Single-cell and single-nucleus RNA-sequencing from paired normal-adenocarcinoma lung samples provides both common and discordant biological insights

Sébastien Renaut[1], Victoria Saavedra Armero[1], Dominique K. Boudreau[1], Nathalie Gaudreault[1], Patrice Desmeules[1], Sébastien Thériault[1], Patrick Mathieu[1], Philippe Joubert[1], Yohan Bossé[1,2]

1) Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval, Quebec City, Canada

2) Department of Molecular Medicine, Université Laval, Quebec City, Canada

Corresponding author
Yohan Bossé, Ph.D.
Scientific Director of the Quebec Heart and Lung Institute
Professor, Laval University
Department of Molecular Medicine
Canada Research Chair in Genomics of Heart and Lung Diseases
Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval
Pavillon Marguerite-d'Youville, Y2106
2725 chemin Sainte-Foy
Quebec City (Quebec)
Canada, G1V 4G5
Tel: 418-656-8711 ext. 3725
email: yohan.bosse@criucpq.ulaval.ca

Abstract

Whether single-cell RNA-sequencing (scRNA-seq) captures the same biological information as single-nuclei RNA-sequencing (snRNA-seq) remains uncertain and likely to be context-dependent. Herein, a head-to-head comparison was performed in matched normal-adenocarcinoma human lung samples to assess biological insights derived from scRNA-seq versus snRNA-seq and better understand the cellular transition that occurs from normal to tumoral tissue. Here, the transcriptome of 160,621 cells/nuclei was obtained. In non-tumor lung, cell type proportions varied widely between scRNA-seq and snRNA-seq with a predominance of immune cells in the former (81.5%) and epithelial cells (69.9%) in the later. Similar results were observed in adenocarcinomas, in addition to an overall increase in cell type heterogeneity and a greater prevalence of copy number variants in cells of epithelial origin, which suggests malignant assignment. The cell type transition that occurs from normal lung tissue to adenocarcinoma was often discordant whether cells or nuclei were examined. In addition, we showed that the ligand-receptor interactome landscape of lung adenocarcinoma was largely different whether cells or nuclei were evaluated. Immune cell depletion in fresh specimens partly mitigated the difference in cell type composition observed between cells and nuclei. However, the extra manipulations affected cell viability and amplified the transcriptional signatures associated with stress responses. In conclusion, research applications focussing on mapping the immune landscape of lung adenocarcinoma benefit from scRNA-seq in fresh samples, whereas snRNA-seq of frozen samples provide a low-cost alternative to profile more epithelial and cancer cells, and yield cell type proportion that more closely match tissue content.

Keywords: Single Cell, Single Nuclei, RNAseq, adenocarcinoma, lung cancer, cell type annotation, Immune cell depletion

Introduction

Single-cell transcriptomics (scRNA-seq) has the ability to inspect the cellular heterogeneity of tissue and cancer with unprecedented details, and as such provide important insights into the cellular origin and cell-specific molecular defects that play a role in disease pathogenesis[1–4]. However, given the pace at which the field is evolving, uncertainties remain with respect to the design and analysis of single-cell transcriptomic datasets in order to gain the most from priceless biological samples. Fresh biospecimens are generally prioritized for cell viability and greater yield of high-quality cells. For tissues, scRNA-seq requires disaggregating the tissue to release individual cells into a single-cell suspension. Differences in dissociation and sample preparation efficiency across cell types are known to affect RNA integrity and can skew cell type proportions. A well-known instance of dissociation bias is observed in human lung tissue, where dissociation of fresh tumor (biopsies or resected specimens) commonly results in a majority of immune cells being sequenced[5–7]. While the aforementioned cell-type dissociation bias can be partly alleviated by enriching the epithelial cell fraction using EPCAM-based cell sorting[6], single cell preparation protocols may also affect cell viability and introduce transcriptional signatures associated with dissociation and stress responses[6,8,9].

Analyzing nuclei (snRNA-seq) instead of cells has been proposed as an alternative for frozen samples and tissues that cannot be readily dissociated[10]. While cellular compositions recovered from scRNA-seq versus snRNA-seq can vary substantially[11], the transition from cell to nuclei sequencing may help to reduce the dissociation bias and transcriptional stress responses, facilitate the study of difficult-to-dissociate tissues and cell types, and allow the assessment of large cells that cannot pass through microfluidics systems. At the same time, reference databases and cell type-specific gene markers, which are readily used to annotate unknown cell populations, have been largely built from scRNA-seq datasets[4] and therefore may not be optimal for snRNA-seq. Cell types and gene expression differences between scRNA-seq and snRNA-seq have been observed in mouse kidneys[12,13] and brain[14,15] as well as in human metastatic breast cancer and neuroblastoma[11]. However, head-to-head comparisons between scRNA-seq and snRNA-seq are still scarce and to the best of our knowledge, this direct comparison has never been evaluated in the context of patient-matched normal lung and tumor tissues.

Lung cancer is highly prevalent and the number one cause of cancer mortality. It thus represents a medically valuable case study to compare the biological signal recovered through cells and nuclei sequencing. A variety of experimental designs and samples have been evaluated by scRNA-seq in patients with lung cancer. This includes lung samples enriched (e.g. FACS-sorted) for immune cells[16,17], lung tumor of mixed histological types[2,7], and non-small cell lung cancer (NSCLC) samples before and after targeted therapy[18] or immunotherapy[19]. More specifically in lung adenocarcinomas (LUAD), the most common histological subtype of lung cancer, which originates from epithelial cells that line the inside of the lungs, resected specimens or biopsies from two to eleven[2,5–7,20] patients have been evaluated, but with a very limited number of paired normal-adenocarcinoma lung samples. Compared with normal lung samples, epithelial cells from lung adenocarcinomas were characterized by a depletion of alveolar cells (AT1 and AT2)[2,6], lost cell identity and more cells annotated as mixed-lineage[5,21], higher transcriptome complexity and cell heterogeneity[6,22], patient-specific cancer cell clusters[18], transcriptional states associated with survival[20,21], and AT2 cells dedifferentiated into a stem-like state[22]. The shift in immune cells from normal to LUAD samples observed in previous studies were similarly informative. It unveiled an increase in B, plasma and T regulatory cells coupled with a decline in natural killer cells as well as reduced signatures of cytotoxicity in T cells, antigen presentation in macrophages, and inflammation in dendritic cells, which are all coherent features of an immunosuppressive tumor microenvironment[6,16]. Finally, differentially enriched ligand-receptor interactions promoting tumorigenesis were also observed between LUADs and normal tissues[6,20].

Herein, specimens derived from the same patients were tested using both scRNA-seq in fresh tissues and snRNA-seq from flash frozen tissues using the 10x Genomics® workflows. The biology captured by both methods was compared in the context of paired tumor-normal human lung samples explanted from patients that underwent surgery for lung adenocarcinoma. This study design revealed the cellular and molecular transition that occurs from normal lung to adenocarcinoma, and evaluated the commonality and discordance in the stemming biological insights gained from cells versus nuclei. In addition, we compared the same paired normal-adenocarcinoma human lung samples using an immune cell depletion protocol that alleviates the cell-type dissociation bias, with the aim of recovering a more representative biological signal.

Results

Single cell/Nucleus dataset preparation

Four patients, two tissue type (Normal/Tumor) and three experimental methods (scRNA-seq, snRNA-seq & immune-depleted scRNA-seq, hereafter labelled as *Cell*, *Nucleus* and *Immune-depleted cell*) were processed for a total of twenty-four samples. 160,621 cells/nuclei passed quality control (53,286; 57,078 and 50,257 for *Cell*, *Nucleus* and *Immune-depleted cell* datasets respectively) with a mean of 6,692 cells per sample (6,661; 7,135 and *6,282* for *Cell*, *Nucleus* and *Immune-depleted cell* datasets respectively, **Fig. 2A**) and a mean of 2,214 genes per cell (1,868; 2,309 and 2,473 genes for *Cell*, *Nucleus* and *Immune-depleted cell* datasets respectively, **Fig. 2B**). The experimental design is presented in **Fig. 1A-B**, while the clinical and cellular characteristics are detailed in **Tables S1** & **S2**, respectively.

From the 61 finest cell types annotations defined by Human Lung Cell Atlas (HLCA)[4], 35 were present in the current dataset at a frequency of >100 cells and we were able to annotate confidently 97.7% of cells at the coarsest level (*immune*, *epithelial*, *endothelial*, *stroma*, **Fig. 2C, Table S3**). This reference-based mapping and annotation approach is consistent with a marker-based approach for both the *Cell* and *Nucleus* datasets (**Fig. S1**). Nevertheless, cell type annotation scores were significantly and consistently lower (smaller fraction of annotated cells) in the *Nucleus* compared to the *Cell* dataset (two-way ANOVA, *p*-value < 2e-16), fine-level compared to high-level annotations (*p*-value < 2e-16) and Tumor compared to Normal tissue (*p*-value < 2e-16).

Cell composition differs from Nucleus in Normal lung tissue.

In **Fig. 3**, the UMAP visualisation showed that the *Cell* dataset from Normal lung tissue was largely dominated by immune cells, with 23,044 immune cells (81.5% of total, **Fig. 3A**). Conversely, the *Nucleus* dataset was dominated by epithelial cells, with 12,556 epithelial cells (69.9%, **Fig. 3B**). In addition, the *Nucleus* dataset contained a larger fraction of unclassified cells compared to the *Cell* dataset (7.3 % vs 0.1 %, Fisher Exact Test [FET], *p*-value < 2e-16).

To further refine the immune community of cells, we sub-setted only the immune cells and labelled the plots with a finer level (level 3) annotation (*Cell*, **Fig. 3C**; *Nucleus*, **Fig. 3D**). We observed that the *Cell* dataset provided a better fine-grained classification as proportionally more cells could be classified into specific cell types. To this effect, the *Nucleus* dataset contained a larger fraction of unclassified cells (41.7 % vs 0.7 %, FET, *p*-value < 2e-16).

