inDrops-2: a flexible, versatile and cost-efficient droplet microfluidics approach for high-throughput scRNA-seq of fresh and preserved clinical samples

Simonas Juzenas[*,1], Vaidotas Kiseliovas[*,2,§], Karolis Goda[*,1], Justina Zvirblyte[1], Alvaro Quintinal-Villalonga[2] , Juozas Nainys[1,§] and Linas Mazutis[1,#]

[1] Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, 10257, Lithuania

[2] Computational and Systems Biology, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, USA

[3] Department of Medicine, Thoracic Oncology Service, Memorial Sloan Kettering Cancer Center, New York, USA

[§] Current address: Atrandi Biosciences, Vilnius, 10257, Lithuania

* Equal contribution

[#] Correspondence

Abstract

The development of a large variety of single-cell analytical methods has empowered researchers to explore diverse biological questions at the level of individual cells. Among these, droplet-based single-cell RNA sequencing (scRNA-seq) methods have been particularly prevalent owning to their high-throughput capabilities and reduced reaction volumes. While commercial systems have contributed to the widespread adoption of droplet-based scRNA-seq, the relatively high cost impose limitations for profiling large numbers of samples. Moreover, as the scope and scale of single cell sequencing methods keeps expanding, the possibility to accommodate diverse molecular biology workflows and inexpensively profile multiple biospecimens simultaneously, becomes highly relevant. Herein, we present inDrops-2: an open-source scRNA-seq platform designed to profile fresh or preserved clinical samples with a sensitivity matching that of state-of-the-art commercial systems, yet at the few folds lower cost. Using inDrops-2, we conducted a comparative analysis of two prominent scRNA-seq protocols – those based on exponential and linear amplification of cDNA – and provide new insights about the technical biases inherited to each approach. Finally, we applied inDrops-2 to simultaneously profile 18 lung carcinoma samples, all in one run following cell dehydration, long-term storage and multiplexing, to obtain a multiregional cellular profile of tumor microenvironment; thus, inDrops-2 offers researchers a possibility to perform large scale transcriptomics studies in a cost-effective manner.

Introduction

Comprehensive molecular characterization of biological samples increasingly relies on the accessibility to single-cell technologies [1]. Over the last few years a large array of platforms and methods for single-cell analysis have been introduced, thereby opening a new era of single-cell -omics [2]. In this venture, single-cell RNA-sequencing (scRNA-seq) techniques have been particularly impactful and have gained an increasing popularity. Rich biological information encoded in gene expression of individual cells that is captured by scRNA-seq methods have played a critical role, for example, in identifying new cells types in human body [3-5], delineating cancer heterogeneity [6-8] and patient response to therapy [9-11], and have advanced our understanding of various biological systems [12-16]. To this date scRNA-seq represents a leading technology for building cell atlases of human body and diseases [17-20] and is likely to remain such in a foreseeable future.

Arguably, among scRNA-seq platforms developed to-date [21-28], the most widely used are plate-based and droplet-based systems, each with unique strengths and weaknesses. The plate-based techniques provide an advantage for targeted applications, for instance, when cells of interest are isolated by FACS into microtiter plates for subsequent full-length scRNA-seq, or copy number variation (CNV) analysis [29-31]. These platforms often provide superior sensitivity, although at higher cost and reduced throughput. In contrast, droplet-based methods rely on 3' or 5' end RNA sequencing and offers a few orders of magnitude higher throughput as well as significantly reduced cost for cell barcoding and sequencing. Historically, two droplet-based techniques originally reported side-by-side in 2015 [21, 22] have paved the way for a high-throughput single-cell transcriptomics. Built on these innovations commercial systems such as 10x Chromium™ [25] (analogue to inDrops) and Nadia™ [32] (analogue to drop-seq) have followed, providing a broad accessibility to scRNA-seq technology.

Commercial systems ensure operational reproducibility and quality, making it a primary choice for single-cell transcriptomics studies. However, the high cost associated with cell barcoding and library preparation often restricts the broader use of the technology, especially among research groups with limited financial resources. On another hand, open-source systems such as inDrops [22] and drop-seq [21], or their modifications [33-35], offer lower operation costs, and can accommodate diverse needs of researchers such as processing unconventional samples [12, 36]. However, open-source systems often exhibit reduced sensitivity (e.g. transcript capture) when compared to commercial peers. For example, it has been reported that the commercial droplet-based systems (i.e. 10x Chromium) outperforms open-source

platforms such as inDrops and drop-seq in terms of UMI and gene capture by at least 2.5-fold [37].

Improving the sensitivity of scRNA-seq is particularly important for resolving subtle transcriptional differences within heterogeneous cell populations [38]. Furthermore, as the scale and complexity of single-cell transcriptomics studies keeps expanding, the need for barcoding and sequencing ever increasing numbers of cells can become challenging. Profiling single cells at $>10^5$ scale using commercially available droplet-based systems, while feasible, can lead to unsustainable financial burden. Open-source systems may provide cheaper alternatives, yet barcoding 10^5-10^6 individual cells requires extended encapsulation during which the cells of interest may undergo undesirable transcriptional changes. Therefore, in addition to high sensitivity, for scRNA-seq method to be broadly applicable it also needs to ensure the preservation of the native cell transcriptome state during the course of the workflow.

In this work we present inDrops-2, an open-source droplet microfluidics platform for performing high-throughput scRNA-seq experiments on live or fixed cells with the transcript and gene detection similar to that of state-of-the-art commercial platform (10x Chromium v3), yet at the 6-fold lower cost and a throughput of 5,000 cells min^{-1}. Using inDrops-2, we implemented two most frequently used scRNA-seq protocols; one based on linear amplification of copy DNA (cDNA) by *in vitro* transcription (IVT), and another based on exponential cDNA amplification by PCR following template switching (TS) reaction. Side-by-side comparison revealed that both scRNA-seq protocols are well-suited for profiling heterogeneous cell populations, yet they also carry important technical differences. The libraries constructed following linear amplification exhibited higher complexity, recovered higher number of genes, yet the protocol is labor intensive and takes 2 days to complete. The libraries constructed with TS-based approach are simpler to implement and display lower technical variability, yet the sequencing results revealed biased towards capturing shorter genes.

To further expand the applicability of inDrops-2, we report a cell preservation protocol for processing clinical samples comprising as little as 20,000 cells. We show that dissociated cells acquired from clinical biospecimens can be stored in a dehydrated state for extended periods of time and later multiplexed by covalently conjugating to DNA oligonucleotides [39] for subsequent transcriptomic analysis. As a proof-of-concept we performed multiplexing of

preserved lung adenocarcinoma samples and applied inDrops-2 to obtain a multiregional cellular profile of tumor microenvironment. We captured not only all major specialized lung epithelial and infiltrating stromal and immune cell phenotypes but also patient specific cell populations displaying clinically relevant phenotypes. In summary, we present inDrops-2, sensitive and cost-efficient method for capturing clinically relevant cell phenotypes from human biospecimens that underwent preservation, long-term storage and multiplexing.

