TAS-Seq: a robust and sensitive amplification method for beads-based scRNA-seq

Shigeyuki Shichino[1], Satoshi Ueha[1], Shinichi Hashimoto[2], Tatsuro Ogawa[1], Hiroyasu Aoki[1], Wu Bin[1], Chang-Yu Chen[1], Masahiro Kitabatake[3], Noriko Ouji-Sageshima[3], Shigeto Hontsu[4], Noriyoshi Sawabata[5], Takeshi Kawaguchi[5], Toshitugu Okayama[6], Eiji Sugihara[7,8], Toshihiro Ito[3], Kazuho Ikeo[6], Taka-aki Sato[7] and Kouji Matsushima[1]

[1]Division of Molecular Regulation of Inflammatory and Immune Diseases, Research Institute of Biomedical Sciences, Tokyo University of Science, Chiba, Japan
[2]Department of Molecular Pathophysiology, Institute of Advanced Medicine, Wakayama Medical University, Wakayama, Japan
[3]Department of Immunology, Nara Medical University, Kashihara, Nara, Japan.
[4]Department of Respiratory Medicine, Nara Medical University, Kashihara, Nara, Japan.
[5]Department of Thoracic and Cardio-Vascular Surgery, Nara Medical University, Kashihara, Nara, Japan.
[6]National Institute of Genetics, Shizuoka, Japan
[7]Research and Development Center for Precision Medicine, University of Tsukuba, Ibaragi, Japan
[8]Center for Joint Research Facilities Support, Research Promotion and Support Headquarters, Fujita Health University, Toyoake, Aichi, Japan

*Correspondence to: Kouji Matsushima
Division of Molecular Regulation of Inflammatory and Immune Diseases, Research Institute of Biomedical Sciences, Tokyo University of Science,
Building 17 second floor, 2641, Yamasaki, Noda, Chiba, JAPAN
Phone: +81-4-7121-4114, FAX +81-4-7121-4114, Zip : 278-0042
E-mail: koujim@rs.tus.ac.jp

Abstract

Single-cell RNA-sequencing (scRNA-seq) is valuable for analyzing cellular heterogeneity. Cell composition accuracy is critical for analyzing cell-cell interaction networks from scRNA-seq data. We developed terminator-assisted solid-phase cDNA amplification and sequencing (TAS-Seq), a scRNA-seq method relying on a terminator, terminal transferase, and nanowell/beads-based scRNA-seq platform that could acquire scRNA-seq data, is highly correlated with flow-cytometric data, has gene-detection sensitivity, and is more robust than

widely-used methods.

Main

Single-cell RNA sequencing (scRNA-seq) has been deciphering cellular subsets in various species, organs, and conditions in an unsupervised manner and drives the construction of single-cell atlas, such as Human Cell Atlas[1]. The primary output of scRNA-seq data in an analyzed sample is the gene-expression pattern of single cells, their classification by gene-expression similarity, and their cellular composition. Cellular composition, i.e., the abundance of transcriptionally distinct cell subsets, is a significant factor reflecting the functions of the analyzed sample; hence, the accuracy of scRNA-seq data cellular composition is essential to elucidate biological issues and build single-cell atlas using scRNA-seq datasets.

10X Genomics Chromium, a microdroplet-based high-throughput scRNA-seq platform, is widely used because it is user-friendly and commercially available[2]. Another major microplate/cell sorter-based scRNA-seq platform is Smart-seq2[3], often combined with a microdroplet-based system to achieve more gene-level sensitivity for every single cell[4]. However, both techniques have cell sampling bias that could affect the cell composition of scRNA-seq datasets. For example, human neutrophils dropout frequently occurs in 10X Chromium system[5], and fragile cells, such as macrophages and some types of stromal cells, tend to be lost during cell sorting because of high-pressure[5,6]. In addition, most high-throughput scRNA-seq methods use template-switching reaction[2] for cDNA amplification. Thus, efficiency is affected by the 5′ structure of RNA[7], limiting the capability of analyzable RNA specimens in scRNA-seq analysis.

Terminal transferase (TdT) is a template-independent polymerase that could efficiently add homopolymer tails against 3′ ends of DNA. TdT-based scRNA-seq methods rely on the homopolymer tailing reaction and capture RNA specimens more uniformly than template-switching methods. Indeed, it has been previously demonstrated that the TdT-based scRNA-seq method Quartz-seq/Quartz-seq2 has high gene-detection sensitivity[8,9]. However, stringent control of the TdT reaction, including controlling reaction time on the second scale and/or primer density on the cDNA-immobilized magnetic beads, is necessary to avoid excessive primer-derived bi-product synthesis and cDNA amplification failure[8-10]. This property leads to difficulties in handling TdT-based scRNA-seq methods.

To overcome these problems, we developed terminator-assisted solid-phase cDNA amplification and sequencing, termed TAS-Seq, a novel TdT-based cDNA amplification method for nanowell/beads-based scRNA-seq methods (Fig. 1a). A nanowell-based system could isolate single cells gently by gravity flow[2] and possibly capture cell composition more precisely. We used BD Rhapsody as a nanowell/beads-based scRNA-seq system because of

its commercial availability.

Stochastic termination of tailing reaction using dideoxycytidine triphosphate (ddCTP): deoxycytidine triphosphate (dCTP) mixture (1:20) was applied to increase the robustness of the TdT reaction. We used potassium cacodylate buffer supplemented with Co^{2+} ion, previously reported as the most efficient buffer system of the TdT tailing reaction[11,12]. On Exonuclease I-treated magnetic beads of BD Rhapsody, we found that ddCTP:dCTP (1:20) effectively suppressed undigested primer-derived products extension (under 200bp) for 5 or 30 minutes of TdT reaction. However, the extensive extension of undigested primer-derived products occurred when ddCTP was not added. In addition, on cDNA-immobilized, Exonuclease I-treated BD Rhapsody beads, ddCTP addition also effectively suppressed undigested primer-derived products extension at ranges up to 45 minutes of TdT reaction with visible cDNA products (Fig. 1c). Using 4000 single cells of the murine lung, TAS-Seq yielded over 1 ug of amplified cDNA with typical size distribution (peaked around 1kbp) by 16 cycles of PCR (Fig. 1d). These results indicated that TAS-Seq could amplify cDNA effectively with well-tolerated TdT reaction time and TdT activity, which might be affected by the lot-to-lot variability of the TdT enzyme.

Because cell hashing by short oligonucleotide-conjugated antibodies is widely used to reduce scRNA-seq cost, we further examined whether TAS-Seq was compatible with the cell hashing method. We pooled 14 samples of BioLegend Hashtag-A labeled $CD45.2^{+}$ cells from a murine subcutaneous tumor model of Lewis lung carcinoma and subjected them to TAS-Seq as previously described[13] (Fig. 1a). TAS-Seq successfully obtained cDNA and hashtag libraries (Fig. 1e). The demultiplexing of 14 libraries by hashtag readout revealed that cell number and genes were detected similarly among 14 samples (Fig. 1f), suggesting the compatibility of TAS-seq with cell hashing technology.

To evaluate the performance of TAS-Seq, we first compared TAS-Seq with a commercial whole-transcriptome amplification (WTA) BD Rhapsody kit, a random priming-based cDNA amplification, using mouse spleen cells (Supplementary Fig. 1a). We found that TAS-Seq detected more genes (3026 genes/50000 reads) than BD WTA kit (1997 genes/50000 reads) and showed similar quality-control metrics of scRNA-seq data (Supplementary Fig. 1b and 1c), suggesting that TAS-Seq could detect more genes than random priming-based approach.

Next, we compared scRNA-seq data of single-cell suspension of adult murine lungs obtained by TAS-Seq with Smart-seq2/10X Chromium v2 data from Tabula Muris Consortium[4] and 10X Chromium v3 data[14] (Fig. 2a). TAS-Seq detected more genes than the other datasets (Fig. 2b). Compared to Smart-seq2 data, TAS-Seq could detect more gene numbers and highly-variable genes (detected by Seurat v2.3.4[15] package) than Smart-seq2

data even in approximately 1/10 sequencing depth (Fig. 2c-2e). Clustering analysis of each dataset by Seurat package revealed that distinct cell subsets of adult murine lung clearly clustered together in all datasets. However, their cell compositions were different (Fig. 2f). Of note, alveolar macrophages were lost in Smart-seq2 data, possibly by cell sorting damage (Fig. 2f).

To evaluate the accuracy of quantification of cell composition of the adult murine lung by each scRNA-seq platform, we compared cell composition obtained by scRNA-seq and flow-cytometric analysis (Supplementary Fig. 2a-2c). We found that TAS-Seq showed highest Pearson's correlation coefficient ($R^2 = 0.962$, $p = 2.45 \times 10^{-11}$), followed by Smart-seq2 ($R^2 = 0.856$, $p = 2.94 \times 10^{-7}$), 10X Chromium v3 ($R^2 = 0.758$, $p = 1.15 \times 10^{-5}$), and 10X Chromium v2 ($R^2 = 0.274$, $p = 3.75 \times 10^{-2}$) (Fig. 2g). We also found that Smart-seq2 and 10X Chromium v2 data over-represented the frequency of endothelial cells and fibroblasts/natural killer cells, respectively (Fig. 2g). In addition, 10X Chromium v2 and v3 data under-represented the frequency of epithelial/endothelial cells and epithelial cells/fibroblasts, respectively (Fig. 2g). Because gene-detection rate was different between cell subsets, we further compared detected gene numbers between TAS-Seq and Smart-seq2 data within each cell subset. TAS-Seq significantly detected more genes than Smart-seq2 within most cell subsets except pericytes and smooth muscle cells (Supplementary Fig. 3).

We further analyzed human lung samples of fibrotic and non-fibrotic areas of a rheumatoid arthritis-associated interstitial lung disease (RA-ILD) patient by TAS-Seq. Cell clustering analysis revealed that TAS-Seq captured the difference of cell composition between fibrotic and non-fibrotic areas from the same patient with minimal batch-effect (Supplementary Fig. 4). Strikingly, TAS-Seq obtained scRNA-seq data highly correlated with flow-cytometric data in RA-ILD samples and precisely detected neutrophils depleted in 10X Chromium v2 dataset of human lungs[5] (Fig. 2h-2j, Supplementary Fig. 5a and 5b). These data indicated that TAS-Seq could capture cell composition of adult murine and human lungs more precisely than Smart-seq2 and 10X Chromium with high gene-detection sensitivity.