We repeated this sub-setting approach for epithelial cells, given their primary role in the onset of lung adenocarcinoma. We observed that *Cell* samples form distinct clusters mainly composed of AT1, AT2 and multiciliated lineages (**Fig. 3E-F**). The *Nucleus* dataset, which had more than five times more epithelial cells than the *Cell* dataset (12,556 versus 2,264), contained similar cell types and mainly in similar proportions, except for a sizable fraction of unclassified cells that appeared largely scattered in the UMAPs (10.9 % unclassified in *Nucleus* versus 1.29% in *Cell*, FET, *p*-value < 2e-16, **Fig. 3E-F**).

In **Fig. 4**, we present, for each cell type (level 3 annotation), the fraction of cells originating from each patient (**Fig. 4A**), the number of cells (**Fig. 4B**) and the number of genes per cell (**Fig. 4C**). In **Fig. 4D-F**, we present the same information for the *Nucleus* dataset and this visualization confirmed that the *Nucleus* dataset has similar cellular composition, except for the over-representation of immune cells in the *Cell* dataset. Both in *Cell* and *Nucleus* datasets, epithelial cell types were dominated by AT1 first and then AT2; endothelial cell types were dominated by capillary; and stromal cell types were dominated by fibroblasts. With respect to the number of genes (transcripts) per cell (**Fig. 4 C, F**), we observed many discordant patterns between *Nucleus* and *Cell* datasets, indicating that similar cell types presented different overall transcriptional signatures based on the experimental method. For example, in the *Cell* dataset, median numbers of genes per cell were low for monocytes (635), but high for T cells (1,709), and the pattern was in the opposite direction for the *Nucleus* dataset (Monocytes = 2,729, T cells = 1,055). For their part, alveolar cells AT1 and AT2 contained 50% more genes expressed in the *Cell* dataset (AT1: 2,479 and AT2: 3,126) compared to the *Nucleus* (AT1: 1,639 and AT2: 2,004), and fibroblast two times as much (2,101 vs 1,061).

The cellular origin of tumoral cells

In **Fig. 5A**, the UMAPs showed that *Cell* sequencing samples from lung tumor tissues were largely dominated by immune cell types (20,410 immune cells vs 5,764 in *Nucleus* dataset), while in **Fig. 5B**, the *Nucleus* dataset were dominated by epithelial cells (27,362 epithelial cells in *Nucleus* vs 1,220 in *Cell* dataset). For both Cell and Nucleus datasets, cells appeared more scattered (i.e., more heterogeneous) in the tumor compared to normal lung (median *silhouette index* $_{\text{(Normal)}}$ = 0.69; median *silhouette index* $_{\text{(Tumor)}}$ = 0.53; two-way ANOVA, *p*-value < 2e-16, **Fig. S2**). This shows a suboptimal cell type assignment of heterogeneous tumor samples to the described lung cell types from the HLCA reference.

In **Fig. 6A-C**, we present, for each level 3 annotation cell type, the fraction of cells from each patient (**Fig. 6A**), the number of cells (**Fig. 6B**), the number of genes per cell (**Fig. 6C**) and in **Fig. 6E-G**, we present the same information for the *Nucleus* dataset. First, we observed that within a coarse level annotation, similar cell types and similar proportions are observed in *Cell* and *Nucleus* datasets. For example, T cells largely dominated the immune cells, fibroblasts dominated the stroma cells and endothelial cell types were relatively rare. With respect to epithelial cells, these were mainly composed of unclassified and AT1 in both *Cell* and *Nucleus* datasets, and secretory epithelial cells appeared to be mainly segregated to patient 3. However, rare cell types were much more common in the *Nucleus* than the *Cell* datasets.

To distinguish malignant and non-malignant cells, we defined a genome-wide summary score (CNV score) that relies on gene expression levels to identify gene deletion and duplication and serves as a proxy to identify cancerous aneuploid cells[23]. This score was the highest for different epithelial cell types depending whether we analysed the *Cell* dataset (rare, multiciliated lineage, AT1, unclassified, **Fig. 6D)** or the Nucleus dataset (multiciliated lineage, secretory and unclassified, **Fig. 6H**). In addition, we also noted that annotation scores were negatively correlated with CNV scores for *Cell* (r^2 = 0.11, *p*-value < 2e-16) and *Nucleus* (r^2 = 0.05, *p*-value < 2e-16) datasets (**Fig. S3**).

The cellular transition to lung adenocarcinoma

Given the known epithelial origin of lung adenocarcinoma and the role of the immune system in effectively controlling the growth of carcinoma cells, we analysed the transition in the proportions of epithelial and immune cells from normal to adenocarcinoma tissue (**Fig. 7A-B**). Alveolar Type 1, AT2 and multiciliated cells decreased in relative abundance in adenocarcinomas, and this was consistent for the *Cell* and *Nucleus* datasets. On the contrary, rare, secretory and unclassified epithelial cell types increased in abundance in adenocarcinoma tissue in a consistent manner between *Cell* and *Nucleus* datasets. For Immune cells, patterns were harder to interpret given the small number of immune cells in the *Nucleus* dataset. Nevertheless, an augmentation of B and T cell lineages in adenocarcinoma was found for both datasets, as well as a sharp drop in natural killer cells in the *Cell* dataset. For macrophages and monocytes, a discordance in the transition from normal to tumor between scRNA and snRNA was observed.

The Ligand-receptor interactome differs between Cell and Nucleus

In **Fig. 8A**, we visualised the incoming and outcoming interactions among 319 ligand-receptor interactions (cell-cell contact) for the *Cell-Normal* dataset. The number of interactions between cell types varies first according to the Cell vs. Nucleus method (two-way ANOVA, F = 90.7, p-value < 2e-16) and then the Normal vs. Tumor tissue type (F = 68.2, p-value = 3.6e-16). In **Fig. 8B**, we show an example of a typical pathway common in *Cell*, rare in *Nucleus* (Major Histocompatibility Complex-I) and its interacting genes, which is more similar between *Normal* vs *Tumor* tissue of the same experimental method (*Cell* vs *Nucleus*). An example pathway, rare in *Cell* but common in *Nucleus* (Protein Tyrosine Phosphatase Receptor Type M) and its self interacting gene is presented in **Fig. 8C**. In this case, each network shows differences according to both the experimental method and tissue.

The effect of immune depletion on Cell sequencing

In order to remove the large fraction of immune cells, we performed immune depletion in Normal and Tumor single-cell suspensions. We confirmed that the *Immune-depleted cell* dataset was enriched in epithelial cells and depleted in immune cells (**Fig. 9A-B)**. As such, both the Normal and Tumor tissues resemble the *Nucleus* dataset in the fact that they harbor a majority of epithelial cells (61.5% and 69.9% of total for the *Immune-depleted cell* and *Nucleus* dataset, respectively), yet they differ given that immune depleted cells harbor proportionally more endothelial (17.8% vs 4%) and stromal (18.4% vs 7.9%) cell types, but less immune cells (1.3% vs 13.0%). In addition, Normal tissues were largely composed of epithelial AT1 and AT2, while Tumor tissues also harbored secretory, rare and unclassified cell types, much like the *Nucleus* dataset (**Fig. 9C-D**). Finally, as we observed for the non-depleted dataset, we saw an increase in the heterogeneity from Normal to Tumor datasets (median Silhouette index for each level 3 cell type annotation: $s_{i\ \text{(Normal)}} = 0.56$, median $s_{i\ \text{(Tumor)}} = 0.2$, two-way ANOVA, p-value < 2e-16, **Fig. S2**).

Finally, we downloaded a set of 512 heat shock and stress response genes that were previously identified as affected by the scRNA-seq method[9]. Ninety four percent (482 genes) of the genes in this core dataset were also present in our current dataset, with varying levels of expression. More specifically, the percentage of cells expressing these genes was largely dependent on the method (**Fig. 9E**, two-way ANOVA, *p*-value < 2e-16). The *Immune-depleted cell* dataset showed the highest expression of the stress response genes, whereas on average a cell from the *Immune-depleted cell* dataset expressed 21% of the 482 genes, compared to 11.0% and 6.9% for the *Cell* and *Nucleus* dataset, respectively. In addition, the proportions of cells expressing this core set of stress response genes was slightly, but significantly (*p*-value = 9.7e-8) higher in Tumor than in Normal (12.4 % and 11.5 %, respectively) tissue. In a similar manner, higher mitochondrial contamination is often considered a sign of lower cell quality or viability[24] and we observed that the percentage of unique sequences (UMIs) assigned to mitochondrial genes in the raw data prior to any filtering was significantly higher (two-way ANOVA, *p*-value = 3.6e-5) in the *Immune-depleted cell* (mean = 15.2 %) and *Cell* (11.2 %) compared to the *Nucleus* (2.6%) dataset, while the tissue type (*p*-value = 0.10) had no significant effect (**Fig. S4**).

Discussion

In this study we generated a dataset of 160,621 cells/nuclei showing commonalities and discordances in biological insights derived from single-cell and single-nucleus RNA-sequencing of paired normal-adenocarcinoma human lung specimens. A distinct portrait of cellular composition was observed per experimental methods that favors scRNA-seq of fresh samples to map the immune landscape of lung adenocarcinoma. On the other hand, snRNA-seq of frozen samples surpassed the relative merits of scRNA-seq to obtain a dataset with cell type proportion that match tissue content and to provide a more cost-effective approach for research applications necessitating a higher number of epithelial and cancer cells (see **Table S4** for a summary of the benefits of each method). In these paired lung samples, we identified gene expression and cell type transitions from normal to tumoral tissue that were not always concordant whether cells or nuclei were examined. The most striking difference was the ligand-receptor interactions that varied more across methods (cells vs. nuclei) rather than tissue types (normal vs. tumor). Immune cell depletion partly alleviated the difference in cell type composition between cells and nuclei, but at the detriment of inducing a stress response. Finally, our analysis revealed that the recently proposed five-level hierarchical cell type annotation system by the Human Lung Cell Atlas[4] will require customization for assigning cell types from nuclei and tumor samples.