Results

1. Optimized inDrops for improved transcript and gene detection

We started our study by reexamining the molecular workflow used in the original inDrops technique [22], with a goal to identify critical parameters that may improve transcript capture and detection. Using a microfluidics setup reported in the past [40] and further detailed in Figure S1A we encapsulated lymphoblast cells (K-562) in 1 nanoliter (nL) droplets at occupancy of ~0.3 together with barcoding hydrogel beads carrying photo-releasable RT primers comprising T7 promoter, cell barcode, unique molecular identifier (UMI) and poly(dT19) sequence (Figure S1B). We adjusted the flow rates of the microfluidics platform to achieve the high-throughput of 1 million droplets per hour, while maintaining high hydrogel bead loading at >85%, and stable droplet formation (Figure S1C). Under this regime, approximately 300,000 single cells can be encapsulated in less than an hour, with a low (~3%) doublet rate. We prepared scRNA-seq libraries following a workflow outlined in **Figure 1A**, while systematically examining each step in the protocol. Specifically, we profiled 1000-3000 cells per condition, by collecting approximately 10,000 droplets, photo-releasing RT primers from the hydrogel beads and performing reverse transcription to produce barcoded-cDNA, followed by second strand synthesis and linear amplification by *in vitro* transcription (IVT). The IVT libraries were fragmented with Lewis acid (Zn^{2+} ions), transcribed into single-stranded cDNA and PCR-amplified with sequencing adapters to obtain final gene libraries compatible with Illumina sequencers. After series of optimizations we arrived at inDrops-2 (IVT) protocol that showed markedly improved transcript and gene detection, and is provided as Supplementary Protocol 1. The most important findings can be summarized as follow. Compared to original inDrops [22], the barcoded-cDNA material purification and primer dimer removal by solid-phase reversible immobilization (SPRI), rather than digestion with nuclease cocktail, had the most significant impact (**Figure 1B**). At sequencing depth of 15,000 reads per cell, inDrops-2 showed increased UMI detection by 2.72-fold (mean ± s.d, 8166 ± 350 vs

2999 ± 461 UMIs, t-test, $P_{FDR} < 1 \times 10^{-300}$) and gene capture by 1.86-fold (mean ± s.d., 3399± 143 vs 1829± 207 genes, t-test, $P_{FDR} < 1 \times 10^{-300}$). Replacing Super Script III with Maxima H-minus reverse transcriptase increased UMI and gene capture further by 1.19-fold and 1.18-fold, respectively (t-test, UMI $P_{FDR} = 1.62 \times 10^{-281}$, gene $P_{FDR} = 1.58 \times 10^{-283}$). The improved UMI/gene detection using inDrops-2 was mirrored in primary cells (Figure S1E) and was not due to preference for a specific RNA biotype (Figure S1F), and overall displayed lower technical variability (**Figure 1C**). The inDrops-2 libraries prepared with Super Script IV (SS-IV) enzyme displayed slightly higher UMI count (SS-IV 9108 ± 198 vs Max 9108 ± 709, t-test, $P_{FDR} = 6.22 \times 10^{-317}$) **(Figure 1D)**, however, given significantly higher cost of the SS-IV enzyme, and relatively mild improvements, it was excluded from our subsequent exercises. The inDrops-2 conducted at 42 ºC tend to show slightly higher UMI/gene counts than corresponding reaction at 50 ºC, irrespectively of the RT enzyme tested (**Figure 1D**). No significant effect on UMI/gene recovery was observed when using different commercially available second strand synthesis or *in vitro* transcription kits, indicating that the critical steps for obtaining high UMI counts are indeed mainly related to reverse transcription reaction (e.g. RT enzyme) and subsequent purification of barcoded-cDNA molecules. In summary, single cell transcriptional profiling with inDrops-2 shows markedly improved UMI/gene detection, nearly 20-fold higher throughput (275 cells s^{-1} vs 15 cells s^{-1}), and higher quality data (**Figure 1E-G**). The step-by-step protocol incorporating aforementioned improvements is accompanying this manuscript as Supplementary Protocol 1.

2. inDrops-2 based on template switching reaction for rapid construction of sequencing libraries

While improved inDrops-2 based on linear amplification enables a substantially higher UMI and gene recovery per single cell, an alternative scRNA-seq strategy commonly used nowadays relies on a template switching (TS) reaction driven by the RT enzyme, and often referred as SMART [41-43]. Owing to the intrinsic terminal transferase activity, the RTase tends to add a few non-templated nucleotides (predominantly cytidines) on 3' end of cDNA, and does so preferably if the RNA template is G-capped [44-46]. A large variety of scRNA-seq methods have exploited this unique RT enzyme feature to incorporate the PCR adapters at 5' mRNA end, and facilitate library preparation for next generation sequencing [25, 29, 43, 47]. Motivated by these and other reports, we sought to adopt TS-based reaction with inDrops and to compare the UMI/gene capture yields to those obtained by linear amplification. To

differentiate the two scRNA-seq approaches, we refer to them as inDrops-2 (IVT) and inDrops-2 (TS).

At first, we aimed to maximize the yields of barcoded-cDNA following inDrops-2 (TS) approach (**Figure 2A)**. For that purpose, we encapsulated K-562 cells in 1 nL droplets together with barcoding hydrogel beads and RT/lysis reaction mix supplemented with template switching oligonucleotide (TSO), that is required for barcoded-cDNA amplification by PCR (see Materials and Methods). We thoroughly optimized each step of the workflow, namely; cell lysis and RT reaction, TSO concentration, temperature, barcoded-cDNA purification, library fragmentation, A-tailing and adapter ligation, and other parameters (Figure S2) to obtain a robust and reproducible scRNA-seq protocol that is described in details as Supplementary Protocol 2. Next, we evaluated the sensitivity of inDrops-2 (TS) by profiling human PBMCs (ATCC) and compared the results to that obtained with a commercial analogue (10X Genomics, Chromium v3 chemistry). Sequencing results presented in **Figure 2B** revealed that at the same sequencing coverage the inDrops-2 (TS) detects nearly the same UMI and gene count in single cells as the current gold-standard in the field. The cell types comprising the biospecimen that were identified with two techniques showed almost identical cell composition (**Figure 2C and 2D**). Altogether these results reassured us that inDrops-2 (TS) can serve as a cost-effective platform for profiling primary cells with high sensitivity and accuracy.

3. Comparison of scRNA-Seq protocols based on linear amplification vs. exponential amplification of cDNA

Having established inDrops-2 (TS), we then asked which scRNA-Seq approach, IVT-based or TS-based, can deliver higher number of unique transcripts and genes. While previous benchmarking studies indicated that scRNA-Seq libraries construction by linear amplification (e.g, CEL-Seq2) recovers higher diversity of genes as compared to PCR-based approaches such as SMART-seq2 [48], yet the head-to-head comparison of IVT-based or TS-based technique, to the best of our knowledge, is lacking. To perform such a comparison, we first formed an emulsion comprising approx. 5000 cells, compartmentalized along with barcoding hydrogel beads (v2) and 25 μM TSO. Upon completion the cDNA synthesis at 42 ºC for 90 min, the emulsion droplets were split into two equal fractions and processed separately following either inDrops-2 (IVT) or inDrops-2 (TS) protocol (**Figure 3A**). Following this strategy, we prepared and sequenced adenocarcinoma and lymphoblast cells.

To perform a comparative analysis between two protocols, we down-sampled sequencing depth of each sample to 20'000 raw reads per cell. We found that both approaches recover similar number of UMIs per cell (**Figure 3B**), however, a significant fraction of transcripts in TS-based libraries corresponded to the ribosomal protein (RP) genes (**Figure 3C**). Considering that in the cell transcriptomics studies the RP transcripts are typically filtered out from the downstream analyses, we removed RP genes. As a result, the differences between two protocols became sharper, with the libraries constructed by linear amplification revealing 25-30% higher UMI and gene detection, respectively (Figure S3A). Interestingly, the inDrops-2 (IVT) was also better at capturing unspliced RNAs as confirmed by higher fraction of reads aligning to introns as well as 3' UTRs, which play a role in post-transcriptional gene expression regulation (**Figure 3D**). As a drawback, the IVT-based approach exhibited slightly higher technical noise as well as higher run-to-run variability (Figure S3B). Additionally, as opposed to sharp enrichment at the 3' end, the gene body coverage in the TS-based approach was shifted, implying a trend towards shorter genes (**Figure 3E**). Indeed, a striking difference between two scRNA-Seq protocols became evident, when all detected genes were binned according to their length (**Figure 3F-G)**. This analysis clearly revealed the bias of template-switching based approach towards shorter genes; a trend that was reproduced in independent experiments on PBMCs as well as using 10X genomics (v3) platform (Figure S3B). Therefore, our results indicate that single cell transcriptome captured with TS-based, or IVT-based, approach is prone to specific technical biases that, when unaccounted, could negatively impact the interpretations of gene expression dynamics in individual cells, as shown by gene set enrichment analysis performed on differentially expressed genes between IVT- and TS-based inDrops-2 (**Figure 3H**). On another side, both approaches provide high confidence for identifying different cell types, making them suitable for cellular composition profiling and tissue atlases efforts (Figure S3C).

