even more amazing boss. Thank you so much for always being there for me for the last 6 years! It has been such a pleasure to work with you and I will always remember that one cigarette in 2014 after your lecture ;)

Furthermore, I would like to especially thank the two people I consider my mentors, and whom I owe so much to, Dr. Christoph Ziegenhain and Daniel Richter. Your almost interdisciplinary help and guidance have helped and shaped me as a researcher as well as a person and I would not have come this far without you. I will always miss the two of you or already do!

Moreover, I would like to thank the "moisty" trinity: my two partners in RNA crime, Aleks Janjic and Lucas "the LEW" Wange, and the first and best minion ever, Johanna Geuder. Thank you guys, for not only being a tremendous help in the lab, but also really good friends over the years.

In short: thanks to everyone in the Enard lab for all your help:

Dr. Ines Hellmann, Ilse Valtierra, Dr. Beate Vieth, Dr. Swati Parekh, Zane Kliesmete and Philipp Janssen for your computational expertise, Karin Bauer for always getting the best price, Ines Bliesener for your great support and last but not least Ellen Zwick and Fiona Edenhofer for your great work.

I am very grateful to the members of the SFB1243 and the DFG for providing me with this great possibility to work in and contribute to the amazing field of AML. In addition I would like to thank Dr. Elizabeth Schroeder-Reiter and Elke Hammerbacher for their great work within the IRTG/SFB and for helping me get started.

Along with that, I would like to thank all may TAC members and collaborators in and outside the SFB for the great work: Prof. Dr. Irmela Jeremias, Dr. Binje Vick, Dr. Tobias Herold, Dr. Klaus Metzeler, Prof Dr. Christiane Fuchs, Enric Redondo-Monte, Dr. Marco Gerlinger and so many more……