Despite the fact that samples originated from the same patients' specimens, scRNA-seq and snRNA-seq varied substantially in their recovered cellular compositions and transcriptional landscape, thus highlighting the considerable impact of methodology on biological inference. While it has been shown previously that cryopreservation of tissue sample (such as performed for snRNA-seq) results in a major loss of epithelial cell types and an underrepresentation of T, B, and NK lymphocytes in the single-nucleus libraries[11,13], it is not necessarily apparent which experimental method is more biologically relevant. Slyper et al.[11] have suggested to analyse both fresh and frozen tissue, but this is often unrealistic in practice. For their part, Denisenko et al.[13] indicate that the apparent discordance in the recovered cellular composition between scRNA and snRNA might be due to either an under-representation of immune cells in snRNA, or an under-representation of other cell types cells in scRNA due to incomplete dissociation. Early pioneering work in lung histology would suggest the latter, whereas cell staining and electron microscopy has revealed that the alveolar regions of normal human lungs are comprised mainly of epithelial, endothelial and interstitial cells, while immune cells (macrophages) comprised a small fraction (~5%) of all cells identified[25]. We thus conclude that in the context of lung adenocarcinoma and patient-matched normal samples, snRNA-seq provides a dataset comprising cell populations more closely matching tissue content.

In addition, we observed a decrease in cell viability in both depleted and non-depleted scRNA-seq, likely due to the longer sample preparation times at room temperature. While this could be partly alleviated by cold-activated proteases[9], it favors snRNA-seq as a experimental protocol to preserve sample integrity. Although immune depletion works well for removing immune cells and therefore might draw a more accurate representation of the lung cellular composition that is closer to snRNA-seq, it requires extra laboratory manipulations and has the adverse effect of affecting both cell viability (**Fig. S4**) and inducing a dissociation transcriptional stress response **(Fig. 9E**), as shown previously[12].

The reference-based annotation used here provides an attractive alternative to unsupervised analysis[26]. We annotated the large majority of cells/nuclei in all tissue types, methods and patients (**Fig. 2, Fig. S5**) while showing that it performed as well as a marker-based approach, at least at the coarsest annotation level (**Fig. S1**). Arguably, the confidence in this reference-based annotation approach depends on several factors. Notably, the comprehensiveness of the reference, the quality and type of query data and the level of cellular granularity required to answer the biological question of interest will dictate the best approach to use. Nevertheless, an unsupervised-marker based approach also depends on several factors such as the clustering algorithm, the gene markers used, and almost always, the expertise and subjectivity of the person annotating the dataset[27,28]. Here, annotation and mapping were done using the same analytical framework for all samples and therefore provided an objective overview of the transcriptional cellular landscape. Fortunately, we were able to use a recently published comprehensive atlas of the lung (HLCA)[4], although thorough cell atlases might not exist for all tissue types, biological conditions and demographic states[29]. The lower annotation scores observed in nuclei and tumor samples and consequently the greater number of unclassified cells, especially at the finer annotation levels suggest that these cells or nuclei have a distinct signature from the current reference cell types. A similar phenomenon was also observed in the HLCA for different disease states[4] and the authors concluded themselves that the HLCA must be viewed as a live resource that will require continuous updates in the future, including samples of diverse ethnic, clinical and experimental (e.g. snRNA-seq) backgrounds.

During the transition from normal to tumoral tissue, we identified a drop in AT1, AT2 and NK cells, concurrently with a rise in immune B and T cells, as previously identified[2,6,16]. In addition, tumoral cells showed an increased transcriptomic heterogeneity and a greater prevalence of copy number variants in epithelial cells. Similarly, it has been described that NSCLC exhibit important interpatient histologic heterogeneity and inferred origin of tumor cells[30]. Here, we showed that epithelial AT1, secretory and multiciliated lineages cell types had higher Copy Number Variants scores than AT2, which suggests malignant assignment. Yet, the distinction between these epithelial cells is not always straightforward, especially in a context of oncogenesis. Along those lines, we noted that annotation scores were negatively correlated with CNV scores which implies that cells with high CNV (likely carcinoma cells) loose their cellular identity and become harder to classify as distinct lung cell types. During the construction of the HLCA, Sikkema *et al.*[4] also noted than a significant fraction of cells from adenocarcinomas did not cluster into the specific fine level cell types. Similarly, Wang et al.[22] argued that cancer cells originate from ‘AT2-like’ cells, but also nuanced this fact and stated that these form a distinct cluster from regular AT2 cells and in fact, have a transcriptional profile closely resembling other epithelial cells. Again, a more refined and thorough reference database will help to solve these questions.

Ultimately, we hope to develop a comprehensive transcriptional resource for the identification of cell-targeted biomarkers and therapeutic targets to treat and prevent LUAD and other ailing aspects of the lung. Accordingly, this study may have clinical significance as immunotherapy is currently revolutionizing the treatment of lung cancer. Response to immune checkpoint inhibitors relies on the existing cell-cell interactions between tumor and T cells (e.g., commercial immunotherapy drugs targeting the interaction between PD-1 in tumor cells and PD-L1 in T cells)[31] and identifying accurate biomarkers of response to immunotherapy is a major challenge in the field of lung cancer[32]. Consequently, this seems like a clinical problem where single-cell genomics can provide a solution. However, here we demonstrated that the ligand-receptor interactome landscape of lung adenocarcinoma is largely different whether cells or nuclei are evaluated. This may lead to conflicting prediction response to these novel immunotherapy agents. Accordingly, at least in the context of lung cancer, the choice between scRNA-seq and snRNA-seq has important implications. Our results favor scRNA-seq on fresh samples to provide a more comprehensive portray and granularity of the immune cells diversity. On the other hand, it may not be representative of the true cellular community, and lead to fewer difficult-to-dissociate tumor cells to assess relevant tumor-immune interactions. More studies will be needed to assess the best methods as well as to overcome other barriers to move single-cell genomics into the clinical setting[33].

Materials and methods

Patients and samples

Lung samples were collected from four patients that underwent curative intent primary lung cancer surgery at the *Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval* (IUCPQ-UL) in 2021-2023, henceforth referred to patient 1, 2, 3 and 4. The four patients were self-reported white French Canadian (European ancestry) with no prior chemotherapy and/or radiation therapy, and all patients were between the age of 59 and 69, former smokers with adenocarcinomas (See **Fig. 1** for overview of experimental design, and **Table S1** for detailed clinical characteristics of patients).

Following surgery, the explanted lobes were immediately transferred to the pathology department. For each patient, two □1 cm^3 fresh tumor samples and two □1 cm^3 non-tumor (normal) lung samples located distant from the tumor were harvested. The first set of tumor/non-tumor samples was transferred in dedicated tubes containing ice-cold RPMI (ThermoFisher, Cat. 11875093) for immediate cell dissociation and single-cells RNA sequencing (scRNA-seq) experiment. The second set of tumor/non-tumor samples was transferred in dedicated tubes, immediately snap-frozen in liquid nitrogen and stored at -80°C until the day of the single-nucleus RNA sequencing (snRNA-seq) experiment. A histologic slide of each specimen was stained (H&E) and reviewed by a pathologist. Staging was performed using the 8th edition of the TNM Classification of Malignant Tumours[34]. Lung tissue samples were obtained in accordance with the Institutional Review Board guidelines. All patients provided written informed consent, and the ethics committee of the IUCPQ-UL approved the study.

Sample preparation for scRNA-seq

Immediately after collection, the weight of each sample was recorded. Samples were transferred to 6-well cell culture plates, washed twice with 3 mL ice-cold PBS (Thermo Fisher, cat. 10010023) to remove excess blood and transferred to a 5 mL glass beaker. Using a 1 mL syringe and 25G needle, 300 µL of Enzyme dissociation mix was injected in the tissue followed by mechanical mincing into small fragments (<1mm³) using spring scissors for 2 minutes. Samples were then transferred to 50 mL Falcon tubes containing 5,7 mL of Enzyme dissociation mix and pipette mixed 5 times using wide bore 1 mL tips. The enzymatic digestion was performed at 37°C, using a Vari-Mix™ test tube rocker at max speed for 35 minutes. Samples were pipette mixed 20 times after 15 and 30 minutes using wide bore 1 mL tips. Enzyme dissociation mix contained: Pronase 1250 µg/mL (Sigma Aldrich, cat. 10165921001), Elastase 18.4 µg/ml (Worthington Biochemical, cat. LS006363), DNase I 100 µg/mL (Sigma Aldrich, cat. 11284932001), Dispase 100 µg/mL (Worthington Biochemical, cat. LS02100), Collagenase A 1500 µg/mL (Sigma Aldrich, cat.10103578001) and Collagenase IV 100 µg/mL (Worthington Biochemical, cat. LS 004186) in HBSS (Thermo Fisher, cat. 14170112). Enzymatic digestion was stopped by adding 1.5 mL of fetal bovine serum (FBS, ThermoFisher, cat. A3840301) followed by pipette mix 5 times using wide bore 1 mL tips. Dissociated cells were filtered through a 70 µm strainer and washed with 7.5 mL ice-cold PBS. Cells were then pelleted at 400g, 4°C for five minutes and supernatant was removed. Three cycles of red blood cells removal were performed as follow: cell pellet resuspended by manual agitation in 500 µL of ACK Lysis Buffer (ThermoFisher, cat. A1049201) and incubated on ice one minute. One mL of ice-cold PBS was added and cells were centrifuged at 400g, 4°C for two minutes and the supernatant was removed. The final pellet was resuspended in 500 µL ice-cold-PBS containing 0.04% Bovine Serum Albumin (BSA, Sigma Aldrich Cat. A7284) and 10% FBS. Cell suspensions were successively passed through 100 µm, 70 µm and 40 µm strainer using quick spin to reach 400g to filtrate each sample. Samples were transferred to 2.0 mL low binding tubes and kept at 4°C. Cell count and viability were performed using a 1:1 mix of cell suspension, Trypan blue (ThermoFisher, cat. 15250061), haemocytometer and conventional light microscopy. Cells suspensions meeting the following criteria were accepted for scRNA-seq library preparation: absence of aggregated cells, a viability >80%, and a total cell count between 400 and 1200 cells/µL. 1x10^5 cells were transferred to a low binding 2 mL tube and kept at 4°C (non-depleted fraction). The remaining cells (from 2 to 5 x10^6 cells) were submitted to CD45 immune cell depletion protocol (single cells depleted fraction) as described below. The characteristics of the lung specimen and the single cell suspension for each sample are given in **Table S2**.