4. Cell preservation for long-term storage and transcriptomic studies of primary cells

Even with high-throughput capabilities offered by droplet microfluidics technology, the barcoding of increasingly high numbers of single cells (>10^5-10^6 scale) may require a sufficiently long (>60 min) encapsulation times during which there is an increased risk that live cells will alter their native transcriptional state. To mitigate such risk, it might it desirable to safeguard cellular transcriptome by fixing the cells so that no transcriptional changes would occur during encapsulation process. In this regard, cell dehydration and preservation in

methanol represents an appealing option as it was shown to retain transcriptional signature of cells, and to be compatible with droplet microfluidics methods [49-51]. Unfortunately, our attempts to adopt aforementioned cell preservation protocols were unsatisfying as we witnessed high variability of UMI / gene capture between the individual runs, and experienced a significant cell loss due to clumping (Figure S4A). Cell recovery was particularly problematic when handling clinical samples comprising low number ($n \leq 100{,}000$) of cells. We reasoned that excessive centrifugation force that is required to pellet the cells during rehydration might be causing cell clumping. Accordingly, after series of independent tests we found that methanol-fixed cells placed on cellulose-based filters (0.65 μm pore size) can be effectively rehydrated without applying an excessive centrifugation (see Methods). We confirmed that RNA integrity of cells in a rehydration buffer containing salts and citrate remains high (RIN > 8) during extended period of time (Figure S4B). Importantly, using rehydration columns cell clumping was significantly reduced thereby increasing the reproducibility of cell recovery.

We then performed inDrops-2 to compare UMI and gene capture of the methanol-preserved *vs* fresh human PBMCs (**Figure 4**). To avoid the RT reaction inhibition by citrate that is present in the rehydration buffer we adjusted the flow rates such that the final concentration of rehydration buffer in a droplet would correspond to 0.01X, or 1.5 mM sodium citrate (see Methods). In addition to benchmarking fresh and methanol-fixed human PBMCs we also tested bone marrow CD34 positive cells. Sequencing results presented in **Figure 4A** confirmed the efficient transcript recovery in fixed cells by inDrops-2, closely matching those of live PBMCs profiled alongside with commercial platform (10X Chromium V3). As expected, non-fixed cells displayed higher fraction of mitochondrial genes (**Figure 4B**), potentially indicating the undergoing transcriptional response during cell handling procedures. Average gene expression levels exhibited high correlation (**Figure 4C**), with gene mapping characteristics matching closely for both fresh and fixed samples (Figure S4C). Performing feature selection, dimensionality reduction and clustering for PBMCs revealed all expected cell populations, including CD4 T, CD8 T, NK, CD14 and CD16 monocytes, dendritic cells and megakaryocytes with similar cell proportions in fixed and fresh samples (**Figure 4D-E**). When using CD34+ samples we reconstructed transcriptional map of hematopoietic stem cell differentiation into known progenitor lineages (Figure S4D). In fresh and fixed samples all major lineages, such as common lymphoid progenitors, erythroid precursor cells, monocytes, dendritic cell precursors and eosinophils, B cell and Mast cell precursors, were present and marker gene expression was in excellent agreement between fixed/fresh samples (Figure S4E). As a final quality measure, we compared mapping statistics and found no significant

differences (Figure S4F). Overall, these results demonstrate that column-based rehydration of methanol-preserved cells ensures: i) minimal cell loss during handling, ii) high number of singlet recovery, iii) efficient UMI/gene detection, iv) accurate recapitulation of the transcriptional signature matching that of live cells, and iiv) alleviates the adverse effects of cell viability decline (i.e., increased mitochondrial gene expression), caused by extended workflows.

5. Scaling inDrops-2 with click chemistry hashtags

Sample multiplexing with hashtags provides an appealing option to increase the scale of scRNA-Seq experiments [39, 52-55]. Taking advantage of methanol-based cell preservation, we applied multiplexing by methyltetrazine-modified DNA oligonucleotides or 'ClickTags' [39]. The distinct feature of ClickTags is that they do not rely on specific cell epitopes for labelling and can be chemically attached to cellular proteins by Diels–Alder reaction, thus making them applicable to a broad range of cells and biospecimens. We sought to profile human lung tumor microenvironment, by conducting multiregional cell composition and gene expression analysis. The surgical samples acquired from lung carcinoma patients were cut into 3 parts, dissociated and FACS sorted into CD45 positive and CD45 negative compartments [56]. Following FACS, the single cell suspensions were preserved in methanol and transferred to -80 ºC for a long-term storage. After 4 weeks, the cells were retrieved and while in methanol hashed with ClickTags (see Methods). In total, 18 samples were hashed, then pooled, and following rehydration processed according to inDrops-2 (TS) workflow.

After sequencing, filtering and quality control step (see Methods) we obtained 32,937 high quality cells with rather consistent number of cells across hashtags (**Figure 5A**). The average UMI and gene count were high, 6959 and 1966, respectively (**Figure 5B** and **5C**) and matched closely the sequencing statistics of fresh LUAD samples (e.g. 5000 UMIs and 2100 genes, on average, at 40,000 reads per cell) [57]. As expected, there was a dependency of ClickTag and UMI counts per cell (**Figure 5D**). Following data normalization, feature selection, dimensionality reduction and visualization with uniform manifold approximation and projection (UMAP) cells were manually annotated using canonical gene markers (**Figure 5E-H**). UMAP colored by FACS-sorting label showed clear separation of CD45 positive and negative cells as expected (Figure 5G). We detected all major specialized lung epithelial and infiltrating stromal and immune cell phenotypes featuring lung adenocarcinoma disease (**Figure 5E, H**), and coinciding with previous reports [7, 58, 59]. Also, we observed high inter-patient variability

with regards to the cellular composition of the tumors (Figure 5F), while inter-regional differences within tumors were not as pronounced (Figure 5I). High resolution analysis in the non-immune cell compartment uncovered lung-specialized epithelial cells such as alveolar epithelial cells (AEC, markers *SFTPA1, HOPX*), club (*SCGB1A1, SCGB3A2*), ciliated (*CAPS, PIFO*), neuroendocrine (*CALCA, UCHL1*) and basal (*KRT17, KRT15*) cells, as well as patient specific *SPINK1*^{high} club cell population and *MMP7*^{high} alveolar epithelial phenotype [**Figure 5E, J**]. Recently, it has been reported that *SPINK1* upregulation leads to adverse outcomes in a multitude of cancers, and enhances proliferation and invasion of LUAD cells *in vitro* [60]. Interestingly, the *SPINK1*^{high} club cells (*SCGB3A1, SCGB3A2*) in our dataset also expressed distal lung marker *NAPSA* and *CEACAM6* (Figure S5), whereas the latter has been implicated in lung cancer progression and poor clinical outcomes [61]. Another interesting finding was two distinct phenotypes of alveolar epithelial cells – both of them expressing canonical AEC markers (i.e. *SFTPA1, SFTPA2, SFTPC*), yet only one population was marked by *MMP7*^{high} and *PRSS2*^{high} (**Figure 5J**). *MMP7* is a widely used biomarker for pulmonary fibrosis, while *PRSS2* is associated with invasive and metastasis promoting features [62]. The co-expression of *PRSS2* and *MMP7* in transitional state epithelial cells has been implicated in idiopathic pulmonary fibrosis [63], thus, our results suggest that *MMP7*^{high} alveolar epithelial cell phenotype might be involved in disease progression and plasticity.