Cell-cell interaction network analysis is a major downstream analysis of scRNA-seq data and is possibly affected by the cell composition accuracy of scRNA-seq datasets. Using CellChat software[16] that considers the abundance of cell subsets, we inferred cell-cell interactions of adult murine lungs using TAS-Seq, Smart-seq2, 10X Chromium v2/v3 datasets, of which total cell number was downsampled to 1732 cells (the cell number of Smart-seq2 dataset). We found that the number of inferred interactions and pathways was highest in TAS-Seq data, throughout from soft to hard thresholds of ligand/receptor genes within cell subsets (minimum percent of expressed cells in each cell subset) (Fig. 2k). Of note, some of the important pathways for lung development, homeostasis, and repair, including sonic hedgehog,

WNTs, bone morphologic proteins (BMPs), fibroblast growth factor (FGFs), and Notch signaling[17], were lost in 10X datasets when the expression threshold became more stringent (Supplementary Fig. 6a), suggesting that TAS-Seq could detect important cell-cell interaction pathways more robustly than 10X platforms when combined with CellChat analysis. In addition, CellChat predicted that alveolar type 2 epithelial cells (AT2 cells) were the major producer/receiver within inferred cell-cell interaction network from TAS-Seq and Smart-seq2 datasets, but not in 10X datasets (Supplementary Fig. 6b). Both vascular endothelial cells and Inmt^{hi} alveolar fibroblasts[18] were also predicted as the other major contributors in TAS-Seq and Smart-seq2 datasets, but one of them was lost in 10X datasets (Supplementary Fig. 6b). Moreover, AT2 cells, Inmt^{hi} alveolar fibroblasts, and vascular endothelial cells were connected stronger within the CellChat-predicted cell-cell interaction network of the TAS-Seq dataset than the other datasets (Fig. 2l and Supplementary Fig. 6c). Because AT2 cell-alveolar fibroblast interaction is thought to be crucial for alveolar homeostasis, repair, and regeneration[19], TAS-Seq is possibly more useful for identifying important intercellular communication of murine lung than Smart-seq2 and 10X Chromium.

Overall, TAS-Seq might be more easy-to-handle than existing TdT-based scRNA-seq methods and might provide high-resolution scRNA-seq data with better accuracy of cell composition and inference of cell-cell interaction network than template-switching-based Smart-seq2 and 10X Chromium. In principle, TAS-Seq could be applied against the other solid-phase-based scRNA-seq and spatial transcriptomics platforms where the captured RNA tends to be degraded, such as 10X Visium™ and HDST[20]. Expanding TAS-Seq application is possibly helpful for better understanding and atlas construction of various biological contexts at the single-cell level.

Methods

Mice. C57BL/6J female mice were purchased from Sankyo Labo Service Corporation (Ibaragi, Japan). All mice were bred at specific pathogen-free facilities at Tokyo University of Science and were 8 weeks old (for lung sample) or 10 weeks old (for spleen sample) at the commencement of experiments.

RNA extraction. NIH/3T3 cells were cultured with DMEM high glucose (Nacalai Tesque, Kyoto, Japan) supplemented with 10% FBS (Cat#2916546, Lot#1608A, MP Bio Japan, Tokyo, Japan) and 10 mM HEPES (Cytiva (Global Life Sciences Technologies Japan), Tokyo, Japan) (DMEM/10%FBS/HEPES), and stored at -80°C with CellBanker 1 (Zenoaq Resource, Fukushima, Japan). Stored cells were thawed and cultured with DMEM/10%FBS/HEPES, and 80% of confluent cells were recovered. Total RNA was extracted from resultant cells using TRIZol Reagent (Thermo Fisher Scientific, Tokyo, Japan) according to the manufacture's instructions. The extracted RNA was dissolved with nuclease-free water (Nacalai Tesque) and stored at -80°C.

Preparation of cDNA-immobilized or un-immobilized BD Rhapsody beads for evaluation of TAS-Seq. BD Rhapsody magnetic beads (BD Biosciences, San Jose, CA, USA) for bulk experiments were collected as follows. Un-trapped BD Rhapsody beads after being loaded onto BD Rhapsody cartridge were collected, washed twice with WTA wash buffer [10 mM Tris-HCl pH 8.0 (Nippon Gene, Tokyo, Japan), 50 mM NaCl (Merck, Tokyo, Japan), 1 mM EDTA (Nippon Gene, Tokyo, Japan), and 0.05% Tween-20 (Merck)], were resuspended with 200 μL of Beads resuspension buffer (BD Biosciences) and stored at 4°C. After removing the supernatant, 1 μg of total RNA from NIH/3T3 cells was diluted with 500 μL of BD Rhapsody Lysis buffer (BD Biosciences) supplemented with 15 mM DTT (BD Biosciences) and added to the beads. Beads were resuspended and incubated for 30 minutes at room temperature (RT) with gentle rotation. Beads were washed once with 500 μL of BD Rhapsody lysis buffer, once with 1 ml of wash buffer B [10 mM Tris-HCl (pH 7.5) (Nippon Gene), 150 mM LiCl (Merck), 1 mM EDTA, and 0.02% Tween-20], and twice with 500 μL of wash buffer B. During the washing step, beads-containing DNA LoBind tubes (Eppendorf Japan, Tokyo, Japan) were replaced twice. After removing the supernatant, reverse transcription was performed for 20 minutes at 37°C using a BD Rhapsody cDNA kit following the manufacture's instruction. After removing the supernatant, Exonuclease I mix (20 μL of 10X Exonuclease I buffer (New England Biolabs, Ipswich, MA, USA), 10 μL of Exonuclease I (New England Biolabs), and 170 μL of nuclease-free water (Nacalai Tesque) in 200 μL reaction) was directly added to the beads and further incubated for 60 minutes at 37°C with

1,200 rpm on a Thermomixer C with Thermotop (Eppendorf Japan). Resultant beads were immediately chilled on ice, the supernatant was removed, and washed with 1 mL of WTA wash buffer, 200 μL of BD Rhapsody lysis buffer (for inactivation of enzyme), once with 1 ml of WTA wash buffer, and twice with 500 μL of WTA wash buffer, resuspended with 200 μL of Beads resuspension buffer (BD Biosciences) and stored at 4℃. During the washing step, beads-containing DNA LoBind tubes (Eppendorf Japan, Tokyo, Japan) were replaced twice. For producing cDNA un-immobilized BD Rhapsody beads, un-trapped BD Rhapsody beads were purchased as above. After removing the supernatant, beads were treated with Exonuclease I mix (20 μL of 10X Exonuclease I buffer, 10 μL of Exonuclease I, and 170 μL of nuclease-free water in 200 μL reaction) for 60 minutes at 37℃ with 1,200 rpm on a Thermomixer C with Thermotop. Resultant beads were immediately chilled on ice; the supernatant was removed and washed with 1 mL of WTA wash buffer, 200 μL of BD Rhapsody lysis buffer (for inactivation of enzyme), once with 1 ml of WTA wash buffer, twice with 500 μL of WTA wash buffer, resuspended with 200 μL of Beads resuspension buffer and stored at 4℃. During the washing step, beads-containing DNA LoBind tubes were replaced twice. For washing BD Rhapsody beads, BD IMagnet Cell Separation Magnet (BD Biosciences) and Dynamag-2 (Thermo Fisher Scientific) were used for collecting BD Rhapsody beads.

Evaluation of terminator-assisted homopolymer tailing reaction and DNA amplification from BD Rhapsody beads. For cDNA un-immobilized beads, beads were split into seven parts, transferred into 1.5 ml DNA LoBind tubes, and subjected to homopolymer tailing reaction by terminal transferase (TdT). After removing the supernatant and washing once with nuclease-free water, the beads were mixed with TdT mixture 1 [1×TdT buffer (Thermo Fisher Scientific), 1.2 mM deoxycytidine triphosphate (dCTP, Roche Diagnostics, Tokyo, Japan), 0.06 mM dideoxycytidine triphosphate (ddCTP, Cytiva), 15 U/μL TdT (Roche), 0.1 U/μL RNase H (QIAGEN, Düsseldorf, Germany)], TdT mixture 2 [1×TdT buffer, 1.2 mM dCTP, 0.06 mM ddCTP, 10 U/μL TdT, 0.1 U/μL RNase H], TdT mixture 3 [1×TdT buffer, 1.2 mM dCTP, 0.06 mM ddCTP, 42 U/μL TdT, 0.1 U/μL RNase H], TdT mixture 4 [1×TdT buffer, 1.2 mM dCTP, 15 U/μL TdT, 0.1 U/μL RNase H], and no TdT control mixture [1×TdT buffer, 1.2 mM dCTP, 0.06 mM ddCTP, 0.1 U/μL RNase H]. TdT reactions were performed using 100 μL/tubes for 5 or 30 minutes (TdT mixture 1 and 4) and for 30 minutes (TdT mixture 2, 3, and no TdT control mixture) at 37℃ with 1,200 rpm on a Thermomixer C with Thermotop. For cDNA immobilized beads, beads were split into four parts, transferred into 1.5 ml DNA LoBind tubes. After removing the supernatant and washing once with nuclease-free water, the three parts of the beads were mixed with the TdT mixture 1, and one

was mixed with the no TdT control mixture. Then, beads were incubated for 15, 30, and 45 minutes (TdT mixture 1) and 45 minutes (no TdT control mixture) at 37℃ with 1,200 rpm on a Thermomixer C with Thermotop. Reactions were chilled on ice immediately after the reaction was completed. After the supernatant was removed, beads were washed with 1 mL of WTA wash buffer, 200 μL of BD Rhapsody lysis buffer (for inactivation of enzyme), once with 1 ml of WTA wash buffer, twice with 500 μL of WTA wash buffer, and resuspended with 100 μL of 10 mM Tris-HCl pH8.0. During the washing step, beads-containing DNA LoBind tubes were replaced twice. Beads were transferred into new 8-strip tubes, the supernatant was discarded, and 12.5 μL of second-strand synthesis mixture [1× KAPA Hifi ReadyMix (KAPA Biosystems, Wilmington, MA, USA) and 0.4 μM 5′universal-9G primer] was added, and second-strand synthesis was performed according to the following program: 95°C for 3 min, 98°C for 20 s, 47°C for 2 min, 72°C for 7 min, and hold at 4°C. Then, 37.5 μL of amplification mix [1× KAPA Hifi ReadyMix, 0.4 μM 3′ universal primer, and 0.267 μM 5′ universal primer] was added and PCR performed using the following program: 95°C for 3 min, 7 cycles (for no cDNA immobilized beads) or 9 cycles (for cDNA-immobilized beads) of 98°C for 20 s, 63°C for 20 s, and 72°C for 5 minutes followed by 72°C for 5 min and hold at 4°C. PCR products were purified once with a 3.0× Pronex size-selective purification system (Promega, Madison, WI, USA) and eluted with 22 μL of 10mM Tris-HCl pH8.0. Amplified products were quantified using a Nanodrop 8000 (Thermo Fisher Scientific), and size distribution was analyzed by Agilent High Sensitivity DNA kit (Agilent Technologies, Santa Clara, CA, USA) with Agilent 2100 Bioanalyzer (Agilent Technologies) with appropriate dilutions.