CD45 immune cell depletion

Cells (from 2 to 5 x10^6 cells) were centrifuged at 300g, 4°C, 10 minutes. The supernatant was removed and the cell pellet was resuspended in 80 µL MACS buffer (0.5% BSA, 2 mM EDTA pH 8.0 in PBS) previously degassed for 1 hour at room temperature. Twenty µL of CD45 microbeads (Miltenyi Cat. 130-045-801) were added and sample was incubated 15 minutes at 4°C followed by addition of 1 mL MACS buffer and centrifugation 300g, 10 minutes at room temperature. Supernatant was removed and pellet resuspended in 2-steps 100 µL + 400 µL MACS buffer. The total volume (500 µL) was applied to a LS Positive Selection Column (Miltenyi Cat. 130-042-401) previously rinsed with 3 mL MACS buffer and installed on a MidiMACS magnetic Separator with a collection tube. Column was rinsed with 3 X 3 mL MACS buffer and all volumes (9.5 mL) were collected which contained the CD45-negative fraction. CD45-negative cells were centrifuged 300g, 10 minutes at room temperature followed by supernatant removal. Cells were washed twice with 1 mL PBS followed by centrifugation at 300g, 10 minutes after each wash. Cells were finally resuspended in 100 µL BSA 0.04%, 10% FBS in PBS and kept at 4°C. Cell count and viability were performed using a 1:1 mix of cell suspension, Trypan blue, haemocytometer and conventional light microscopy. Cells suspensions meeting the following criteria were accepted for scRNA-seq library preparation: absence of aggregated cells, a viability >80%, and a total cell count between 400 and 1200 cells/µL.

Sample preparation for snRNA-seq

Nuclei suspension was prepared from ~30 mg snap frozen tissue using Chromium Nuclei Isolation Kit as per manufacturer protocol (10x Genomics Cat. 1000494). Nuclei count and integrity were performed using a 1:1 mix of nuclei suspension and methylene blue 0.25% (Ricca Chemical, Cat. 48504), haemocytometer and conventional light microscopy. Nuclei suspensions meeting the following criteria were accepted for snRNA-seq library preparation: absence of aggregated nuclei, nuclei with circular shape and intact membrane (without blebbing) >80%, and a total nucleus count between 400 and 1200 nuclei/µL. Nuclei suspension were kept at 4°C until proceeding with 10x Genomics snRNA-Seq library preparation protocol.

10x Genomics sn/scRNA-seq library preparation

For each sample, approximatively 15,000 nuclei or cells were loaded into each channel of a Chromium Next Gel Beads-in-emulsion (GEM) Chip G (10x Genomics Cat. 1000127) as per manufacturer instruction for GEM generation and barcoding. Given the cell capture efficiency of around 65%, 10,000 cells per library were therefore expected. The Chip was run on the Chromium Controller, GEMs were aspirated and transferred to a strip tube for cDNA synthesis, cDNA amplification and library construction using Chromium Next GEM single-cell 3’ Library Kit v3.1 (10x Genomics Cat. 1000128) and Single Index Kit T Set A (10x Genomics Cat. 2000240) as per manufacturer instruction. The library average fragment size and quantification was performed using Agilent Bioanalyzer High Sensitivity DNA kit (Agilent Cat. 5067-4626) and a final concentration determination was performed using NEBNext® Library Quant Kit for Illumina (New England Biolabs Cat. E7630) prior to library sequencing.

Next generation sequencing

Libraries were individually diluted to 10 nM, pooled and sequenced on an Illumina NextSeq2000 system following manufacturer's recommendations. Sequencing was realized on a P3 (100 cycles) cartridge, aiming for 200 to 500 million reads per library (sample). Run parameters for paired-end sequencing were as follows: read 1, 28 nucleotides; read 2, 91 nucleotides; index 1, 8 nucleotides; and index 2, 0 nucleotide.

Single cell/nucleus data preparation

Demultiplexing, alignment and transcript counting was performed using the *Cellranger* software (v7.1.0, 10x Genomics) on our local server (Lenovo ThinkSystem SR650, 40 cores and 384GB RAM). The BCL files from the Illumina sequencing run were first demultiplexed into FASTQ files using the *cellranger mkfastq* command. Read alignment and UMI counting were then executed with the *cellranger count* command (see alignment and cell statistics in **Table S5**). We used GRCh38 as the reference transcriptome available on Gencode, release 43 (GRCh38.p13).

Data quality control

The most up-to-date bioinformatics procedure defined by the R (v4.3.0)[35] library *Seurat* (v4.3.0)[24] was used to create an object for each sample and calculate values for *nCount* (number of Unique Molecular Identifiers [UMI] per cell), *nFeatures* (number of genes expressed per cell) and *percent.mt* (fraction of UMIs aligning to mitochondrial genes) parameters. Using the R library *scuttle* (v1.10.1)[36], we determined outlier values for *nCount*, *nFeatures* and *percent.mt* based on the median absolute deviation and sub-setted each sample accordingly. Note that for the *percent.mt* parameter, if necessary, we further capped this outlier value at twenty-five percent per sample.

For each sample, we then performed normalization and variance stabilization using the function *SCTransform*, which also has the benefit to regress out the *percent.mt* effect from the underlying count data. Then, using the R library *DoubletFinder* (v2.0.3)[37], we identified and removed doublets (assuming a five percent doublet rate), which occur when multiple cells are captured into a single oil droplet during the GEM generation.

Reference-based cell type annotation and mapping

On each of these curated samples, cellular annotation was performed using the R library *Azimuth* (v0.4.6)[26] and the most recent version of the Human Lung Cancer Atlas (HLCA v2)[4]. Note that in the subsequent methodology, *cell* annotation refers to the annotation of a uniquely barcoded GEM sample stemming from either a scRNA-seq or a snRNA-seq dataset.

The HLCA is a comprehensive and curated reference dataset constructed using a diverse set of 107 healthy lung samples (584,444 cells) and which allows to identify the transcriptional signature of 61 hierarchical cell types, from the coarsest possible annotations (level 1: *Immune*, *Epithelial*, *Endothelial* and *Stroma*), recursively broken down into finer levels (levels 2-5). In addition, this reference-based mapping approach allows to robustly and sensitively compare samples of broad cellular compositions, while also identifying specific and rare cell populations[24,26,38]

Specifically, for each sample (query), the algorithmic approach first identifies anchors between the reference and query (that is, pairs of cells from each dataset that are contained within each other's neighborhoods) and uses these anchors to integrate the query dataset onto the reference. Then, the embeddings of the query data onto the reference Principal Components (50 PCs) are calculated and visualised directly onto the reference two-dimensional Uniform Manifold Approximation and Projection (UMAP). Finally, annotation scores [0:1], which reflect the confidence in the annotation, were used to label cell types, whereas cells with annotation scores < 0.5 were labelled as *unclassified*.

Copy number variations analysis

For each scRNA-seq and snRNA-seq tumor sample, we performed an analysis of Copy-Number Variants (CNVs) in order to identify malignant aneuploid cells based on the premise that gene CNVs can be identified using the difference between the mean log expression level of non-cancerous reference cells (here immune cells) and the log gene expression level of a cell of interest. This was performed using the R library *infercnv* (v1.17.0)[23] and a general index (CNV score) for each cell was then defined as the mean sum of square of scaled [-1;+1] standardized log fold-change values.

Biological dataset comparisons

We integrated twenty-four samples into six different datasets (*Cell-Normal*, *Nucleus-Normal*, *Cell-Tumor*, *Nucleus-Tumor*, *Immune-depleted cell-Normal*, *Immune-depleted cell-Tumor*), in order to quantify biological similarities and differences among datasets (see **Fig. 1F-G** for summary of comparisons and accompanying figures). Given that the same reference dimensionality reduction (PCA) and visualisation space (UMAP) was used for each sample, we could simply merge expression data, metadata and projections into objects that accounts for technical variation among sample in order to quantify patterns. For each individual cell, we also calculated a Silhouette index[39] to evaluate the goodness of fit of the clustering, whereas the index is calculated from the UMAP embeddings and the clusters correspond to specific cell type (level 3) annotations. We then tested the effect of the experimental method and tissue type on the Silhouette index using a two-way Analysis of Variance (ANOVA).

Ligand-receptor analysis

In order to infer and visualise the intercellular communication among cell populations, we used the R library *cellchat* (v 1.6.1)[40]. We quantified the cell-cell interaction pathways in normal and tumor tissue (*cell* and *nucleus* dataset) to describe the cellular transition during oncogenesis and quantify how the experimental method and tissue type affected the results. We limited this analysis to level 3 annotation and excluded infrequent cell types (<500 cells in total) and cells that were unclassified at the level 3 annotation. We quantified the number of interactions from and to each cell type and tested the effect of the experimental method and tissue type using a two-way ANOVA.

Stress-related genes

To quantify the effect of our *Cell*, *Nucleus* and *Immune Depleted Cell* experimental methods on the overall stress responses of the cell populations, we analysed the expression pattern of a core set of 512 heat shock and stress response genes that were previously identified to be affected by the scRNA-seq sample preparation method[9]. We quantified the proportions of cells that expressed these genes for each sample and tested the effect of the experimental method, tissue type and patient using a two-way ANOVA.

Supplementary Information

Authors' contributions

PD, ST, PM, PJ and YB conceived the study. PD and PJ oversaw the sample pathology. SR and YB wrote the manuscript. VA, DB, NG conducted the single-cell experiments and sequencing. SR analyzed the data. All authors read and approved the final manuscript.

Funding

This work was supported by the IUCPQ Foundation and a generous donation from Mr. Normand Lord, President & CEO, St-Augustin Canada Electric.