Other non-immune cells in the LUAD atlas included lymphatic (*CCL21, NR2F2*) and tumor endothelial cells (*CLDN5, CLEC14A*), mesothelium (*MSLN, UPK3B*), smooth muscle cells (SMC, *ACTA2, TAGLN*) and diverse group of fibroblasts (*COL1A2, FN1*) which included two transcriptionally distinct groups involved in inflammation (**Figure 5E**). Complement-high fibroblasts had an unusually high expression of complement system constituents (i.e. *C7, C3, CFD*) indicative of inflammatory processes in the tumor microenvironment (Figure S5). Another fibroblast population was enriched for *HAS1* (**Figure 5J**), similarly to an invasive fibroblast population recently discovered in fibrotic lungs [64]. Moreover, this population upregulated a potent chemokine for monocytes *CCL2* and other inflammatory factors, such as cytokines *CXCL1, CXCL2* and *IL6*. Thus, we hypothesize that these rare inflammatory fibroblast populations in LUAD could be of interest for future in-depth investigation.

Consistent with previous LUAD atlases [7, 58, 59] the myeloid compartment comprised mast cells (*TPSB2, TPSAB1*), monocytes (*S100A9, FCN1*), conventional type 1 dendritic cells (*CLEC9A, CST3*), activated dendritic cells (*CCR7, CCL22*), monocyte-like dendritic cells (*CCL17, CLEC10A*), alveolar macrophages (*MARCO, FABP4*) as well as M1- and M2-like

subpopulations of tumor-associated macrophages (TAM). The lymphoid compartment consisted of innate lymphoid cells (ILC, expressed *CD3D*), B cells (*CD79A, MS4A1*), plasma cells (IGHG4, JCHAIN), plasmacytoid dendritic cells (*LILRA4, CLIC3*), NK cells (*NKG7, GZMB*) and a large group of diverse T cell phenotypes. Specifically, within the T cell population, we captured CD4 regulatory T cells (*FOXP3, CTLA4*), naïve CD4 T cells (*IL7R, CCR7*), effector memory CD8 T cells (*CD52, S100A4*), cytotoxic CD8 T cells (*GZMA, CCL4*) and CXCL13-high CD4 T cells (**Figure 5E**). Interestingly, the tumor sample comprising CXCL13-high T cell phenotype coincided with high count of B cells, therefore, supporting recent findings that CXCL13 acts as a potent attractant for B and other immune cells [65]. Moreover, the abundance of PDCD1-high CXCL13 producing CD8 T cells predicts response to PD-1 blockade therapy and correlates with increased overall survival in non-small cell lung cancer [66]. Overall, these findings clearly illustrate the power of inDrops-2 to recover clinically relevant cell phenotypes from biospecimens that underwent preservation, long-term storage and multiplexing, in an inexpensive and efficient manner.

Discussion

Advancements in high-throughput single-cell RNA-seq technologies [1] and computational methods [67] have opened new possibilities for investigating the gene expression programs and cellular composition in both normal and pathological conditions at unprecedented resolution and scale. As the range of scRNA-seq applications continues to expand across different domains of biomedical and biological sciences [16, 17, 68], there is a constant need for systems that not only deliver high throughput and sensitivity, but are also cost-effective. This is particularly relevant for analysis of biospecimens characterized by high heterogeneity, such as human tissues or cancer, necessitating the profiling of a large number of cells. Moreover, diverse needs of researchers often require versatile scRNA-seq platforms that can accommodate a broad range of samples and workflows.

Here we present an open-source scRNA-seq platform, inDrops-2, which enables high-throughput single cell transcriptomic studies with the transcript and gene detection matching that of state-of-the-art commercial platforms (i.e. 10X genomics Chromium v3), yet at the 6-fold lower operational cost. The system is highly flexible and customizable, providing a straightforward option to implement user-specific workflows. For instance, we implemented two popular scRNA-seq protocols; one based on linear amplification of cDNA using *in vitro* transcription and named as inDrops-2 (IVT), and another based on template switching reaction

followed by exponential cDNA amplification, or inDrops-2 (TS). While both techniques were found to be well suited for identifying different cells types and detected similar transcript count, yet they also showed important differences. The scRNA-seq workflow based on exponential cDNA amplification following TS-reaction shows lower technical variability and requires less labor to obtain sequencing results. However, due to limited efficiency of template switching reaction, the scRNA-seq libraries tend to be enriched in shorter (≤14 kb) genes, including ribosomal proteins. Given the median length size for protein-coding genes in humans to be 26 kb [69], the aforementioned technical bias might skew conclusions in certain biological contexts. For example, it has been reported that longer genes tend to be associated with cell development, complex diseases and cancer, while short genes are common to biological processes that require fast response such as immune system [70]. There are indications that transcript length also plays a role in aging [71]. Therefore, the scRNA-seq libraries constructed following linear amplification of cDNA by *in vitro* transcription, although being labor intensive and relying on advanced molecular biology skillset, exhibit higher complexity and capture more genes per single cell, including non-coding RNAs, thus making it a potentially better choice for studying gene expression at a whole genome level.

In addition to improving scRNA-seq performance on live cells, we implemented a cell preservation procedure that safeguards intracellular mRNA from degradation and alleviates the technical challenges of working with primary cells that are prone to uncontrollable transcriptional changes during handling. Importantly, while our procedure is loosely built on previous reports [49-51], yet in contrast to others we explore gentle cell rehydration process to ensure minimal cell clumping and loss, making it applicable to clinical samples of limited availability (n ≤ 50,000 cells). The data quality of rehydrated cells was high, and matched current standards in the field, with UMI/gene capture similar to live cells using commercial platforms (10X Chromium V3). Taking one step further we explored chemical hashing (indexing) of dehydrated cells based on Dies-Alder reaction [39]. As a proof-of-concept we performed multiregional profiling of lung carcinomas by hash-tagging 18 samples that were preserved after acquiring them from 3 patients. We could clearly differentiate the patients based on their disease profile (Figure S5) and identified all major specialized lung epithelial, infiltrating stromal and immune cell phenotypes [7, 58, 59]. We also identified cells with potential significance in immunotherapy, such as the CXCL13 producing CD4 T cells. These findings illustrate the broad applicability of inDrops-2 for future biomedical studies, where characterization of complex diseases at single-cell level are of high relevance. These results underscore the broad potential of inDrops-2 in biomedical research, especially in scenarios

where a detailed single-cell phenotypic characterization is highly important. Noteworthy, chemical cell hashing with DNA oligonucleotides not only minimized technical batch effects, but also benefits data analysis by facilitating unbiased removal of cell doublets all while preserving the clinically relevant cellular phenotypes.

Being an open and flexible platform inDrops-2 can accommodate diverse workflows matching user-specific needs and has a potential to further democratize single cell technologies. Indeed, it has recently been shown that inDrops-based platforms such as VASA-Seq [35] outperforms commercial and plate-based methods in terms of gene and transcript capture, while spinDrops [72] benefits applications based on target cell enrichment by on-chip sorting. Beyond transcriptomics, inDrops-based platforms have been successfully tailored to probe other -omic modalities in individual cells such as open chromatin by HyDrop [33] or genome by Microbe-Seq [73], to name a few. To conclude, inDrops-2 represents an affordable, highly sensitive, and adjustable open-source platform that should expand single cell -omic applications, and further enhance the scalability of experiments by providing a possibility to inexpensively profile multiple clinical samples by a high-throughput transcriptomic analysis.

ACKNOWLEDGEMENTS

This work received funding from European Regional Development Fund [01.2.2-LMT-K-718-04-0002] under grant agreement with the Research Council of Lithuania. Part of this work was also funded by The Alan and Sandra Gerry Metastasis and Tumor Ecosystems Center. S.J. was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 101030265. We are grateful to Ignas Masilionis for wet-lab assistance, and the members of the SCRI at the Sloan Kettering Institute for their valuable support and kind assistance.