Single-cell preparation. Lung cells were prepared as described previously with some modifications. Briefly, mice were anesthetized with isoflurane, lungs were perfused with PBS (Nacalai Tesque), and the left lung was collected. Human lung samples were collected from lung cancer patients with pulmonary fibrosis who underwent curative surgical resection from August 2015 to December 2019 at Nara Medical University Hospital. Informed consent was obtained from all patients who participated in the study herein. Removed lung samples were determined as non-fibrosis and fibrosis areas without lung cancer under thin-section computed tomography by two independent respiratory specialists. Murine or human lung samples were minced into 0.5 mm^2 with a razor blade and digested with Liberase solution [RPMI-1640 (Nacalai Tesque) supplemented with 10% FBS, 10 mM HEPES pH7.2-7.4, 0.25 mg/ml Liberase TM (Roche), and 2 kU/mL DNase I (Merck)] at 37 °C for 60 minutes. For murine samples, the cell suspension was agitated 20 times with an 18G needle (Terumo, Tokyo, Japan) after 20 minutes incubation, agitated 20 times with 21G needle (Terumo) after

40 minutes incubation, and agitated 10 times with 200 μL pipette tip. For human samples, the cell suspension was agitated 20 times with an 18G needle (Terumo) every 20 minutes incubation. Cell suspensions were passed through a 70 μm cell strainer (BD Biosciences), centrifuged at 4°C for 500 ×g for 5 minutes, and their supernatant was discarded. Cells were resuspended with 25% Percoll PLUS (Cytiva), agitated with an 18G needle five times. After under layered 65% Percoll PLUS, cell suspensions were centrifuged at 20°C for 1,000 ×g for 20 minutes, and the middle layer was collected. Resultant cell suspensions were diluted thrice with preparation medium [RPMI-1640 supplemented with 5% FBS and 10 mM HEPES], centrifuged 500 ×g at 4°C for 7 minutes, and their supernatants were discarded. Resultant cells were resuspended with 500 μL of preparation medium. Tumor cells of the subcutaneous model of lewis lung carcinoma were collected as described previously[13]. Spleen cells were collected from the subcutaneous model of lewis lung carcinoma as described previously[13] with some modifications. Anti-CD4 antibodies (clone GK1.5, BioLegend, 200 μg/head) were intraperitoneally injected at days 5 and 9 post tumor inoculation, and the spleen was harvested at day 12 post tumor inoculation. The spleen was mushed on a cell strainer with 5 ml of preparation medium, and resultant cells from the spleen were harvested and pooled. Then, spleen cells were suspended with ACK lysis buffer [155 mM NH_4Cl, 10 mM $KHCO_3$, 0.1 mM EDTA, pH 7.3] and incubated for 10 min at room temperature. The cell suspension was washed two times with PBS, filtered with 40 μm strainer, suspended with CellBanker 1, and gradually freezed at -80 °C by using CoolCell (Thermo Fisher Scientific) for long term storage. Stored spleen cells were thawed and washed just before use for processing by BD Rhapsody. Each single-cell suspension cell concentration was counted using Flow-count fluorospheres (Beckman Coulter, Brea, CA, USA) and a Gallios flow cytometer (Beckman Coulter).

Flow cytometry. For murine lung cells, single-cell suspensions were blocked with Fc block (anti-CD16/32, clone: 2.4G2, BioXcell, West Lebanon, NH, USA) and stained with appropriate antibody mixtures diluted with PBS supplemented with 2% FBS. After washing with PBS supplemented with 2% FBS, cells were suspended with PBS supplemented with 2% FBS and 0.5 μg/ml propidium iodide. For human lung cells, single-cell suspensions were washed once with PBS and stained with LIVE/DEAD Fixable Aqua Dead Cell Stain Kit (Thermo Fisher Scientific) at 4°C for 30 minutes. Cells were washed once with PBS supplemented with 2% FBS; cells were blocked with 2% normal mouse serum and stained with appropriate antibody mixtures diluted with PBS supplemented with 2% FBS. After washing with PBS supplemented with 2% FBS, cells were suspended with PBS supplemented with 2% FBS. Data were collected on a Gallios flow cytometer and analyzed using FlowJo software v10.6.2 (BD Biosciences). A detailed list of used antibodies is shown in

Supplementary Table 5.

cDNA synthesis and Exonuclease I treatment by BD Rhapsody system. For cell hashing, $CD45^+$ tumor-infiltrating leukocytes were stained with 2.5 μg/ml of Totalseq anti-mouse Hashtag-A antibodies (A0301-A0314, BioLegend, San Diego, CA, USA) at 4°C for 25 minutes and washed thrice with Cell Staining Buffer (BioLegend) and pooled equally as described previously[13]. Obtained single-cell suspensions were subjected to a BD Rhapsody system with BD Rhapsody Targeted & Abseq Reagent kit (BD Biosciences) following the manufacture's instructions. 10000 mouse lung cells and 20000 mouse $CD45^+$ tumor-infiltrating leukocytes were subjected to the BD Rhapsody workflow, and 6000 human lung cells were subjected to the BD Rhapsody Express workflow. After reverse transcription, Exonuclease I treatment of the resultant BD Rhapsody beads was performed at 37°C for 60 minutes with 1,200 rpm on a Thermomixer C with Thermotop. Resultant beads were immediately chilled on ice; the supernatant was removed and washed with 1 mL of WTA wash buffer, 200 μL of BD Rhapsody lysis buffer (for inactivation of enzyme), once with 1 ml of WTA wash buffer, twice with 500 μL of WTA wash buffer, resuspended with 200 μL of Beads resuspension buffer and stored at 4°C. During the washing step, beads-containing DNA LoBind tubes were replaced twice. For the spleen cell sample subjected to BD WTA kit (BD Biosciences), half of the BD Rhapsody beads were split just after Exonuclease I treatment, and the enzyme was heat-inactivated at 80°C for 20 minutes.

Amplification of cDNA by BD Rhapsody WTA kit. Half of the Exonuclease I-treated BD Rhapsody beads from the spleen cell sample were subjected to BD Rhapsody kit for cDNA amplification following the manufacture's instructions.

Amplification of cDNA and hashtag libraries by TAS-Seq from BD Rhapsody beads. Half of the reverse-transcribed, Exonuclease I-treated BD Rhapsody beads were subjected to TAS-Seq workflow for cDNA and/or hashtag library amplification. After removing the supernatant and washing once with nuclease-free water, the beads were mixed with 200 μL of TdT mixture [1×TdT buffer, 1.2 mM dCTP, 0.06 mM ddCTP, 15 U/μL TdT, 0.1 U/μL RNase H] and incubated for 30 minutes at 37°C with 1,200 rpm on a Thermomixer C with Thermotop. Reactions were chilled on ice immediately after the reaction was completed. After the supernatant was removed, beads were washed with 1 mL of WTA wash buffer, 200 μL of BD Rhapsody lysis buffer, once with 1 ml of WTA wash buffer, twice with 500 μL of WTA wash buffer, and resuspended with 100 μL of 10 mM Tris-HCl pH8.0. During the washing step, beads-containing DNA LoBind tubes were replaced twice. Beads were split into two parts and

transferred into new 8-strip tubes, the supernatant was discarded, and 25 µL of second-strand synthesis mixture [1× KAPA Hifi ReadyMix and 0.4 µM 5′ WTA-9G primer (for spleen cells and tumor-infiltrating leukocytes) or 5′ LibA-9G primer (for human and mouse lung samples)] was added, and second-strand synthesis was performed according to the following program: 95°C for 3 min, 98°C for 20 s, 47°C for 2 min, 72°C for 7 min, and hold at 4°C. Then, 75 µL of 1st round of whole-transcriptome amplification (WTA) mix [1× KAPA Hifi ReadyMix, 0.4 µM 3′ universal primer, and 0.267 µM 5′ WTA primer] (for spleen cells), [1× KAPA Hifi ReadyMix, 0.4 µM 3′ universal primer and 0.267 µM 5′ LibA primer] (for mouse and human lung cells), or [1× KAPA Hifi ReadyMix, 0.4 µM 3′ universal primer, 0.267 µM 5′ universal primer and 0.267 µM 5′ hashtag primer] (for tumor-infiltrating leukocytes) was added, split samples into two tubes (50 µL each), and PCR performed using the following program: 95°C for 3 min, seven cycles of 98°C for 20 s, 63°C for 20 s, and 72°C for 5 min, followed by 72°C for 5 min and hold at 4°C. PCR products with no hashtag libraries (spleen, mouse lung, and human lung cells) were combined and purified twice with 0.65× AmPure XP beads (Beckman Coulter) and eluted with 21 µL of nuclease-free water. PCR products with hashtag libraries (for tumor-infiltrating leukocytes) were combined, and cDNA product was purified by 0.65× AmPure XP beads, and unbounded fraction was isolated. Hashtag product was purified from the unbounded fraction by adding additional 0.7× AmPure XP beads (final 1.35×). Then, cDNA and hashtag libraries were further purified by 0.65× and 1.35×AmPure XP beads, respectively, and eluted with 21 µL of nuclease-free water. For amplification of the cDNA libraries, 2nd round of WTA mix [25 µL of 2× KAPA Hifi ReadyMix, 2 µL of 10µM 3′ universal primer, and 2 µL of 10µM 5′ WTA primer (for spleen cells and tumor-infiltrating leukocytes) or 5′ LibA primer (for human and mouse lung samples)] was added to the cDNA libraries, and PCR performed using the following program: 95°C for 3 min, 9 cycles (for mouse and human lung cells) or 13 cycles (for spleen cells) of 98°C for 20 s, 63°C for 20 s, and 72°C for 5 min followed by 72°C for 5 min and hold at 4°C. For amplification of the hashtag libraries, 2nd round of hashtag-amplification mix [25 µL of 2× KAPA Hifi ReadyMix, 2 µL of 10 µM 3′ universal primer, and 2 µL of 10 µM 5′hashtag primer] was added to the cDNA libraries, and PCR performed using the following program: 95°C for 3 min, 9 cycles of 98°C for 20 s, 63°C for 20 s, and 72°C for 45 sec, followed by 72°C for 5 min and hold at 4°C. Amplified products were purified two times with 0.65× AmPure XP beads (for cDNA libraries) or 1.35× AmPure XP beads (hashtag libraries) and eluted with 30 µL of 10 mM Tris-HCl pH8.0. Then, barcoded PCR mix for hashtag library [1× KAPA Hifi ReadyMix, 0.4 µM 3′ i5-UDI0033 primer, 0.4 µM i7-UDI0033 primer, 5 ng of purified hashtag library] was prepared and PCR performed using the following program: 95°C for 3 min, 9 cycles of 98°C for 20 s, 63°C for 20 s, and 72°C for 45 sec, followed by 72°C for 5

min and hold at 4°C. Amplified products were purified by double size selection with 0.8× → 0.4 × (final 1.2×) AmPure XP beads and eluted with 25 µL of 10 mM Tris-HCl pH8.0. Amplified products were quantified using a Nanodrop 8000, and size distribution was analyzed by Agilent High Sensitivity DNA kit with an Agilent 2100 Bioanalyzerwith appropriate dilutions. Primer sequences used for this study were shown in **Supplementary Table 6**.