Data availability statement

The datasets generated by *Cellranger* will be available as open-access downloadable files on Zenodo upon acceptance (zenodo.org/records/10144050). All analytical codes used to produce the results of this study will be made available at https://github.com/Yohan-Bosse-Lab/scRNA

Acknowledgments

The authors would like to thank the research staff at the IUCPQ biobank for their valuable assistance. P.M. is the recipient of the Joseph C. Edwards Foundation granted to Université Laval. P.J. is the recipient of a Junior 2 Clinical Research Scholar award from the Fonds de recherche Québec - Santé (FRQS). Y.B. holds a Canada Research Chair in Genomics of Heart and Lung Diseases

References

1. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).
2. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).
3. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
4. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
5. Laughney, A. M. *et al.* Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nat. Med.* **26**, 259–269 (2020).
6. Sinjab, A. *et al.* Resolving the Spatial and Cellular Architecture of Lung Adenocarcinoma by Multiregion Single-Cell Sequencing. *Cancer Discov.* **11**, 2506–2523 (2021).
7. Zilionis, R. *et al.* Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. *Immunity* **50**, 1317-1334.e10 (2019).
8. van den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
9. O’Flanagan, C. H. *et al.* Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biol.* **20**, 210 (2019).
10. Krishnaswami, S. R. *et al.* Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.* **11**, 499–524 (2016).

11. Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).

12. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *J. Am. Soc. Nephrol. JASN* **30**, 23–32 (2019).

13. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).

14. Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PloS One* **13**, e0209648 (2018).

15. Lake, B. B. *et al.* A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci. Rep.* **7**, 6031 (2017).

16. Leader, A. M. *et al.* Single-cell analysis of human non-small cell lung cancer lesions refines tumor classification and patient stratification. *Cancer Cell* **39**, 1594-1609.e12 (2021).

17. Guo, X. *et al.* Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* **24**, 978–985 (2018).

18. Maynard, A. *et al.* Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. *Cell* **182**, 1232-1251.e22 (2020).

19. Liu, B. *et al.* Temporal single-cell tracing reveals clonal revival and expansion of precursor exhausted T cells during anti-PD-1 therapy in lung cancer. *Nat. Cancer* **3**, 108–121 (2022).

20. Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* **11**, 2285 (2020).

21. Marjanovic, N. D. *et al.* Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* **38**, 229-246.e13 (2020).

22. Wang, Z. *et al*. Deciphering cell lineage specification of human lung adenocarcinoma with single-cell RNA sequencing. *Nat. Commun.* **12**, 6500 (2021).

23. Tickle, T., Tirosh, I., Georgescu, C., Brown, M. & Haas, B. *inferCNV of the Trinity CTAT Project*. (Klarman Cell Observatory, Broad Institute of MIT and Harvard, 2019).

24. Hao, Y. *et al*. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

25. Crapo, J. D., Barry, B. E., Gehr, P., Bachofen, M. & Weibel, E. R. Cell Number and Cell Characteristics of the Normal Human Lung.

26. Butler, A., Darby, C., Hao, Y., Hoffman, P. & Satija, R. *Azimuth: A Shiny App Demonstrating a Query-Reference Mapping Algorithm for Single-Cell Data*. (2022).

27. Xie, B., Jiang, Q., Mora, A. & Li, X. Automatic cell type identification methods for single-cell RNA sequencing. *Comput. Struct. Biotechnol. J.* **19**, 5874–5887 (2021).

28. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

29. Snyder, M. P. *et al.* The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).

30. Chen, Z., Fillmore, C. M., Hammerman, P. S., Kim, C. F. & Wong, K.-K. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat. Rev. Cancer* **14**, 535–546 (2014).

31. Garon, E. B. *et al.* Pembrolizumab for the Treatment of Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **372**, 2018–2028 (2015).

32. Mino-Kenudson, M. *et al.* Predictive Biomarkers for Immunotherapy in Lung Cancer: Perspective From the International Association for the Study of Lung Cancer Pathology Committee. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **17**, 1335–1354 (2022).

33. Lim, J. *et al.* Transitioning single-cell genomics into the clinic. *Nat. Rev. Genet.* **24**, 573–584 (2023).

34. Brierley, J. D., Gospodarowicz, M. K. & Wittekind, C. *TNM Classification of Malignant Tumours*. (John Wiley & Sons, 2017).

35. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2023).

36. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).

37. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**, 329-337.e4 (2019).

38. Stuart, T. *et al*. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

39. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

40. Jin, S. *CellChat: Inference and analysis of cell-cell communication from single-cell and spatial transcriptomics data*. (2023).

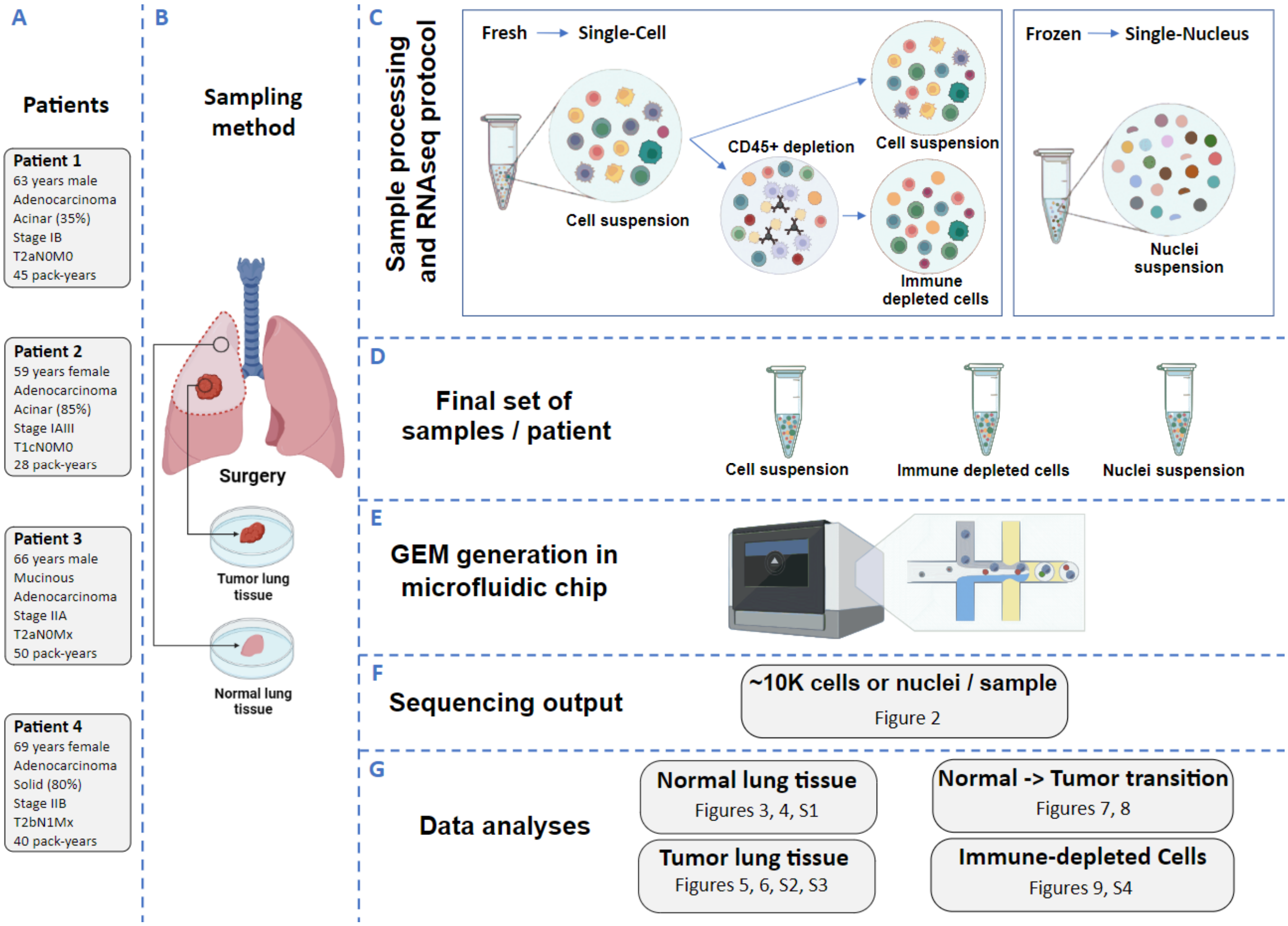

Figure 1 | Overview of the experimental design. For each patient (**A**), a tumor specimen and a normal (non-malignant) lung specimen harvested from a site distant from the tumor were resected (**B**). The research specimens were immediately divided into smaller fragments. For both normal and tumor lung specimens, a fragment was frozen in liquid nitrogen and stored at -80°C until further processing for snRNA-seq. For fresh specimens, the fragments proceeded directly to dissociation into single-cell suspensions. A subsample of the dissociation mix underwent immune cell depletion (**C**). The final set of samples (**D**) were then loaded in wells of the microfluidic chip (**E**) in order to generate the transcriptome of approximately 10,000 cells or nuclei per sample (**F**). Dataset comparisons performed with accompanying figures (**G**).