AUTHOR CONTRIBUTIONS

SJ, KG, VK and JZ data analysis and interpretation; KG, VK and JN method development, single-cell RNA-seq experiments; AQV biospecimen acquisition, logistics and processing; L.M. study design, supervision and funding acquisition. LM wrote the initial draft of the manuscript; SJ, KG, JZ and LM revised the manuscript. All authors have read and approved the final manuscript.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

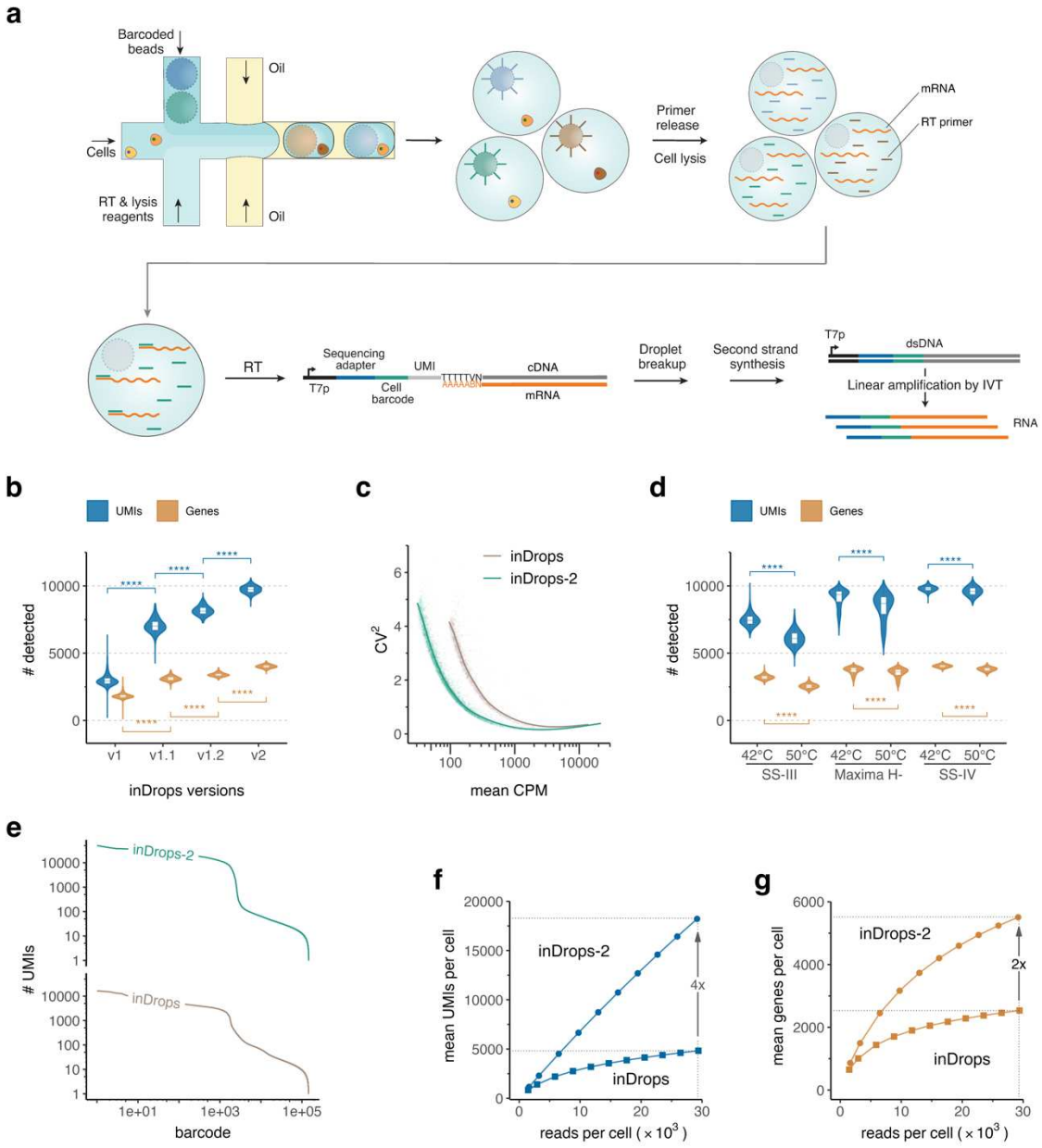

Figure 1. The overview of inDrops-2 performance. a) Schematics of inDrops-2 technique using linear amplification of barcoded cDNA. **b)** Detection of unique molecular identifiers (UMI) and genes in K-562 cell line, using different versions of inDrops. **c)** Comparison of technical variability between inDrops and inDrops-2. where CPM refers to counts per million. **d)** Evaluation of reverse transcription enzymes and temperature on the UMI and gene detection. **e)** Barcode rank plots derived from scRNA-seq data produced with inDrops and inDrops-2. **f)** Mean UMI count per cell as a function of sequencing depth. **g)** Mean gene count per cell as a function of sequencing depth.

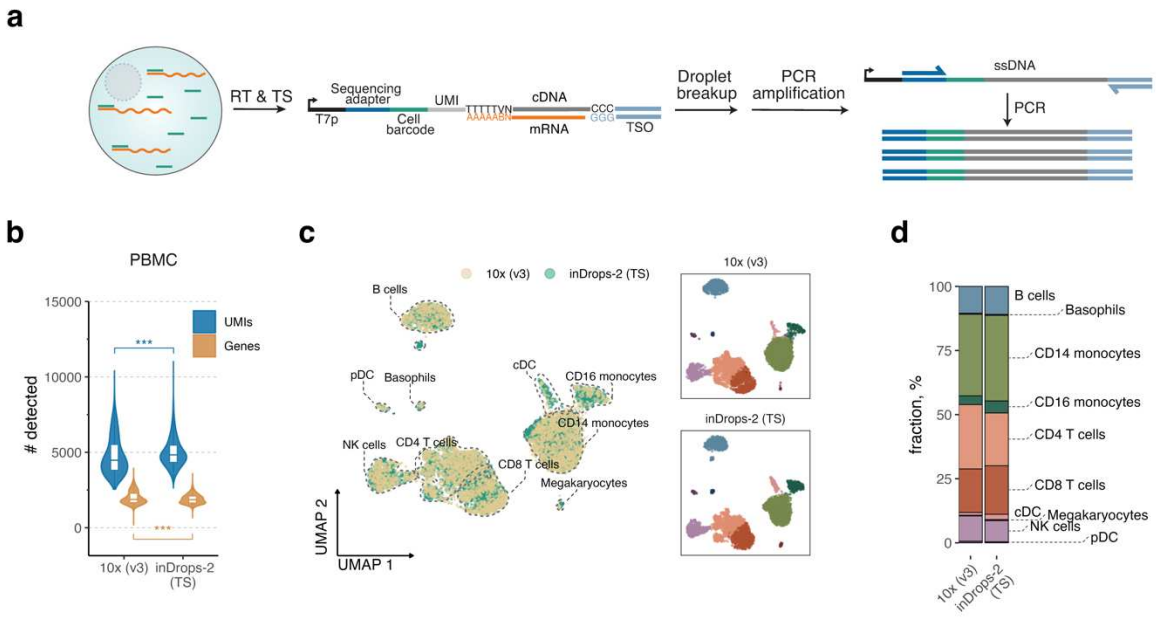

Figure 2. The overview of inDrops-2 using template switching approach. a) Schematics of inDrops-2 (TS) approach based on exponential cDNA amplification following template switching (TS) reaction. **b)** Comparison of unique molecular identifiers (UMI) and gene detection in PBMC cells at sequencing depth of 20'000 reads per cell. **c)** UMAP clustering based on individual cell types profiled with 10X Genomics (v3) and inDrops-2 (TS) platforms. **d)** Comparison of cell type fractions recovered with 10X Genomics (v3) and inDrops-2 (TS) platforms.