Illumina library construction and sequencing. Illumina libraries were constructed from 100 ng of amplified cDNA libraries using the NEBNext Ultra II FS library prep kit for Illumina (New England Biolabs) with some modifications. Briefly, fragmentation, end-repair, and A-tailing were performed using the following program: 32°C for 5 min, 65°C for 30 min, and hold at 4°C. Then, 2.5 µL of 3.3 µM illumine adapter was used for adapter ligation. Ligated products were purified by double size selection with 10 µL → 25 µL AmPure XP beads and eluted with 15 µL of nuclease-free water. Nine cycles of Barcoding PCR were performed using i5-UDI00XX and i7-UDI00XX primers. Resultant products were purified twice by double size selection with 0.5× → 0.3 × (final 0.8×) AmPure XP beads and eluted with 30 µL of 10 mM Tris-HCl pH8.0. Size distribution of amplified products was analyzed by Agilent High Sensitivity DNA kit with Agilent 2100 Bioanalyzer or MultiNA system (Shimazu, Kyoto, Japan) with appropriate dilutions. Resultant libraries and barcoded hashtag libraries were quantified using the KAPA Library Quantification Kit (KAPA Biosystems). Primer sequences used for this study were shown in **Supplementary Table 6.** The primers were purchased from Eurofins Genomics (Tokyo, Japan) or Integrated DNA Technologies (Coralville, IA USA). Sequencing was performed by Illumina Novaseq 6000 sequencer (Illumina, San Diego, CA, USA) following the manufacturer's instructions. Pooled library concentration was adjusted to 1.75 nM, and 12% PhiX control library v3 (Illumina) was spiked into the library.

Fastq data preprocessing and generation of the single-cell gene-expression matrix. Pair-end Fastq files (R1: cell barcode reads, R2: RNA reads) of TAS-Seq and BD WTA kit data were processed as follows. Adapter trimming of sequencing data was performed using cutadapt 2.10[21]. Filtered reads were chunked into 16 parts for parallel processing by using Seqkit 0.14.0[22]. Filtered cell barcode reads were annotated by Python script provided by BD Biosciences with minor modification for compatibility to Python 3.7. Reference RNA sequences were built by concatenating cDNA and ncRNA fasta files of the Ensembl database (build GRCm38 release-101 for mouse data and GRCh38 release-101 for human data)[23]. Associated cDNA reads were mapped to reference RNA using bowtie2-2.4.2[24] by the following parameters: -p 2 --very-sensitive-local -N 1 -norc -seed 656565 -reorder. Then, cell

barcode information of each read was added to the bowtie2-mapped BAM files by the python script and pysam 0.15.4 (https://github.com/pysam-developers/pysam), and read counts of each gene in each cell barcode were counted using mawk. Resultant count data was converted to a single-cell gene-expression matrix file. The inflection point of the knee-plot (total read count versus the rank of the read count) was detected using DropletUtils package[25] in R 3.6.3 (https://cran.r-project.org/). Cells of which total read count was over inflection point were considered as valid cells. Because unique-molecule identifiers (UMIs) of BD Rhapsody beads are 8-base UMIs directly before polyT stretch, which might not be sufficient to exert theoretical UMI diversity by the distortion of base frequencies and to avoid UMI collision (more than 10-base UMIs is necessary for scRNA-seq datasets)[26], we did not use BD Rhapsody UMIs for TAS-Seq data and BD WTA data.

Background subtraction of TAS-Seq expression matrix by distribution-based error correction (DBEC). To reduce background read counts of each gene that were possibly derived from RNA diffusion during cell lysis step within BD Rhapsody cartridge and reverse transcription, we performed distribution-based error correction (DBEC) that is included in BD Rhapsody targeted scRNA-seq workflow. To estimate background and signal read count distribution, we used the Gaussian mixture model previously used to estimate the gene-expression distribution of scRNA-seq datasets[27]. First, genes of which $\log_2(x+1)$-transformed maximum expression over 8 were selected, and biexponential transformation was applied to each gene count by using FlowTrans package[28] in R 3.6.3. Next, Gaussian mixture components (model E, from one to three components) were detected using mclust package[29] in R 3.6.3, and the average expression of each component was calculated. Genes of which the maximum average expression of each component was over 5.5 were selected. Then, if the difference of the average expression of each component against their maximum expression was greater than 5, the expression level of the components was considered to be background gene expression and converted expression of the components to 0.

Single-cell clustering and annotation. Clustering of single cells of each dataset was performed using Seurat v2.3.4[15] in R 3.6.3. For Tabula Muris Smart-seq2 data, the expression value of ERCC spike-ins was excluded. Seurat object for each dataset was created using CreateSeuratObject function (min.cells=5, min.genes=500). scRNA-seq library metrics, including mitochondrial gene-count proportion, ribosomal protein gene-count proportion, and ribosomal RNA count proportion, were calculated using R 3.6.3 and visualized using geom_hex (bins=100) function in ggplot2 package[30]. Gene count and $\log_{10}$ converted read count distribution was visualized using the RidgePlot function in the Seurat package with

default parameters. Cells of which mitochondrial gene proportion was over 0.4 were filtered out by FilterCells function in Seurat v2.3.4. The expression data was normalized by normalizeData function (scale.factor = 1,000,000 (for TAS-Seq, BD WTA and Smart-seq2 data according to the analytical parameter used by Tabula Muris[4]) or 10,000 (for 10X data)), and scaled with ScaleData function in Seurat v2.3.4. In mouse datasets, read counts or UMI counts of each cell within each dataset were regressed as a confounding factor within the ScaleData function. In human datasets, read counts and percentage of mitochondrial gene proportion of each cell within each dataset were regressed as a confounding factor within the ScaleData function. Highly-variable genes of each dataset were identified using the FindVariableGenes function in Seurat v2.3.4. with the following parameters: mean.function = ExpMean, dispersion.function = LogVMR, x.low.cutoff = 0.1, x.high.cutoff = Inf, y.cutoff = 0.5. Then, principal component analysis (PCA) against identified highly-variable genes and projection of PCA onto entire data was performed using RunPCA (number of calculated PCs were 100) and ProjectPCA functions in Seurat v2.3.4. Enrichment of each PC was calculated using the JackStraw function (num.replicate = 100), and PCs that were significantly enriched statistically ($p \leq 0.05$) were selected for clustering and dimensional reduction analyses. Cell clustering was performed using FindClusters function (resolution = 2.0 (for 10X Chromium v2 and Smart-seq2 data), 5.0 (for TAS-Seq data of murine and human lungs), and 4.0 (for 10X Chromium v3 data)) in Seurat v2.3.4 against the significant PCs, and dimensional reduction was performed using python wrapper of Fast Fourier transform-accelerated interpolation-based t-stochastic neighbor embedding (FIt-SNE)[31] v1.2.1 (perplexity = 100, df = 0.9, random_seed=42, max_iter=1000, and all the other parameters were set as defaults) through reticulate package (https://github.com/rstudio/reticulate) in R 3.6.3. Statistically significant marker genes of each identified cluster were identified using parallelized FindMarkers function in Seurat v2.3.4 (test.use="wilcox", only.pos=TRUE, min.pct=0.1, logfc.threshold=0.25, adjusted p (Bonferroni correction) ≤ 0.05). Then, each identified cluster was manually annotated by their marker genes previously reported as cell subset-defining marker genes, and the lineage marker double-positive cells were annotated as doublets. Next, we further sub-clustered cell subsets that were not fully separated into known cell subsets (dendritic cell (DC) subsets Smart-seq2 and 10X Chromium v2, T cell subsets in Smart-seq2, monocyte/DC/interstitial macrophage subset in TAS-Seq, gamma delta T cells/innate lymphoid cells in TAS-Seq, monocyte/interstitial macrophage subset in 10X Chromium v3 datasets) using Seurat v2.3.4 by the similar workflow of whole-cell data, and incorporate their annotation into Seurat object of whole-cell data (detail analysis parameters and associated codes were deposited at https://github.com/s-shichino1989/TASSeq-paper.). Cell subset annotations and their compositions were visualized in 2D FIt-SNE space and

stacking plot, respectively. All of the identified marker genes are shown in **Supplementary Table 7**, and cell cluster annotations are shown in **Supplementary Table 1 and 3**.

Cell composition correlation analysis between flow cytometric data and scRNA-seq data. The percentage of the abundance of specific cell subsets against total cells were calculated, and Pearson's correlation coefficients, linear regression, and associated p-values between the cell composition of flow-cytometric data and each scRNA-seq data were calculated using GraphPad Prism 9.1.2 (Graphpad Software, La Jolla, CA, USA).

Inference of cell-cell interaction network from scRNA-seq data. Inference of cell-cell interaction network of each murine lung scRNA-seq dataset was performed using CellChat 1.0.0 package[16]. First, total cell number was downsampled to 1737 cells (the total cell number of Smart-seq2 dataset) by SubsetData function of Seurat v2.3.4, re-normalized using NormalizeData function (scale.factor = 1000000) of Seurat v2.3.4., and cellchat objects of each dataset were created from raw expression data, normalized expression data, and associated cell-annotation metadata extracted from the Seurat objects. Identification of overexpressed interactions, calculation of communication probability between cell subsets, and identification of overexpressed pathways was performed according to the CellChat default workflow (https://github.com/sqjin/CellChat) changing the threshold of ligand-receptor gene-expression abundancy within cell subsets (changing thresh.pc parameter = 0.05, 0.15, 0.25, 0.5, and 0.75 in the identifyOverExpressedGenes function in CellChat 1.0.0). Outgoing/incoming signaling strength was calculated and visualized using the netAnalysis_signalingRole_scatter function in CellChat 1.0.0. Circle plot visualization of cell-cell interaction network and the strength of the communication between each cell subset were performed using the netVisual_circle function in CellChat 1.0.0. The compareInteractions function in CellChat 1.0.0 calculated the number of inferred cell-cell interactions of each dataset after merging cellchat objects. Heatmap visualization of identified cell-cell interaction pathways was performed using pheatmap package in R 3.6.3.