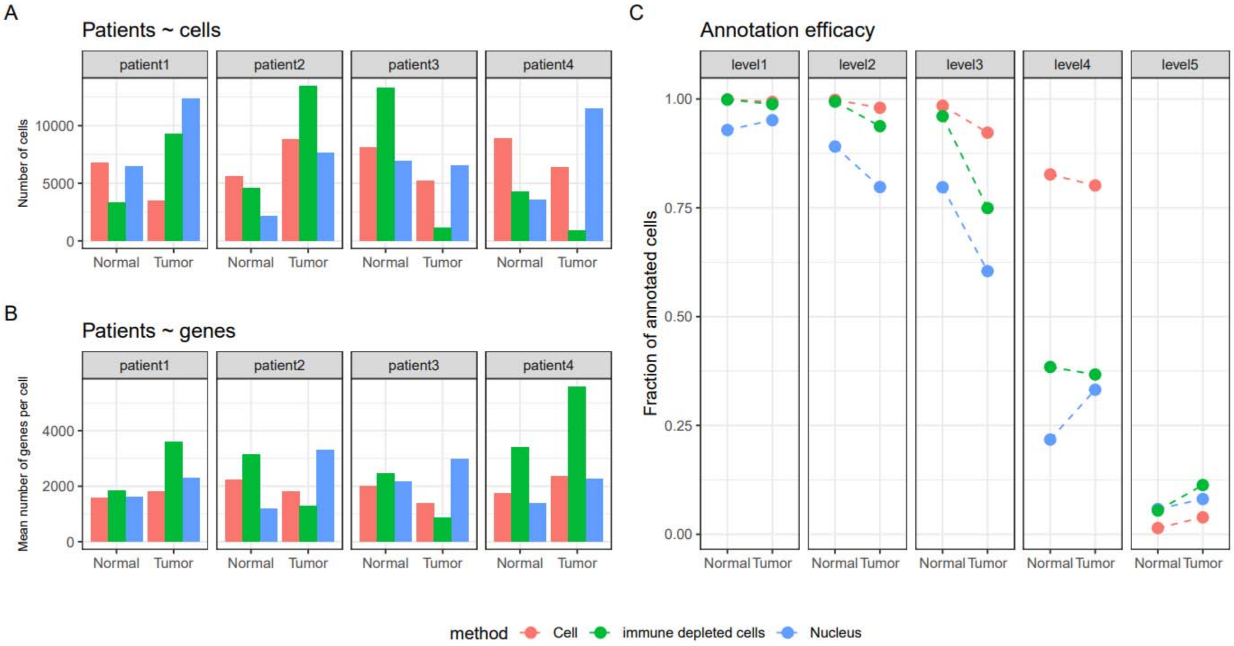

Figure 2 | Overview of the 160,621 cells/nuclei that passed quality control obtained from lung tumors and distal normal lung samples. A. Number of cells retained after quality control for each patient, each experimental method (*Cell*, *Nucleus*, *Immune-depleted cell*) and tissue type (*Normal*, *Tumor*). **B.** Mean number of genes per cell, per patient, method and tissue type. **C.** The fraction of annotated cells for each of the five-level HLCA hierarchical cell annotation reference framework, per method and tissue type.

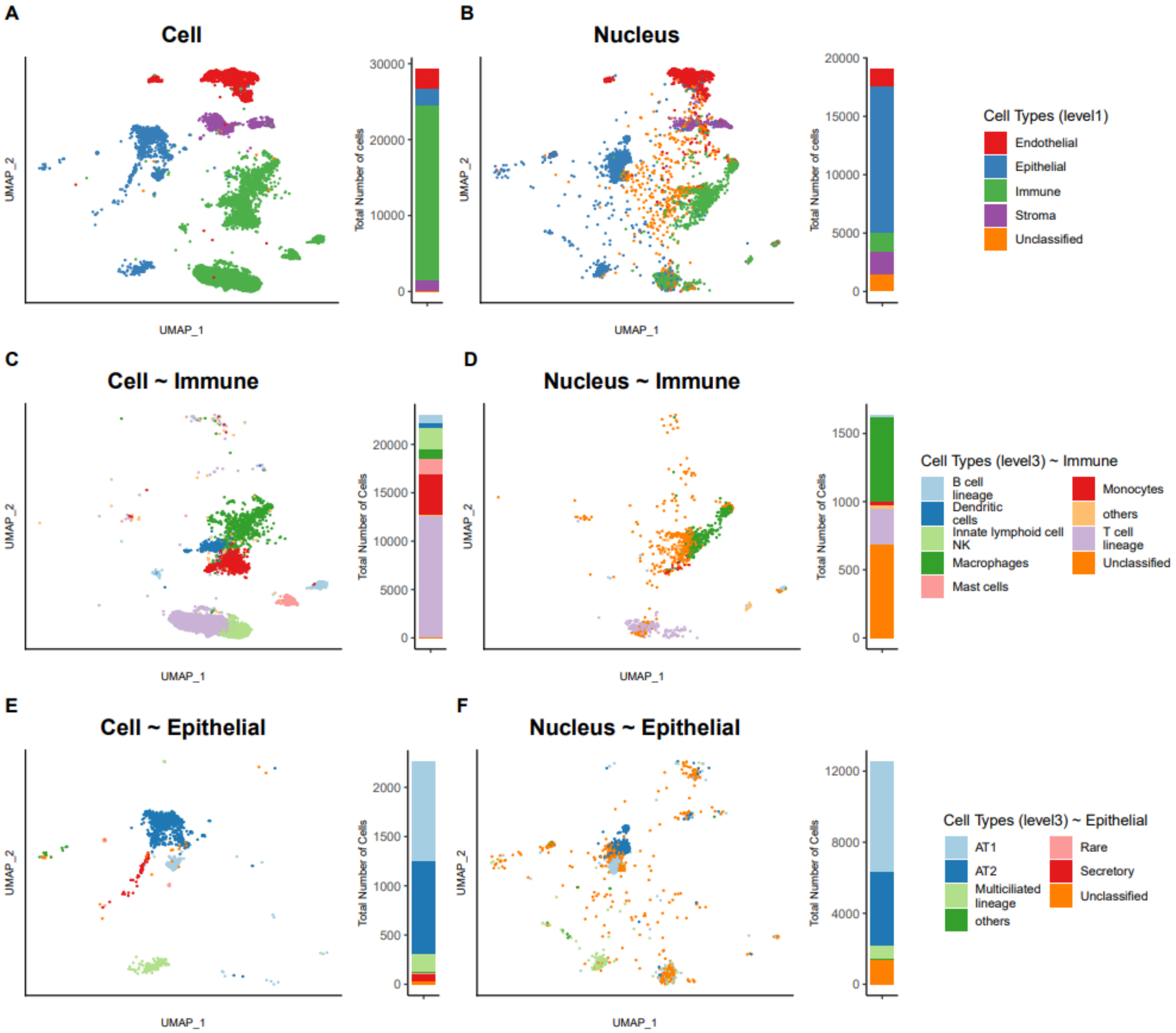

Figure 3 | UMAP representations and cell types annotations (Normal tissue) for *Cell* (**A**) and *Nucleus* (**B**) datasets with general cell types (level 1) annotation. Finer-grained annotation (level 3) for the subset of immune cells (**C**) or nuclei (**D**) and for the subset of epithelial cells (**E**) or nuclei (**F**). To the right of each UMAP, stacked bar plots indicate the proportion of each cell type in the specific dataset. Cell types present at < 1% are labelled as others.

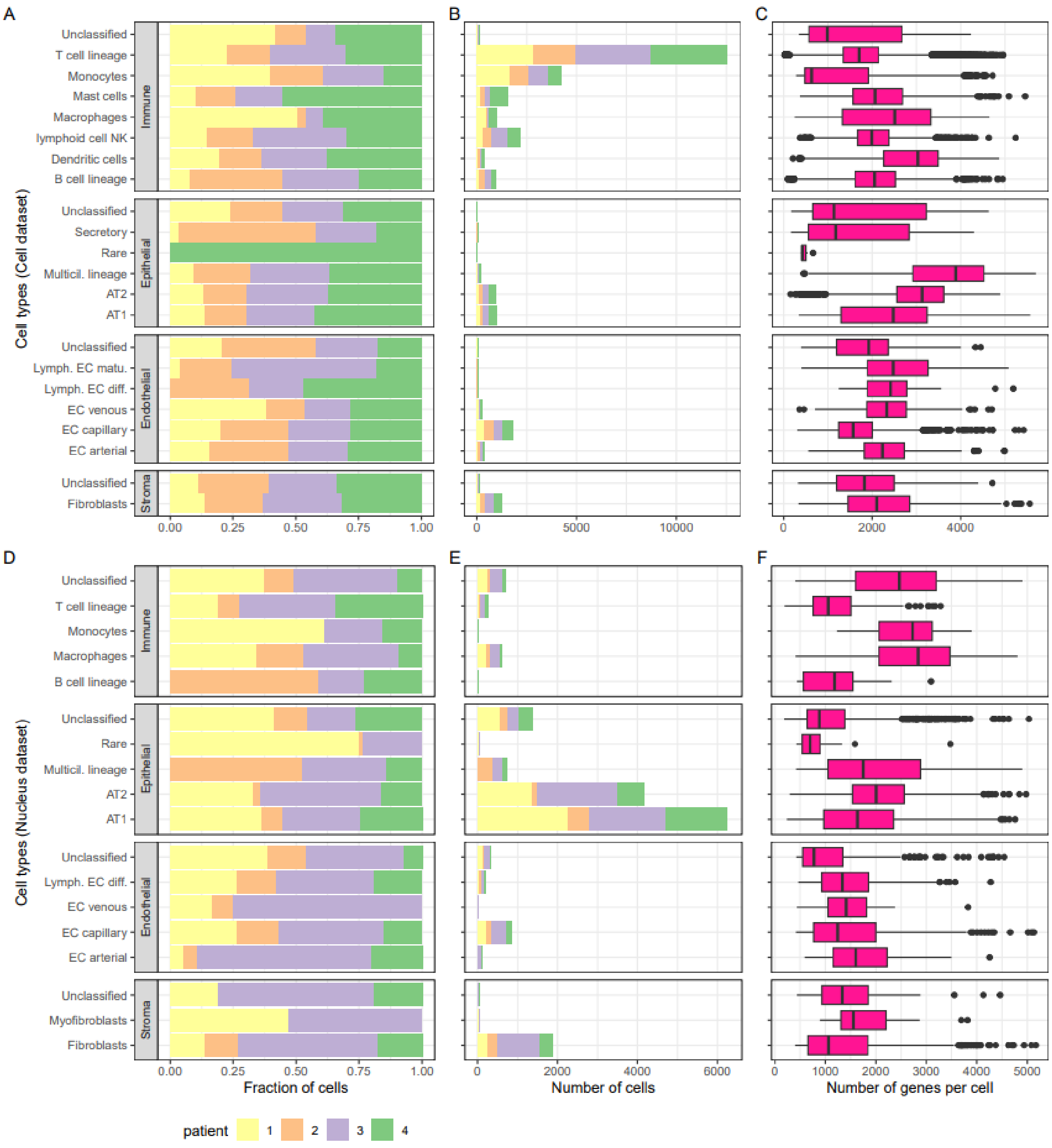

Figure 4 | Cell types characteristics (Normal tissue). For each of the four coarse (level 1) cell types annotation (*Immune*, *Epithelial*, *Endothelial*, *Stroma*) further refined into finer categories (level 3): the fraction of cells (**A**: *Cell* dataset, **D**: *Nucleus*) and the number of cells (**B**: *Cell*, **E**: *Nucleus*) originating from each patient. Box plots of the number of genes expressed per cell (**C**: *Cell*, **F**: *Nucleus*), with plot center, box and whiskers corresponding to median, IQR and 1.5□×□IQR, respectively. Note that only cell types with > 20 cells were retained for clarity in this visual representation.