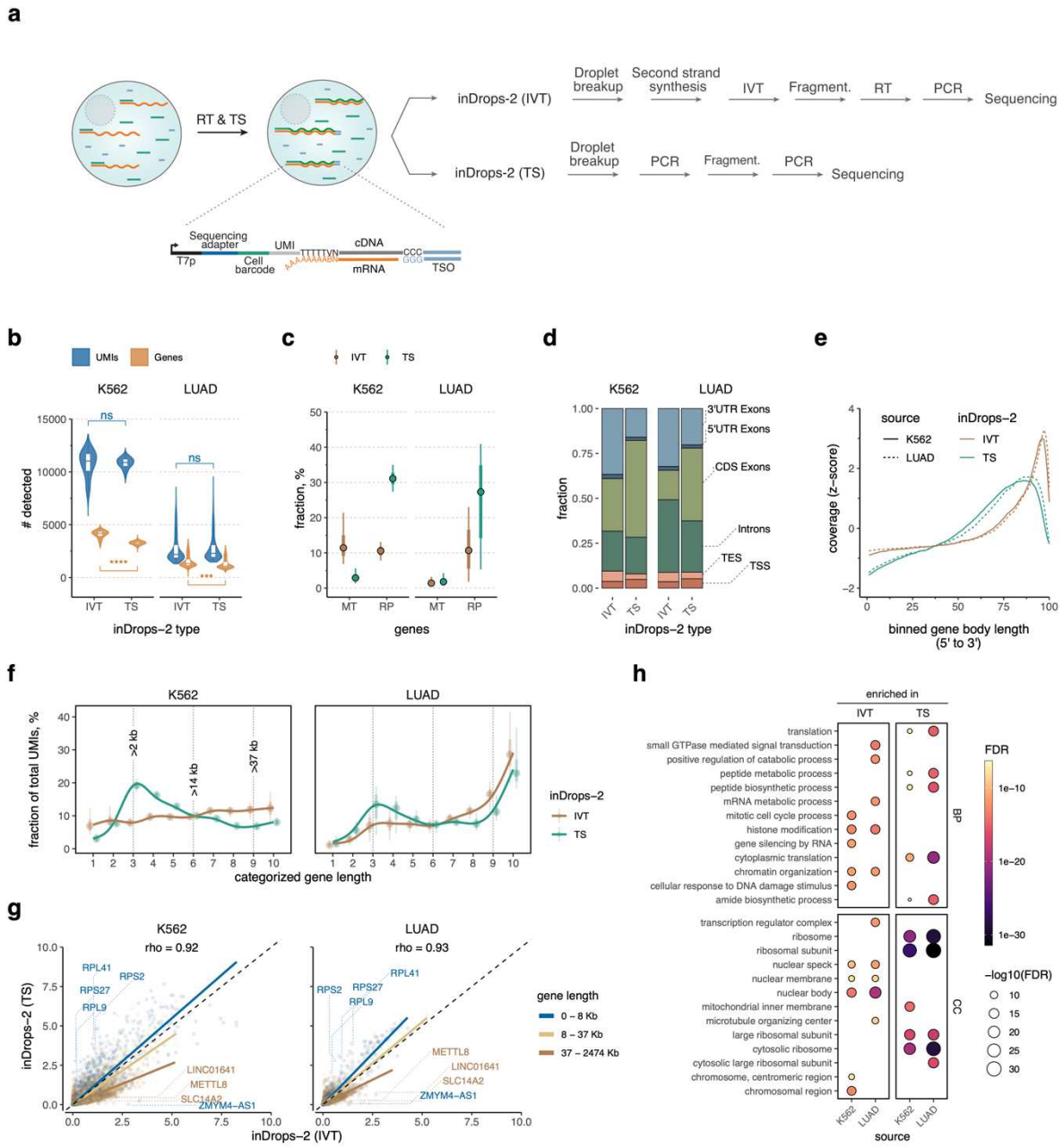

Figure 3. Comparison of scRNA-seq libraries prepared with linear and exponential amplification of cDNA. a) Schematics of the experiment. After RT reaction comprising template switching oligonucleotide emulsion droplets were split in two equal fractions and sequencing libraries prepared according to inDrop-2 (IVT) and inDrop-2 (TS) protocol. **b)** Number of UMIs and genes detected in lymphoblast (K-562) and lung adenocarcinoma (LUAD) cells in scRNA-seq libraries prepared by linear amplification of cDNA using *in vitro* transcription (IVT) reaction, and by exponential amplification of cDNA following template switching (TS) reaction. **c)** Fraction of genes corresponding to mitochondrial (MT) and ribosomal proteins (RP) in IVT-based and TS-based scRNA-seq libraries. **d)** Fraction of reads mapping to different regions of a gene. **e)** Sequencing coverage across the gene body. **f)** Fraction of UMIs as a function of binned gene length. **g)** Correlation analysis between inDrop-2 (IVT) and inDrop-2 (TS) protocols. Detection of transcripts encoded by longer genes (dark brown) is improved in IVT-based approach as compared to TS-based scRNA-seq libraries. Note, the median length of protein-coding gene in humans is 26 kb. **h)** Gene set enrichment analysis performed on differentially expressed genes between IVT-based and TS-based scRNA-seq libraries.

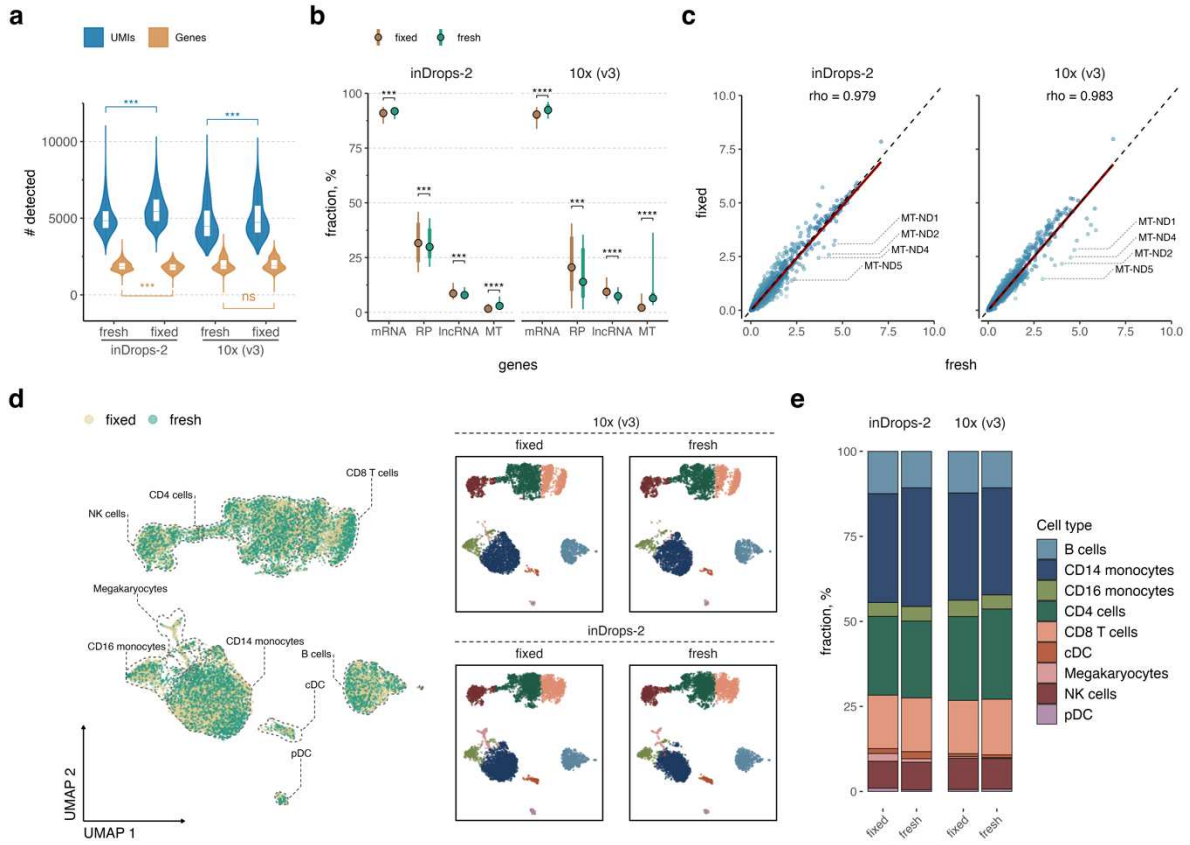

Figure 4. scRNA-seq of fresh and methanol-fixed PBMC cells using inDrops-2 and 10X Genomics (v3) platforms. a) UMI and gene detection in fresh and methanol-fixed PBMCs. **b)** Fraction of UMIs along RNA biotypes, **c)** correlation analysis and **d)** UMAP of fresh and methanol-fixed PBMCs. **e)** Cell types and their fractions recovered in fresh and fixed PBMC samples.