Statistical analyses. The significance of the difference of the read count number between TAS-Seq and Smart-seq2 datasets, and the difference of detected gene number between TAS-Seq and BD WTA datasets, and the difference of detected gene number of each cell subset between TAS-Seq and Smart-seq2 datasets were calculated using wilcox_test function of coin package[32] in R 3.6.3. The significance of the enrichment of PCs and marker genes of each cell cluster was calculated using Seurat v2.3.4 package in R 3.6.3. The significance of the correlation between the cell composition of flow-cytometric data and each scRNA-seq data

was calculated using GraphPad Prism 9.1.2 (Graphpad Software, La Jolla, CA, USA). All statistical analyses were conducted with a significance level of $\alpha = 0.05$ ($p \leq 0.05$).

Study approval. All animal experiments were reviewed and approved by the Animal Experiment Committee of Tokyo University of Science (approval number: S17034, S18029, S19024, and S20019). All human studies were approved by the Ethics Committee of Nara Medical University (Approval No. 1973) and Tokyo University of Science (Approval No. 18018).

Data Availability

Raw data, annotated gene-expression matrix, and associated metadata from these experiments have been deposited in the NCBI gene expression omnibus (GEO); accession GSE180149. Public data used for this study is available at https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733 (Tabula Muris data), and GSE145998 (10X Chromium v3 data), https://www.nature.com/articles/s41586-020-2922-4#Sec33; supplementary table 2 (cell abundancy data of human lungs of 10X Chromium v2 and Smart-seq2 data).

Code Availability

The mapping pipeline for TAS-Seq data is available at https://github.com/s-shichino1989/TASSeq. All the R code used for this study and rDBEC R package (includes functions of distribution-based error correction and utility functions for this study) are available at https://github.com/s-shichino1989/TASSeq-paper.

Acknowledgments

We thank J. Yasuda for their excellent technical assistance. This work was supported by the Japan Agency for Medical Research and Development PRIME program (S.S., JP21gm6210025) and the Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research on Innovative Areas program (Inflammation Cellular Sociology, 17H06392, K.M.), and Grant-in-Aid for Young Scientists (19K16620, S.S.).

Authorship contributions

S.S., S.U., and K.M. designed the study. S.S. performed all the experiments except Lewis lung carcinoma experiment. C.C. and W.B. performed Lewis lung carcinoma experiment. T.Okayama., E.S., K.I, T.S. performed sequencing experiments and contributed to the preprocessing of fastq data. S.S., N.O.S., M.K., and T.I. performed the experiment of human lung samples. S.Hontsu performed diagnosis of the existence of fibrosis in RA-ILD samples. N.S. and T.K performed surgical operation of RA-ILD patient. S.S., S.U., S. Hashimoto., T. Ogawa., H. A., W. B., C.C., M. K., T. I., T. Okayama., E. S., K. I., T. S., and K. M. wrote the manuscript. K.M. supervised the study.

Competing interests

Competing interests; S.S. reports advisory role for ImmunoGeneTeqs, Inc; stock for ImmunoGeneTeqs, Inc, S.U. reports advisory role for ImmunoGeneTeqs, Inc; stock for ImmunoGeneTeqs, Inc, IDAC Theranostics, Inc. H.A. reports stock for ImmunoGeneTeqs, Inc., K.M. reports consulting or advisory role for Kyowa-Hakko Kirin, ImmunoGeneTeqs, Inc; research funding from Kyowa-Hakko Kirin, and Ono; stock for ImmunoGeneTeqs, Inc, IDAC Theranostics, Inc., T.I. reports consulting or advisory role for ROHTO Pharmaceutical Co., Ltd; research funding from ROHTO Pharmaceutical Co., Ltd.

Figure Legends

Figure 1. Principles, proof-of-concept, and cell hashing compatibility of TAS-Seq.

a, Diagram of the TAS-Seq library preparation workflow . First, cell load, lyse cells, mRNA trap, and cDNA synthesis were performed according to the standard workflow of a BD Rhapsody system. After reducing free primers by Exonuclease I treatment, dC-tailing reaction was performed by TdT/RNaseH with ddCTP:dCTP (1:20) and Co^{2+} supplementation. Tailing reaction was stochastically stopped by ddCTP incorporation into 3′ termini **(1)**. After inactivation and washing, beads were separated into four parts of PCR tubes, and second strand synthesis was performed using 5′universal-dG9 (5'BDWTA-dG9 or 5'LibA-dG9) primer and PCR master mix **(2)**. Reactions were immediately chilled on ice, and 1st PCR was performed by directly adding PCR master mix and appropriate primers (3' and 5′ universal primer (if only amplify cDNA) or 3', 5′ universal primer and HTO primer (if use cell hashing antibodies)) **(3)**. Resultant reactions were pooled, and size-selection was performed using AmPure XP beads, and amplified by 2nd PCR **(4)**. Then, sequencing libraries were generated using resultant cDNA and hashtag libraries and sequenced by Illumina Novaseq 6000. **b,** TAS-Seq tolerance against TdT reaction time and TdT activity. TdT reaction with ddCTP:dCTP (1:20) or dCTP only with Co^{2+} supplementation was performed from 5 minutes or 30 minutes with different TdT enzyme amounts against exonuclease I-treated BD Rhapsody beads. Second strand synthesis and 1st PCR were performed, and all of the products were purified and their size distributions analyzed. Note that the length of primer-derived bi-products (arrows) peaked at around 136 bp and did not extend over 200 bp in every reaction time. **c,** TAS-Seq tolerance against TdT reaction time. TdT reaction with ddCTP:dCTP (1:20) with Co^{2+} supplementation was performed 30 or 45 minutes against cDNA-synthesized, exonuclease I-treated BD Rhapsody beads. Second strand synthesis and 1st PCR were performed, and all of the products were purified, and their size distributions analyzed. Note that the length of primer-derived bi-products (arrows) peaked at around 136 bp and did not extend over 200 bp in every reaction time. In addition, amplified cDNA was also visible (arrowheads). **d,** Size distribution of TAS-Seq amplified cDNA (total 16 cycles of PCR) library of single cells derived from the murine lung. cDNA size was over 400 bp and peaked at 994 bp. **e,** Size distribution of TAS-Seq amplified cDNA and hashtag libraries from $CD45.2^{+}$ cells of subcutaneous tumor model of lewis lung carcinoma. cDNA library peaked at 1050bp, and the associated hashtag library peaked at 239 bp (after barcoding). **f,** Heatmap representation of normalized and $\log_2$-centered hashtag count of each cell. Row means each cell, and column means each hashtag. The normalized $\log_2$ hashtag count difference between the first and second most counted hashtags was calculated for each cell. Cells were ranked in ascending order by the difference, and the top 4.26% cells were identified as doublets. Then,

each cell remaining was assigned to the most counted hashtag. Of note, gene-detection distribution, shown by the ridgeline plot, was similar among each hashtag-assigned sample. Assigned cell number against each hashtag was plotted. **b and c,** Representative results of two independent experiments are shown.

Figure 2. TAS-Seq accurately detects cell composition of murine and human lungs with high gene-detection sensitivity.

a. Diagram of the sample preparation. Single-cell suspension was processed from the left lung of 8-week-old female C57BL/6J mice, and TAS-Seq constructed scRNA-seq library. **b.** Ridgeline plot representation of the distribution of detected gene number of each dataset of the murine lung. Plots were ordered from the back to the front by ascending order of the mean of the detected gene number. **c.** Scatter plot of the read number/detected gene number of each cell of TAS-Seq and Smart-seq2 datasets. **d.** Violin plot representation of the read number distribution of each cell of TAS-Seq and Smart-seq2 datasets. Box plot shows the mean of the read number with upper and lower quantile. ****p = 0, W = 687860 (Wilcoxon rank-sum test). **e.** The number of highly-variable genes identified by Seurat v2.3.4 package in TAS-Seq and Smart-seq2 datasets. **f.** Visualization of cell clustering results of each scRNA-seq dataset of the murine lung by Seurat v2.3.4 package in 2D FIt-SNE space. The stacking plot shows the composition of each annotated cell. Each annotated cell was colored commonly between FIt-SNE and stacking plots. Detail of the cell annotations and associated marker genes of each dataset are represented in **Supplementary Table 1**. **g.** Comparison of cell composition between flow-cytometric data and scRNA-seq datasets of the murine lung. Pearson's correlation coefficients and associated p-values were calculated. The gating scheme for identifying each cell subset by flow cytometry is shown in Supplementary Figure 3. **i.** Comparison of cell composition between flow-cytometric data and TAS-Seq data of RA-ILD lungs. Pearson's correlation coefficients and associated p-values were calculated. The gating scheme for identifying each cell subset by flow cytometry is shown in **Supplementary Figure 5.** The composition of neutrophils of flow-cytometric and TAS-Seq data is shown on the right side. **j.** Composition of neutrophils in Smart-seq2 and 10X Chromium v2 datasets of human lungs reported by Travaglini *et al.*. Note that human lung neutrophils were not detected in 10X Chromium v2. **k.** Changes of the number of inferred interactions and pathways of cell-cell interaction network of each scRNA-seq dataset of murine lung predicted by CellChat when the threshold of genes of which minimum fraction of expressed cells within each cell subset. **l.** Circle plot visualizations of cell-cell interaction network of TAS-Seq, Smart-seq2, 10X v2/v3 datasets within commonly-detected particular cell subsets which were strongly contributed to the network at least one dataset. Circle sizes are normalized to the cell number of each subset. Edge width represents communication strength (wider edge means stronger communication between source and target cell subsets), normalized among all datasets. Edge colors are similar to the color of their source cell subsets. See also **Supplementary Figure 6c** for the cell-cell interaction network of all cell subsets. Abbreviations: dendritic cell (DC), conventional DC (cDC), plasmacytoid DC (pDC), endothelial cell (Endo), lymphatic

endothelial cell (LEC), vascular endothelial cell (VEC), epithelial cell (Epi), alveolar type 1 epithelial cell (AT1), alveolar type 2 epithelial cell (AT2), fibroblast (FB), innate lymphoid type 2 cell (ILC2), macrophage (Mac), alveolar macrophage (AM), interstitial macrophage (IM), monocyte (Mo), natural killer cell (NK), smooth muscle cell (SMC), gamma-delta T cell (gdT).

Supplementary Figure 1. TAS-Seq outperformed the BD Rhapsody WTA kit in terms of gene-detection sensitivity.

a. Diagram of the experimental workflow. Using 1600 frozen mouse adult spleen cells, cDNA synthesis was performed by the same BD Rhapsody cartridge, and resultant beads were separated into two partsthem which were processed by TAS-Seq and BD commercial WTA kit, respectively. **b.** Scatter plot representation of the read number/detected gene number of each cell of TAS-Seq (salmon, 736 cells) and BD WTA (blue, 715 cells) datasets of mouse spleen cells. A Violin plot of the detected gene number of each dataset is also shown on the right side of the scatter plot. $p = 2.2 \times 10^{-163}$, $W = 45757.5$ by Wilcoxon rank-sum test. **c.** Hexagonal pseudocolor plot of library quality metrics of the TAS-Seq and BD WTA datasets. Blue, green, and red indicate more cells located within the same hexagonal area. Note that mitochondrial gene proportion, ribosomal protein gene proportion, and ribosomal RNA proportion were similar between the two datasets, suggesting comparable performance on the library metrics between TAS-Seq and BD WTA kit.