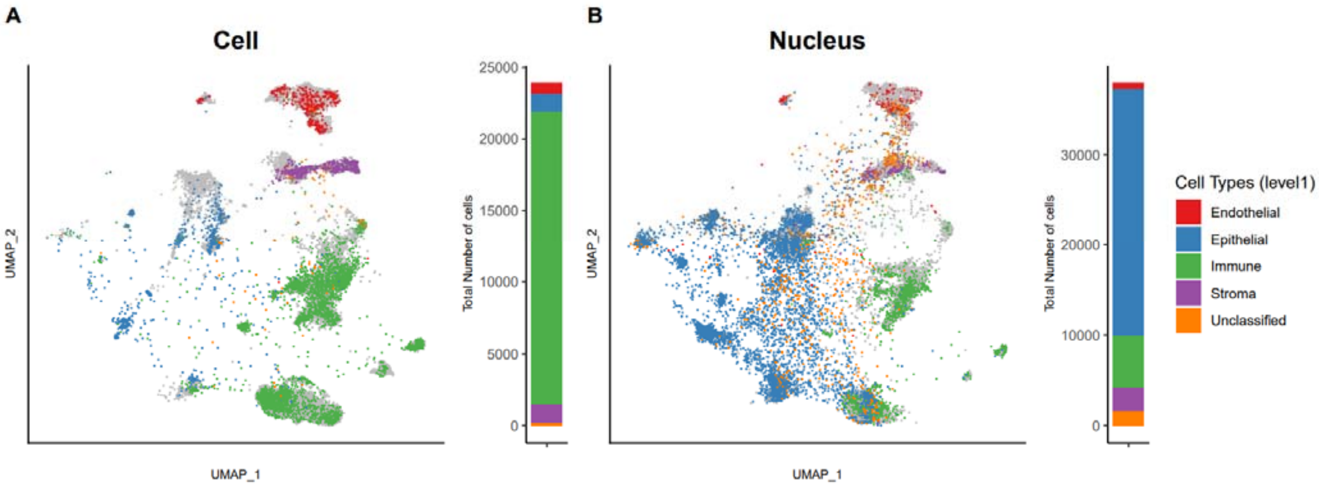

Figure 5 | UMAP representations and cell types annotations (Tumor tissue) for *Cell* (**A**) and *Nucleus* (**B**) datasets with general cell types (level 1) annotation. Tumor samples are overlaid on top of Normal samples (in gray). To the right of each UMAP, stacked bar plots indicate the proportion of each cell type in the specific dataset.

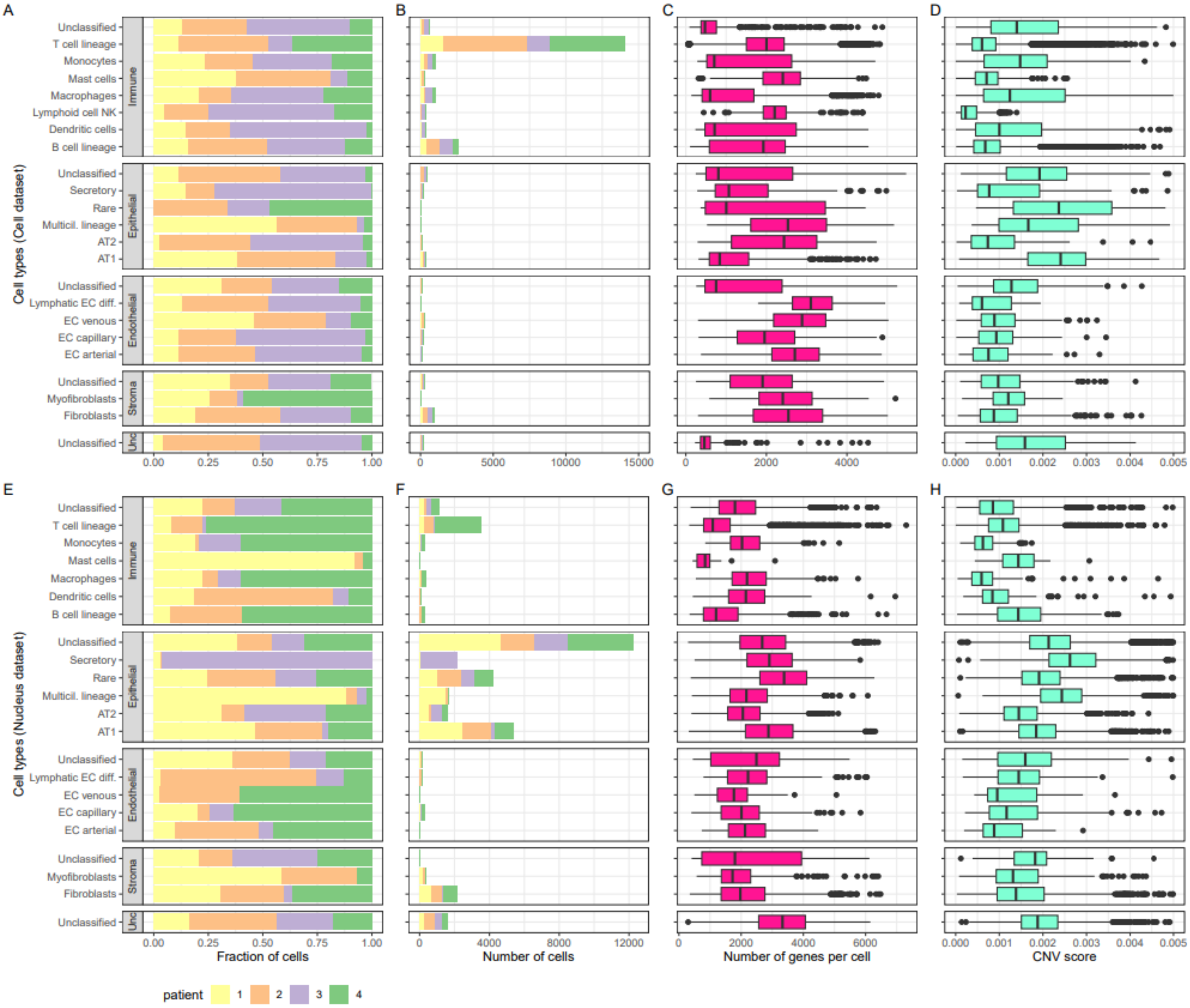

Figure 6 | Cell types characteristics (tumor tissue). For each of the four coarse (level 1) cell types annotations (*Immune*, *Epithelial*, *Endothelial*, *Stroma*) and unclassified (*unc*), further refined into finer categories (level 3 cell types): the fraction of cells (**A**: *cell* samples, **E**: *nuclei* samples) and the number of cells (**B**: *cell*, **F**: *nucleus*) originating from each patient. Box plots of the number of genes expressed (**C**: *cell*, **G**: *nucleus*) and the CNV score (**D**: *cell*, **H**: *nucleus*), with plot center, box and whiskers corresponding to median, IQR and 1.5□×□IQR, respectively. Note that only cell types with > 20 cells were retained for clarity in this visual representation.

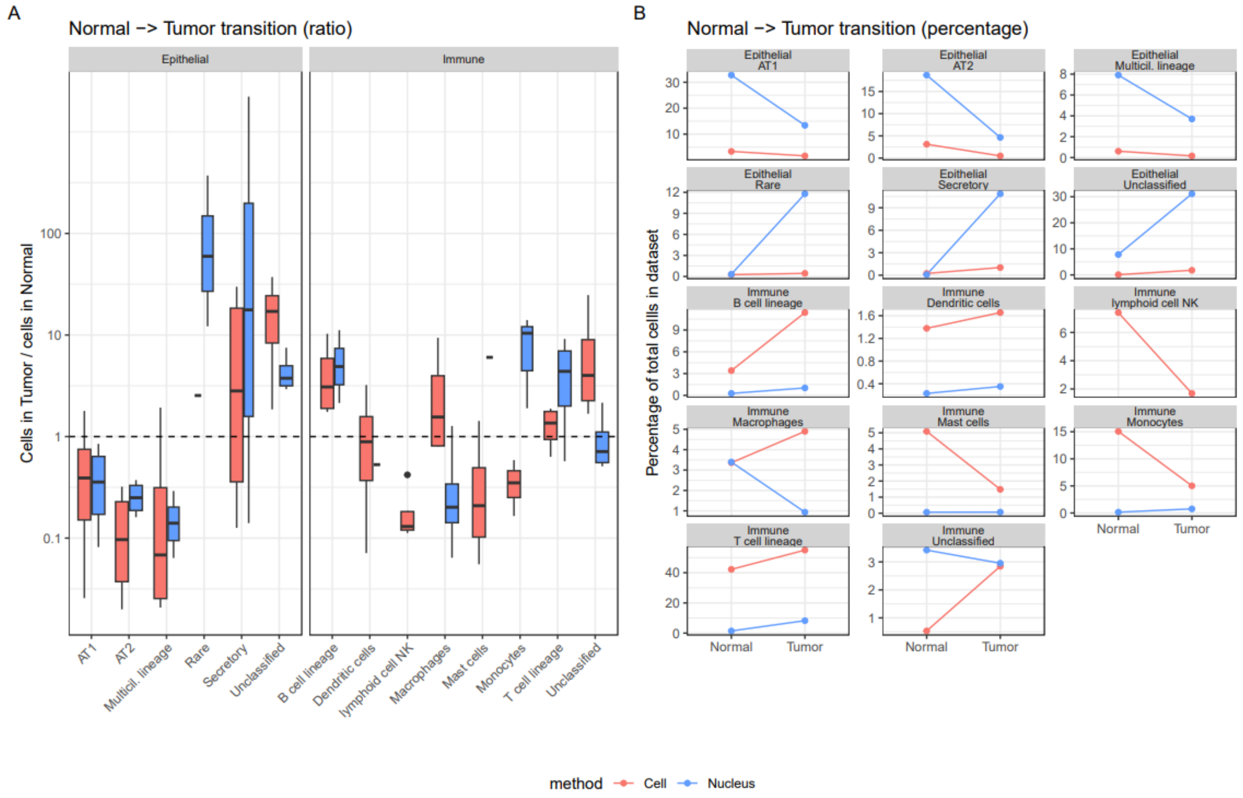

Figure 7 | Normal - tumor transition. **A.** For each specific (level 3) Epithelial or Immune cell type, the fraction of cells they represent in the *Tumor* dataset divided by the fraction of cells they represent in the *Normal* dataset (ratios above 1 represent an increase in the Tumor dataset), with plot center, box and whiskers corresponding to median, IQR and 1.5□×□IQR, respectively **B.** The percentage of specific (level 3) Epithelial or Immune cell types in *Tumor* and *Normal* dataset. Note that only cell types with > 20 cells were retained for clarity in this visual representation.