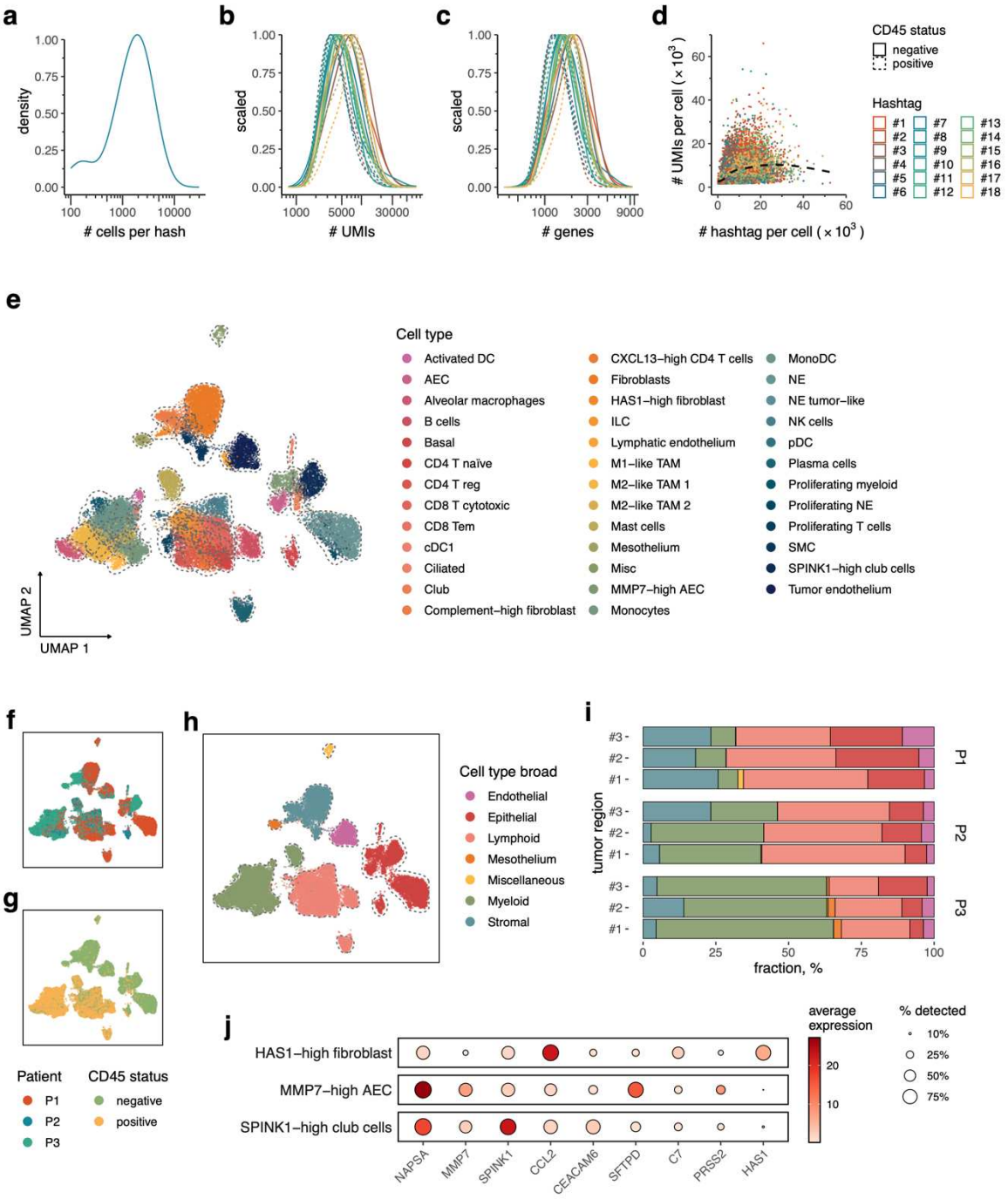

Figure 5. scRNA-seq of methanol-fixed and hashtag-indexed lung carcinoma cells. Probability distributions of **a)** cell number, **b)** total UMI count, and **c)** number of detected genes per individual hashtag. **d)** Relationship between UMI and hashtag counts per cell. **e)** An annotated UMAP of all lung carcinoma cells (n=32,937) displays high heterogeneity of cellular phenotypes in the tumor microenvironment. **f)** and **g)** UMAP colored by patient ID and CD45 status, respectively. **h)** UMAP representation colored by broad cell type categories. **i)** Sample composition analysis by broad cell type shows inter-patient compositional variability. **j)** Marker gene expression in selected cell populations. AEC – alveolar epithelial cells, cDC1 – conventional dendritic cells type I, ILC – innate lymphoid cells, NE – neuroendocrine cells, pDC – plasmacytoid dendritic cells, SMC – smooth muscle cells, TAM – tumor associated macrophages.

REFERENCES

1. Vandereyken, K., et al., *Methods and applications for single-cell and spatial multi-omics.* Nature Reviews Genetics, 2023. **24**(8): p. 494-515.
2. Zhu, C., S. Preissl, and B. Ren, *Single-cell multimodal omics: the power of many.* Nature Methods, 2020. **17**(1): p. 11-14.
3. Plasschaert, L.W., et al., *A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte.* Nature, 2018. **560**(7718): p. 377-381.
4. Villani, A.-C., et al., *Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors.* Science, 2017. **356**(6335).
5. Keren-Shaul, H., et al., *A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease.* Cell, 2017. **169**(7): p. 1276-1290.e17.
6. Azizi, E., et al., *Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment.* Cell, 2018. **174**(5): p. 1293-+.
7. Laughney, A.M., et al., *Regenerative lineages and immune-mediated pruning in lung cancer metastasis.* Nat Med, 2020. **26**(2): p. 259-269.
8. Nowicki-Osuch, K., et al., *Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition.* Science, 2021. **373**(6556): p. 760-767.
9. Sade-Feldman, M., et al., *Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma.* Cell, 2018. **175**(4): p. 998-1013.e20.
10. Jerby-Arnon, L., et al., *A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade.* Cell, 2018. **175**(4): p. 984-997.e24.
11. Xue, J., et al., *Rapid non-uniform adaptation to conformation-specific KRAS G12C inhibition.* Molecular Cancer Therapeutics, 2019. **18**(12).
12. Briggs, J.A., et al., *The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution.* Science, 2018. **360**(6392).
13. Zilionis, R., et al., *Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species.* Immunity, 2019. **50**(5): p. 1317-1334.e10.
14. *Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris.* Nature, 2018. **562**(7727): p. 367-372.
15. Liao, M., et al., *Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19.* Nature Medicine, 2020. **26**(6): p. 842-844.
16. Shaw, R., X. Tian, and J. Xu, *Single-Cell Transcriptome Analysis in Plants: Advances and Challenges.* Molecular Plant, 2021. **14**(1): p. 115-126.
17. Elmentaite, R., et al., *Single-cell atlases: shared and tissue-specific cell types across human organs.* Nature Reviews Genetics, 2022. **23**(7): p. 395-410.
18. Regev, A., et al., *The Human Cell Atlas.* eLife, 2017. **6**.
19. Rozenblatt-Rosen, O., et al., *The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution.* Cell, 2020. **181**(2): p. 236-249.
20. Jones, R.C., et al., *The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans.* Science, 2022. **376**(6594).
21. Macosko, E.Z., et al., *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.* Cell, 2015. **161**(5): p. 1202-14.