Supplementary Figure 2. Gating scheme for identification of murine lung cell subsets by flow cytometry. Single-cell suspension of 8-week-old C57BL/6J female murine lung, subjected to TAS-Seq analysis, was analyzed by flow cytometry. **a.** Gating scheme of murine lung endothelial cells, epithelial cells, smooth muscle cells (SMC)/pericytes, and fibroblasts. **b.** Gating scheme of murine lung myeloid cell subsets. **c.** Gating scheme of murine lung lymphoid cell subsets.

Supplementary Figure 3. Comparison of detected gene number of each cell subset of murine lung between TAS-Seq and Smart-seq2 datasets. Violin plot representing the distribution of detected gene number of each dataset among commonly-detected cell subsets in TAS-Seq (salmon) and Smart-seq2 (green). Boxplot shows mean, upper and lower quantile of detected genes. **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$ by Wilcoxon rank-sum test. Exact p-values and W statistics are shown in **Supplementary Table 2**.

Supplementary Figure 4. TAS-Seq data of the lungs of a human RA-ILD patient. Visualization of cell clustering results of TAS-Seq data of human RA-ILD lungs by Seurat v2.3.4 package in 2D FIt-SNE space. The stacking plot showed the composition of each annotated cell. Each annotated cell was colored commonly between FIt-SNE and stacking plot. Detail of the cell annotation and associated marker genes are represented in **Supplementary Table 3**.

Supplementary Figure 5. Gating scheme for identification of human RA-ILD lung cell subsets by flow cytometry. Single-cell suspension of non-fibrotic and fibrotic lung samples from a human RA-ILD patient, subjected to TAS-Seq analysis, was analyzed by flow cytometry. **a.** Gating scheme of human lung endothelial cells, epithelial cells, smooth muscle cells (SMC)/pericytes, and fibroblasts. **b.** Gating scheme of human lung leukocytes.

Supplementary Figure 6. Difference of CellChat-inferred Cell-cell interaction network of murine lungs between TAS-Seq, Smart-seq2, and 10X Chromium v2/v3 datasets. a. Heatmap representation of detected pathways within the network at several thresholds of minimum expression of genes in each cell subset (from 0.05 to 0.5). Detected pathways are colored by magenta, and undetected pathways are colored by grey. Commonly-detected pathways are separately shown in **Supplementary Table 4** to show the difference between datasets. Fibroblast growth factor (FGF), bone morphologic protein (BMP), sonic hedgehog (HH), NOTCH, and WNT signaling are highlighted by red arrows. **b.** Scatter plot of incoming (target) and outgoing (source) signaling strength within cell-cell interaction network of each cell subset (minimum expression of genes in each cell subset ≥ 0.15, a minimum number of expressed cells ≥ 10, a threshold of the significance of the interaction ≤ 0.05). Dot size represents the sum of the number of incoming and outgoing signaling of each cell subset. Vascular endothelial cells, alveolar type 2 cells, and $Inmt^{hi}$ alveolar fibroblasts were strongly connected with the TAS-Seq and Smart-seq2 dataset network, and cell subsets were more strongly connected in TAS-Seq dataset than the Smart-seq2 dataset. The contribution of alveolar type 2 cells was depleted in 10X v3 and v3 datasets. **c.** Circle plot visualizations of all of the cell-cell interaction networks of TAS-Seq, Smart-seq2, 10X v2/v3 datasets. Circle sizes are normalized to the cell number of each subset. Edge width represents communication strength (wider edge means stronger communication between source and target cell subsets), normalized among all datasets. Edge colors are the same as the color of their source cell subsets. Abbreviations: dendritic cell (DC), conventional DC (cDC), plasmacytoid DC (pDC), endothelial cell (Endo), lymphatic endothelial cell (LEC), vascular endothelial cell (VEC), epithelial cell (Epi), alveolar type 1 epithelial cell (AT1), alveolar type 2 epithelial cell (AT2), fibroblast (FB), innate lymphoid type 2 cell (ILC2), macrophage (Mac), alveolar macrophage (AM), interstitial macrophage (IM), monocyte (Mo), natural killer cell (NK), smooth muscle cell (SMC), gamma-delta T cell (gdT).

Supplementary Table 1. Cell annotations, associated marker genes, and associated references for mouse datasets.

Supplementary Table 2. Exact p-value of Wilcoxon rank-sum test of Supplementary Figure 3.

Supplementary Table 3. Cell annotations, associated marker genes, and associated references for human RA-ILD lung datasets.

Supplementary Table 4. List of the commonly-detected pathways in CellChat analysis.

Supplementary Table 5. List of the antibodies used for this study.

Supplementary Table 6. List of the primer sequences used for this study.

Supplementary Table 7. All of the identified marker genes of each dataset by Seurat analysis.

References

1. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**(2017).
2. 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).
3. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-8 (2013).
4. Tabula Muris, C. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372 (2018).
5. Travaglini, K.J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619-625 (2020).
6. Takeda, A. *et al.* Single-Cell Survey of Human Lymphatics Unveils Marked Endothelial Cell Heterogeneity and Mechanisms of Homing for Neutrophils. *Immunity* **51**, 561-572 e5 (2019).
7. 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. *J Biol Chem* **294**, 18220-18231 (2019).
8. 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).
9. 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**, R31 (2013).
10. Huang, H. *et al.* Non-biased and efficient global amplification of a single-cell cDNA library. *Nucleic Acids Res* **42**, e12 (2014).
11. Deng, G. & Wu, R. An improved procedure for utilizing terminal transferase to add homopolymers to the 3' termini of DNA. *Nucleic Acids Res* **9**, 4173-88 (1981).
12. Chirpich, T.P. The effect of different buffers on terminal deoxynucleotidyl transferase activity. *Biochim Biophys Acta* **518**, 535-8 (1978).
13. Chen, C.-Y. *et al.* Combining an alarmin HMGN1 peptide with PD-L1 blockade facilitates stem-like CD8+ T cell expansion and results in robust antitumor effects. *BioRxiv* (2021).
14. Koenitzer, J.R., Wu, H., Atkinson, J.J., Brody, S.L. & Humphreys, B.D. Single-Nucleus RNA-Sequencing Profiling of Mouse Lung. Reduced Dissociation Bias and Improved Rare Cell-Type Detection Compared with Single-Cell RNA Sequencing. *Am J Respir Cell Mol Biol* **63**, 739-747 (2020).
15. 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).

16. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat Commun* **12**, 1088 (2021).
17. Zepp, J.A. & Morrisey, E.E. Cellular crosstalk in the development and regeneration of the respiratory system. *Nat Rev Mol Cell Biol* **20**, 551-566 (2019).
18. Liu, X. *et al.* Definition and Signatures of Lung Fibroblast Populations in Development and Fibrosis in Mice and Men. *BioRxiv* (2020).
19. Hogan, B.L. *et al.* Repair and regeneration of the respiratory system: complexity, plasticity, and mechanisms of lung stem cell function. *Cell Stem Cell* **15**, 123-38 (2014).
20. Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* **16**, 987-990 (2019).
21. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).
22. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).
23. Yates, A.D. *et al.* Ensembl 2020. *Nucleic Acids Res* **48**, D682-D688 (2020).
24. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
25. 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).
26. Petukhov, V. *et al.* dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biol* **19**, 78 (2018).
27. Korthauer, K.D. *et al.* A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol* **17**, 222 (2016).
28. Finak G, M.-P.J., Gottardo R. flowTrans: Parameter Optimization for Flow Cytometry Data Transformation. *R package version 1.36.0.* (2019).
29. Scrucca, L., Fop, M., Murphy, T.B. & Raftery, A.E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* **8**, 289-317 (2016).
30. H, W. *ggplot2: Elegant Graphics for Data Analysis.*, (Springer-Verlag New York, 2016).
31. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S. & Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods* **16**, 243-245 (2019).
32. Hothorn, T., Hornik, K., van de Wiel, M.A.V. & Zeileis, A. Implementing a Class of Permutation Tests: The coin Package. *J Stat Softw* **28**, 1-23 (2008).