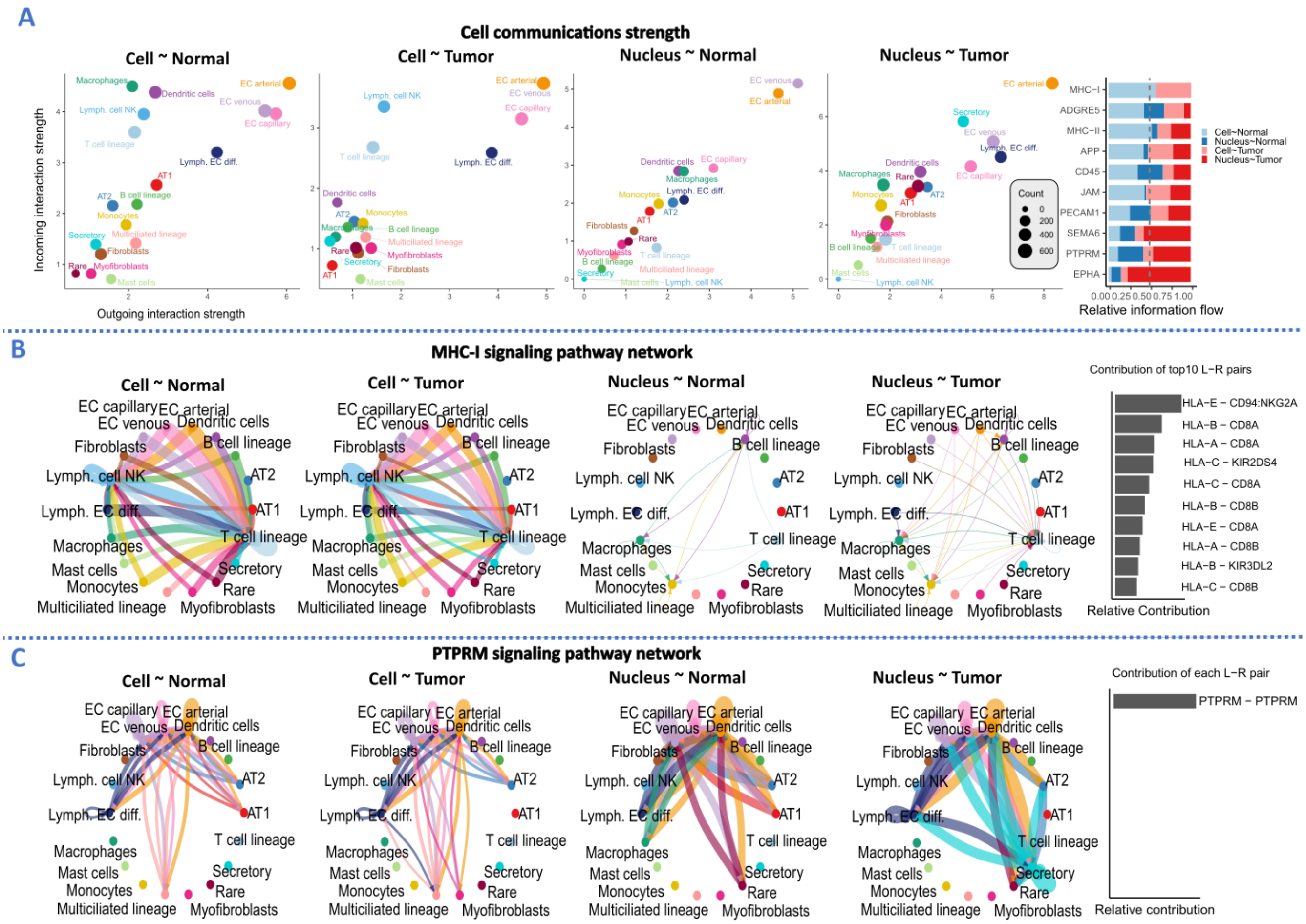

Figure 8 | The ligand-receptor interactome. **A.** Scatter plots of ingoing and outgoing interactions per tissue type and method for common cell types (see methods) among all comparisons. To the right are the top 10 interacting pathways**. B**: An example of pathway common in cell, rare in nucleus (MHC-I) with the contribution of the top10 ligand-receptor interacting genes (bar plot to the right). **C:** An example of pathway rare in cell, common in nucleus (PTPRM) with the ligand-receptor interacting gene (bar plot to the right).

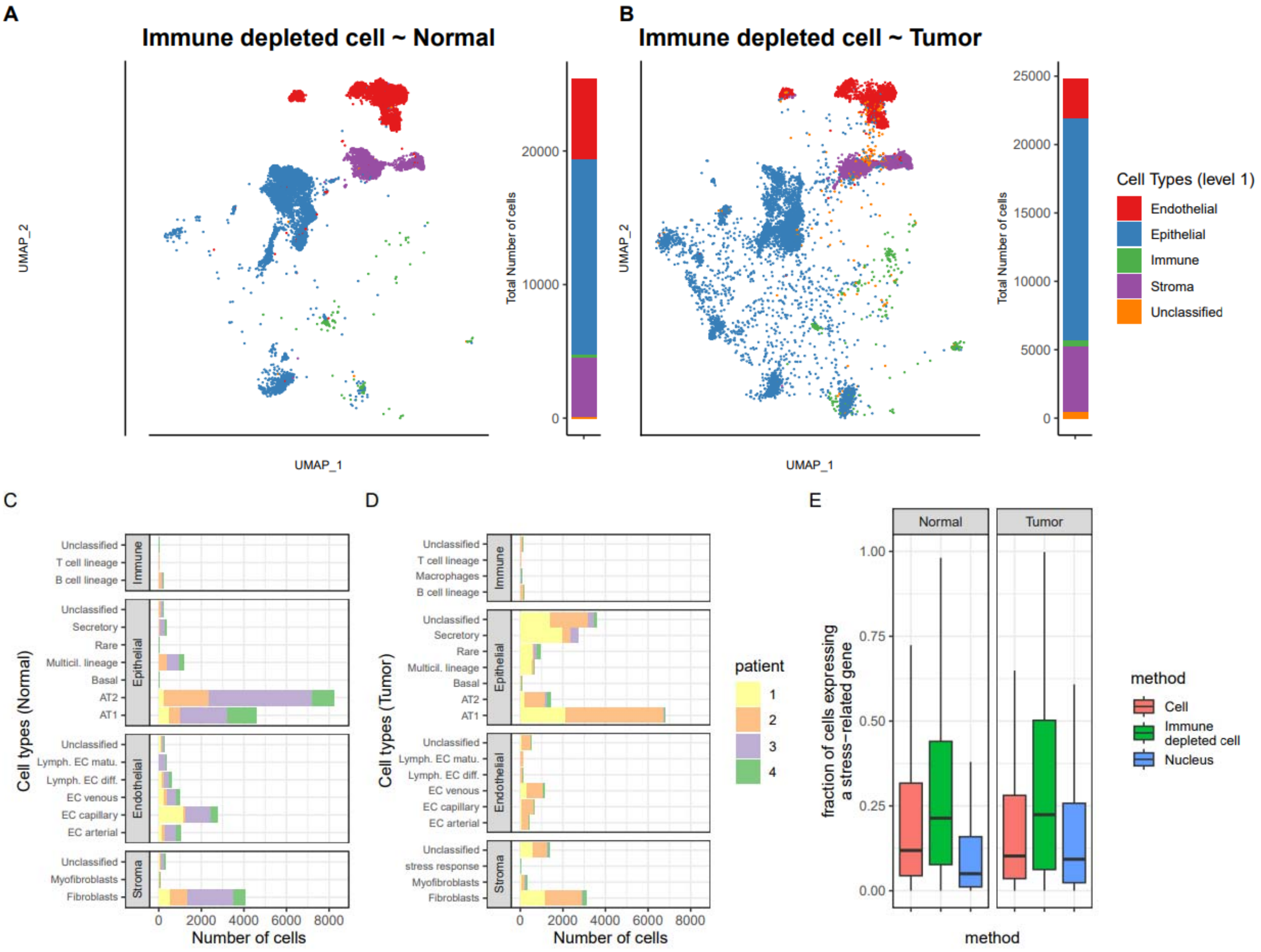

Figure 9 | UMAP representations and cell types annotations (immune depleted cells) for Normal (**A**) and Tumor (**B**) tissue samples with general cell types (level 1) annotation. To the right of each UMAP, stacked bar plots indicate the proportion of each cell type in the specific dataset. Number of cells in the Normal (**C**) and Tumor (**D**) tissue, per patient. **E**: The percentage of cells expressing a stress-related gene signature as a function of the experimental method and tissue type.