22. Klein, A.M., et al., *Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.* Cell, 2015. **161**(5): p. 1187-1201.
23. Picelli, S., et al., *Full-length RNA-seq from single cells using Smart-seq2.* Nature Protocols, 2014. **9**(1): p. 171-181.
24. Hashimshony, T., et al., *CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification.* Cell Reports, 2012. **2**(3): p. 666-673.
25. Zheng, G.X.Y., et al., *Massively parallel digital transcriptional profiling of single cells.* Nature Communications, 2017. **8**(1).
26. Gao, R., et al., *Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer.* Nature Communications, 2017. **8**(1).
27. Bose, S., et al., *Scalable microfluidics for single-cell RNA printing and sequencing.* Genome Biology, 2015. **16**(1).
28. Drake, R.S., et al., *Profiling Transcriptional Heterogeneity with Seq-Well S3: A Low-Cost, Portable, High-Fidelity Platform for Massively Parallel Single-Cell RNA-Seq*, in *Single Cell Transcriptomics*. 2023. p. 57-104.
29. Hagemann-Jensen, M., et al., *Single-cell RNA counting at allele and isoform resolution using Smart-seq3.* Nature Biotechnology, 2020. **38**(6): p. 708-714.
30. Haber, A.L., et al., *A single-cell survey of the small intestinal epithelium.* Nature, 2017. **551**(7680): p. 333-339.
31. Gao, R., et al., *Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes.* Nature Biotechnology, 2021. **39**(5): p. 599-608.
32. Davey, K., et al., *A flexible microfluidic system for single-cell transcriptome profiling elucidates phased transcriptional regulators of cell cycle.* Scientific Reports, 2021. **11**(1).
33. De Rop, F.V., et al., *Hydrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads.* eLife, 2022. **11**.
34. Habib, N., et al., *Massively parallel single-nucleus RNA-seq with DroNc-seq.* Nat Methods, 2017. **14**(10): p. 955-958.
35. Salmen, F., et al., *High-throughput total RNA sequencing in single cells using VASA-seq.* Nature Biotechnology, 2022. **40**(12): p. 1780-1793.
36. Karaiskos, N., et al., *The Drosophila embryo at single-cell transcriptome resolution.* Science, 2017. **358**(6360): p. 194-199.
37. Zhang, X., et al., *Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems.* Molecular Cell, 2019. **73**(1): p. 130-142.e5.
38. Svensson, V., et al., *Power analysis of single-cell RNA-sequencing experiments.* Nature Methods, 2017. **14**(4): p. 381-387.
39. Gehring, J., et al., *Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins.* Nature Biotechnology, 2019. **38**(1): p. 35-38.
40. Zilionis, R., et al., *Single-cell barcoding and sequencing using droplet microfluidics.* Nature Protocols, 2017. **12**(1).
41. Matz, M., et al., *Amplification of cDNA ends based on template-switching effect and step-out PCR.* Nucleic Acids Research, 1999. **27**(6): p. 1558-1560.
42. Zhu, Y.Y., et al., *Reverse Transcriptase Template Switching: A SMART™ Approach for Full-Length cDNA Library Construction.* BioTechniques, 2001. **30**(4): p. 892-897.
43. Ramsköld, D., et al., *Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells.* Nature Biotechnology, 2012. **30**(8): p. 777-782.

44. Schmidt, W., *CapSelect: a highly sensitive method for 5' CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs.* Nucleic Acids Research, 1999. **27**(21): p. 31e-31.
45. Tang, D.T.P., et al., *Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching.* Nucleic Acids Research, 2013. **41**(3): p. e44-e44.
46. Wulf, M.G., et al., *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, 2019. **294**(48): p. 18220-18231.
47. Hahaut, V., et al., *Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq.* Nature Biotechnology, 2022. **40**(10): p. 1447-1451.
48. Mereu, E., et al., *Benchmarking single-cell RNA-sequencing protocols for cell atlas projects.* Nature Biotechnology, 2020. **38**(6): p. 747-755.
49. Alles, J., et al., *Cell fixation and preservation for droplet-based single-cell transcriptomics.* BMC Biology, 2017. **15**(1).
50. Chen, J., et al., *PBMC fixation and processing for Chromium single-cell RNA sequencing.* Journal of Translational Medicine, 2018. **16**(1).
51. Katzenelenbogen, Y., et al., *Coupled scRNA-Seq and Intracellular Protein Activity Reveal an Immunosuppressive Role of TREM2 in Cancer.* Cell, 2020. **182**(4): p. 872-885.e19.
52. Peterson, V.M., et al., *Multiplexed quantification of proteins and transcripts in single cells.* Nat Biotechnol, 2017. **35**(10): p. 936-939.
53. Stoeckius, M., et al., *Simultaneous epitope and transcriptome measurement in single cells.* Nature Methods, 2017. **14**(9): p. 865-+.
54. Shahi, P., et al., *Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding.* Sci Rep, 2017. **7**: p. 44447.
55. McGinnis, C.S., et al., *MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices.* Nature Methods, 2019. **16**(7): p. 619-626.
56. Quintanal-Villalonga, Á., et al., *Protocol to dissociate, process, and analyze the human lung tissue using single-cell RNA-seq.* STAR Protocols, 2022. **3**(4).
57. Chan, J.M., et al., *Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer.* Cancer Cell, 2021. **39**(11): p. 1479-1496.e18.
58. Sinjab, A., et al., *Resolving the Spatial and Cellular Architecture of Lung Adenocarcinoma by Multiregion Single-Cell Sequencing.* Cancer Discov, 2021. **11**(10): p. 2506-2523.
59. Bischoff, P., et al., *Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma.* Oncogene, 2021. **40**(50): p. 6748-6758.
60. Xu, L., et al., *SPINK1 promotes cell growth and metastasis of lung adenocarcinoma and acts as a novel prognostic biomarker.* BMB Rep, 2018. **51**(12): p. 648-653.
61. Son, S.M., et al., *Therapeutic Effect of pHLIP-mediated CEACAM6 Gene Silencing in Lung Adenocarcinoma.* Sci Rep, 2019. **9**(1): p. 11607.
62. Sui, L., et al., *PRSS2 remodels the tumor microenvironment via repression of Tsp1 to stimulate tumor growth and progression.* Nat Commun, 2022. **13**(1): p. 7959.
63. Kobayashi, Y., et al., *Persistence of a regeneration-associated, transitional alveolar epithelial cell state in pulmonary fibrosis.* Nat Cell Biol, 2020. **22**(8): p. 934-946.

64. Habermann, A.C., et al., *Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis.* Sci Adv, 2020. **6**(28): p. eaba1972.
65. Ukita, M., et al., *CXCL13-producing CD4+ T cells accumulate in the early phase of tertiary lymphoid structures in ovarian cancer.* JCI Insight, 2022. **7**(12).
66. Thommen, D.S., et al., *A transcriptionally and functionally distinct PD-1(+) CD8(+) T cell pool with predictive potential in non-small-cell lung cancer treated with PD-1 blockade.* Nat Med, 2018. **24**(7): p. 994-1004.
67. Stuart, T. and R. Satija, *Integrative single-cell analysis.* Nature Reviews Genetics, 2019. **20**(5): p. 257-272.
68. Stubbington, M.J.T., et al., *Single-cell transcriptomics to explore the immune system in health and disease.* Science, 2017. **358**(6359): p. 58-63.
69. Piovesan, A., et al., *GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics.* Database, 2016. **2016**.
70. Lopes, I., et al., *Gene Size Matters: An Analysis of Gene Length in the Human Genome.* Frontiers in Genetics, 2021. **12**.
71. Stoeger, T., et al., *Aging is associated with a systemic length-associated transcriptome imbalance.* Nature Aging, 2022. **2**(12): p. 1191-1206.
72. De Jonghe, J., et al., *spinDrop: a droplet microfluidic platform to maximise single-cell sequencing information content.* Nature Communications, 2023. **14**(1).
73. Zheng, W., et al., *High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome.* Science, 2022. **376**(6597).