Figure 1

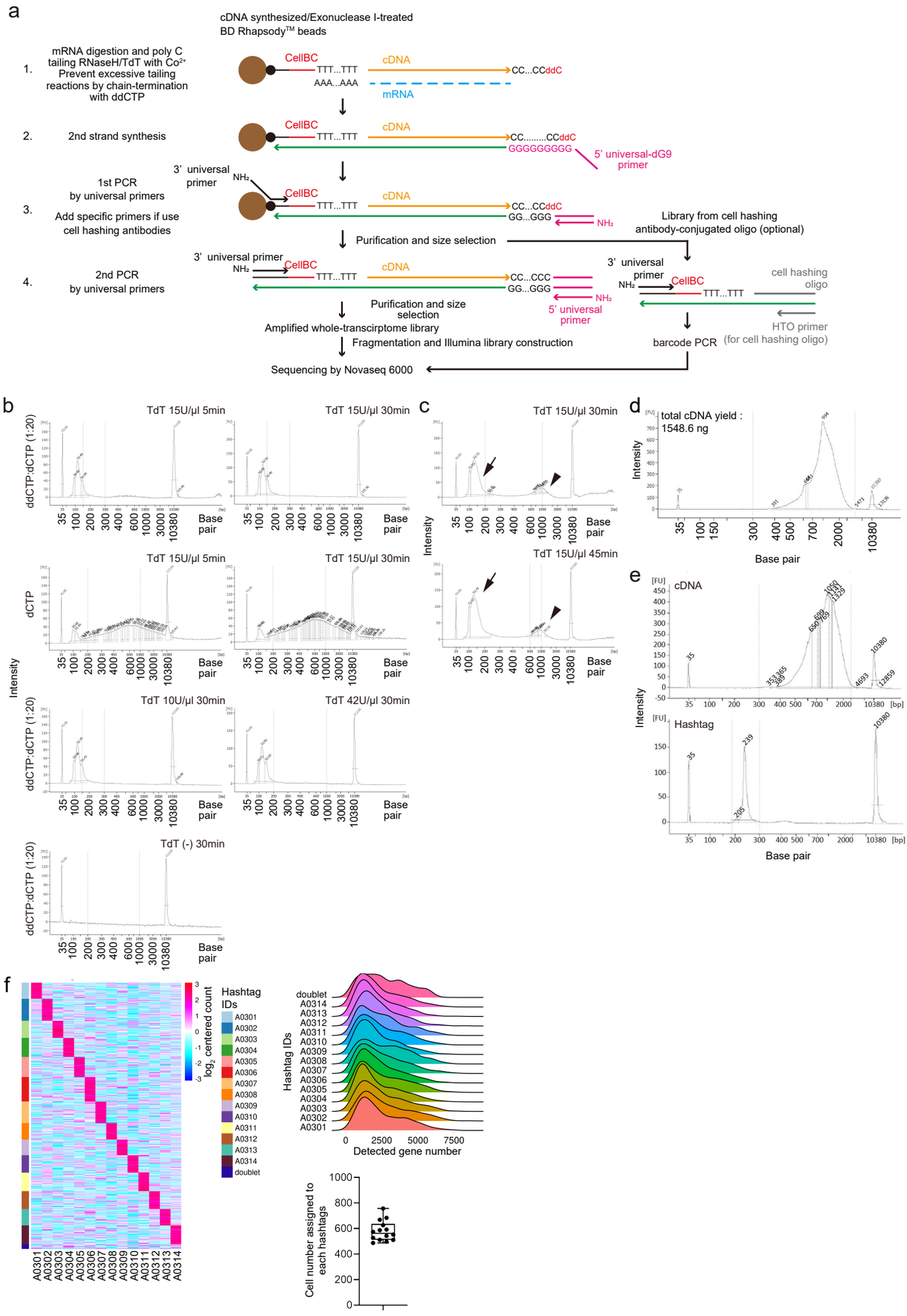
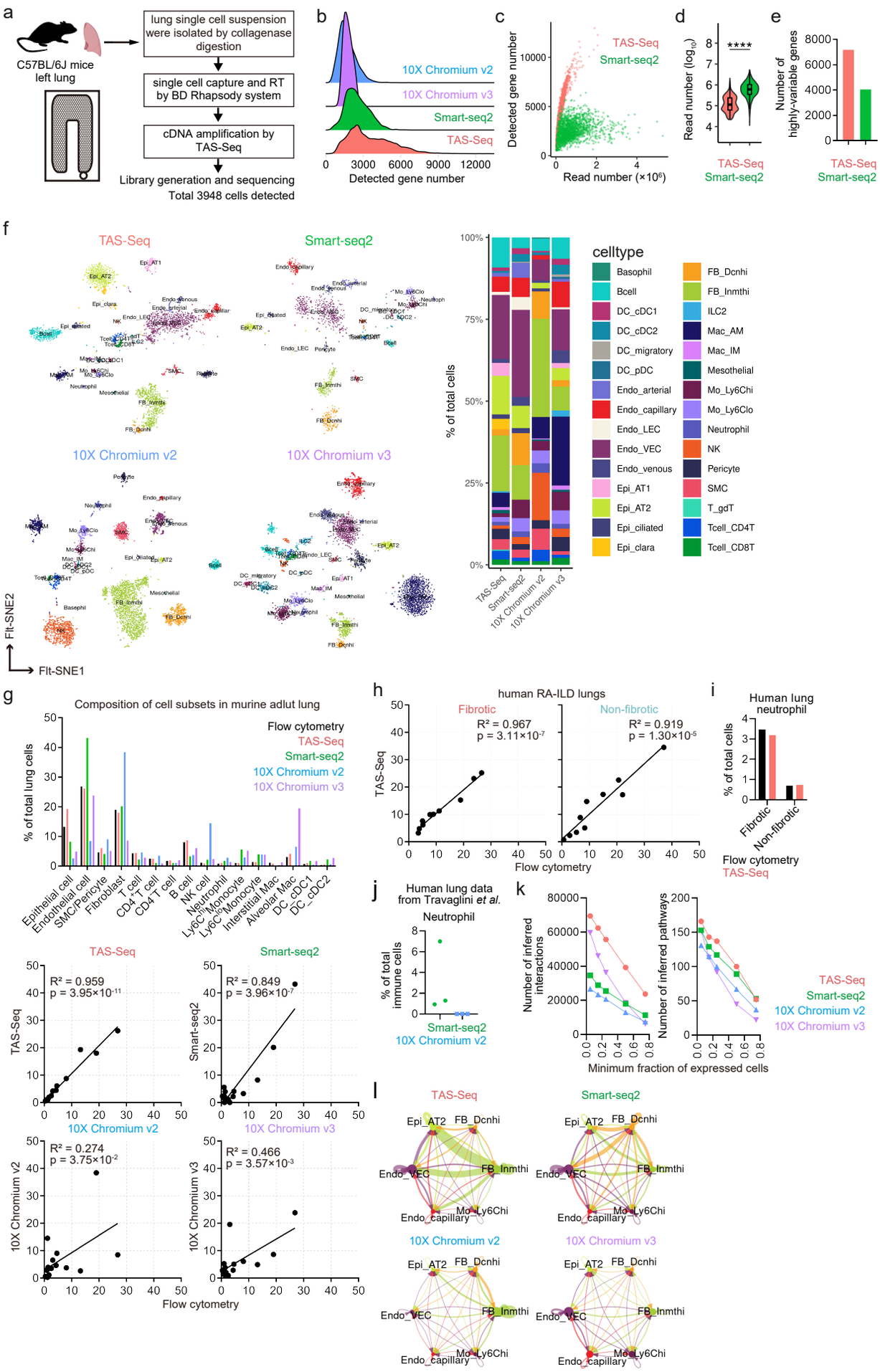
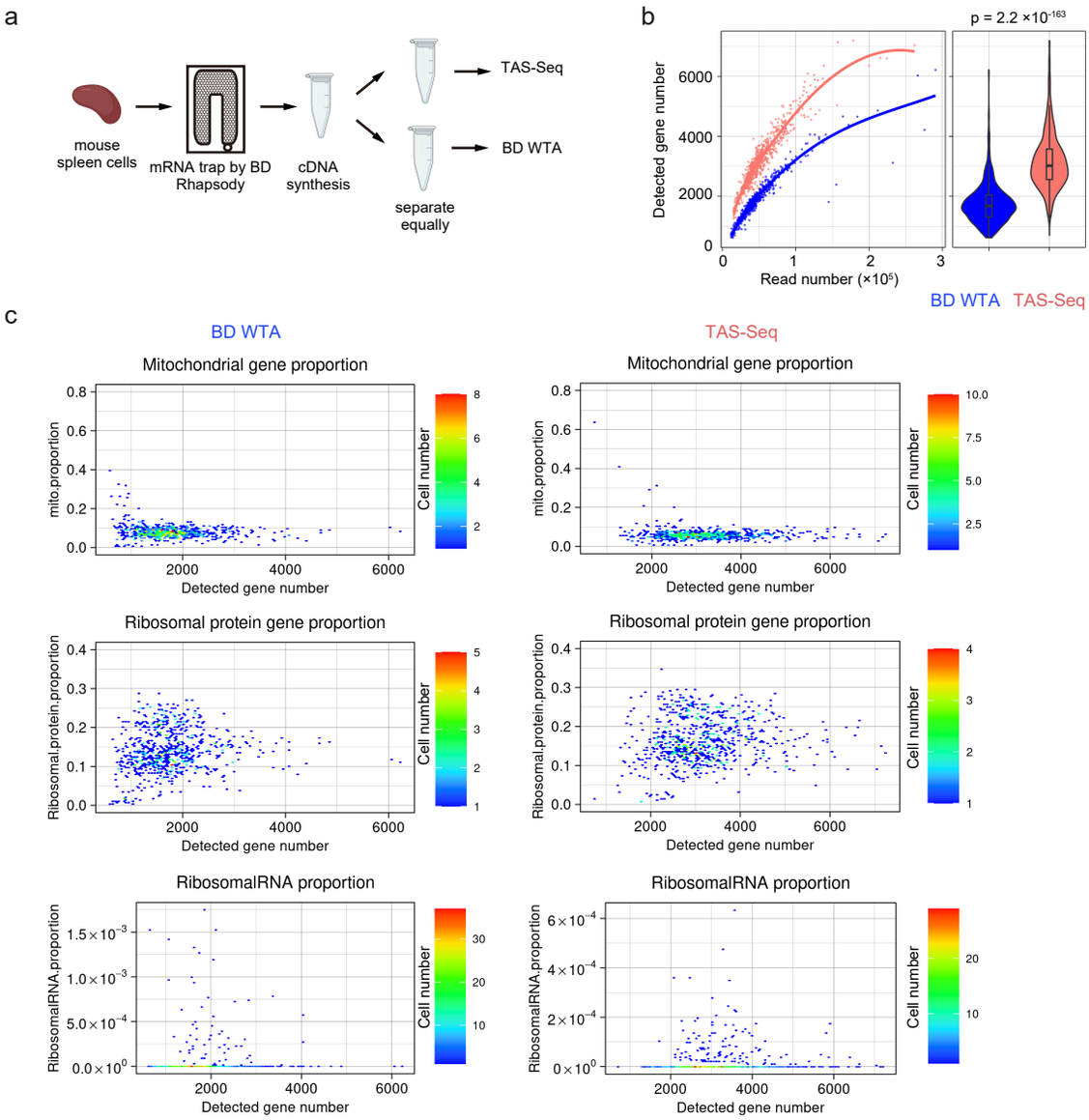


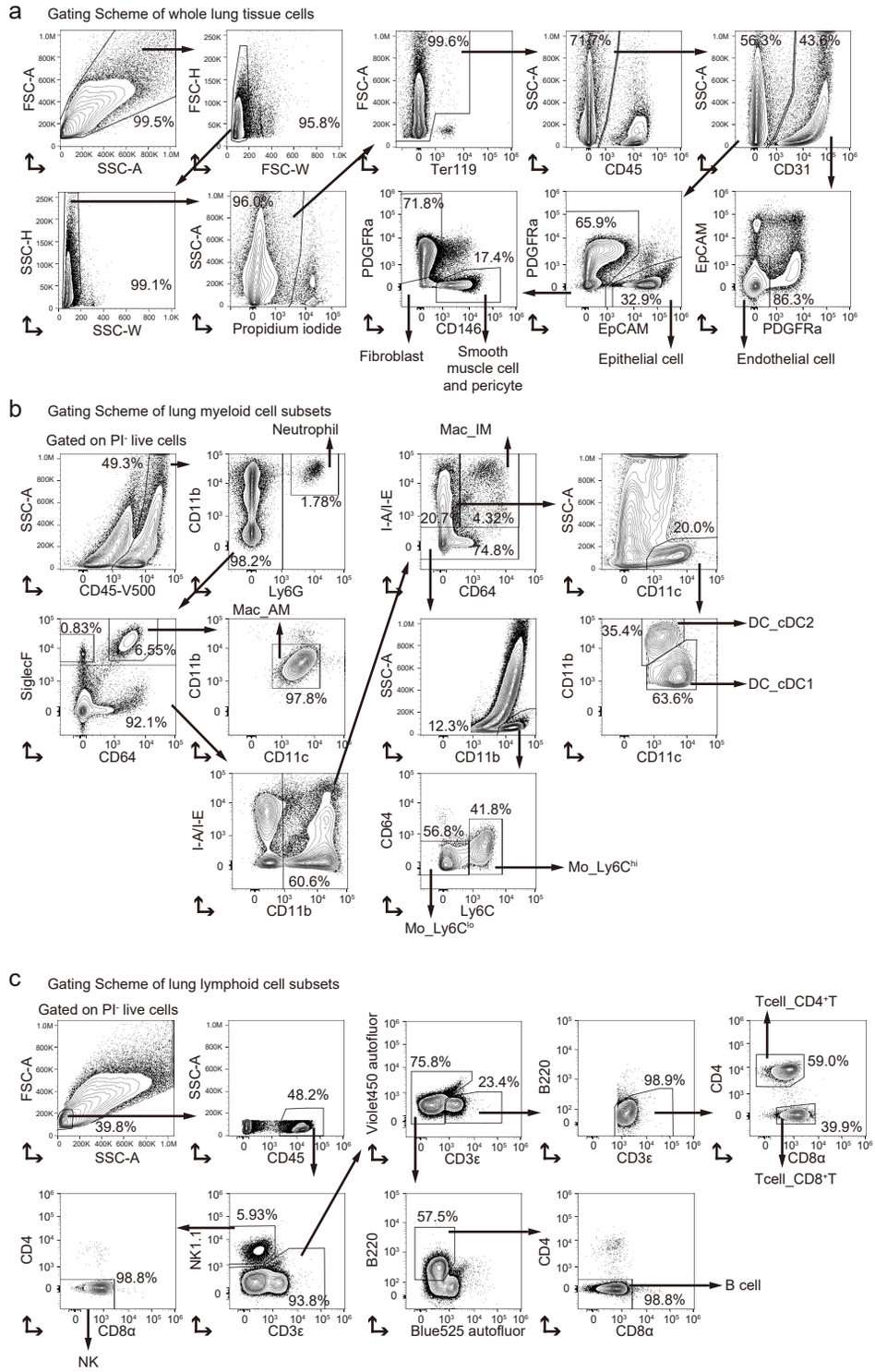
Figure 2



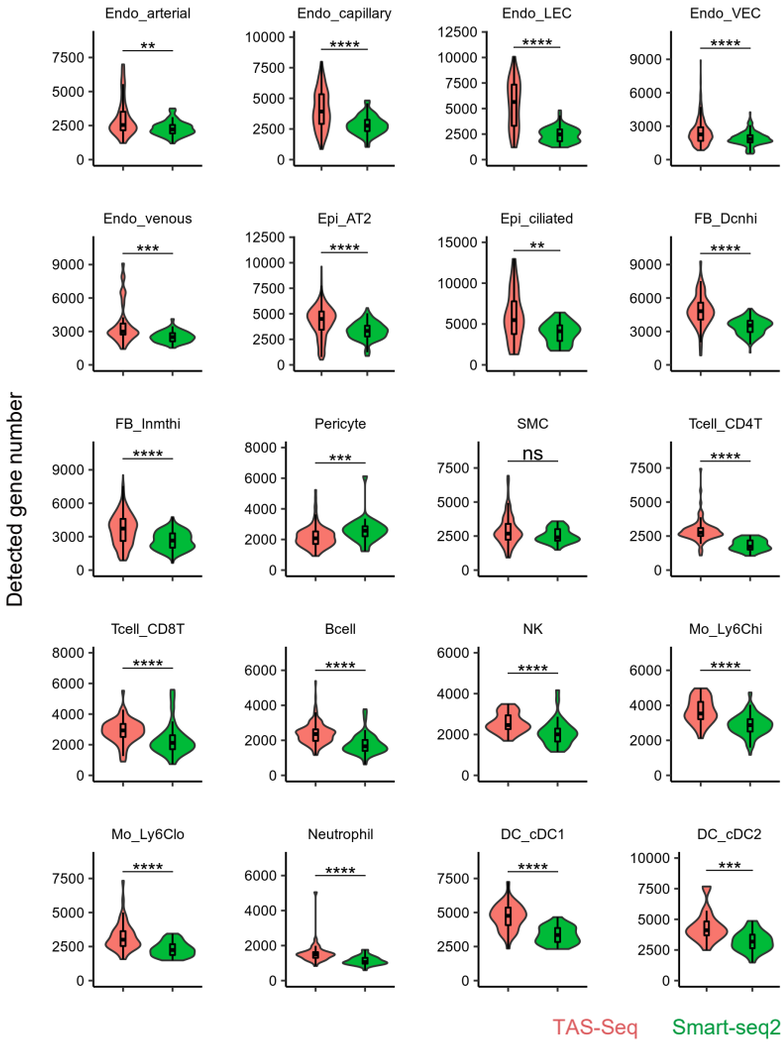
Supplementary Figure 1



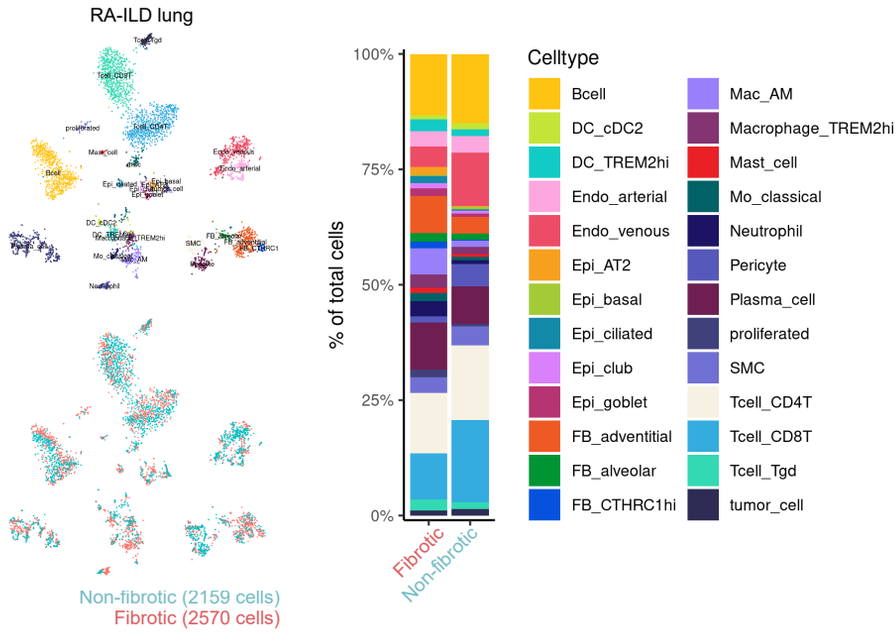
Supplementary Figure 2



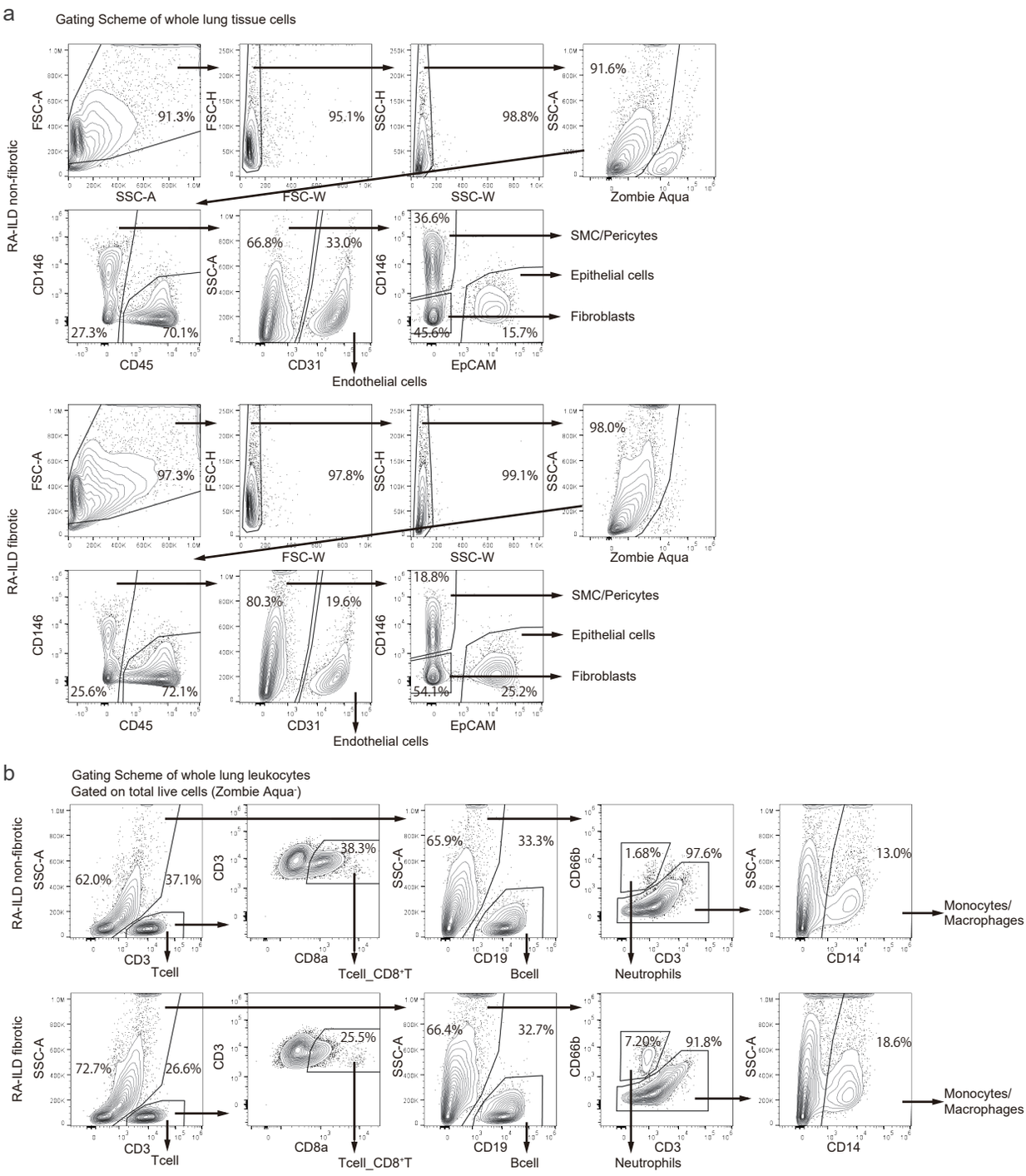
Supplementary Figure 3



Supplementary Figure 4



Supplementary Figure 5



Supplementary Figure 6