Differential Abundant Cell Population Analysis in COVID-19PBMC and Immune Checkpoint Blockade Single Cell RNASequencing Data

Differential Abundant Cell Population Analysis in COVID-19 PBMC and Immune Checkpoint Blockade Single Cell RNA Sequencing Data

Gege Qian

A Master's Thesis Submitted to the Immunology Program of

The Harvard Medical School to Fulfill the Requirement for the

Degree of Master of Medical Sciences in Immunology

Harvard University

Boston, Massachusetts

May, 2021

Dissertation Advisor: Dr. Shirley.X.Liu Author: Gege Qian **Differential Abundant Cell Population Analysis in COVID-19 PBMC and Immune Checkpoint Blockade Single Cell RNA Sequencing Data**

Abstract

Understanding immunological changes underlying tumors and disease microenvironments of other disorders, such as SARS-COV-2, has been challenging because it involves measuring genomic, epigenomic, and molecular changes in a myriad of cells. With the advent of single-cell technologies, it is now possible to assess transcriptome and chromatin accessibility at the single-cell resolution. The technology is currently being deployed in cancer and immunological disorders to study underlying immunological changes. These applications have also exposed the need for new statistical methods to handle increasing data complexity in single-cell experiments.

One such application is characterizing the transcriptomic profile to identify the differential cell population abundance between two biological conditions, which is probably the most fundamental application of the scRNA-Seq analysis. However, the current single-cell approach performs the analysis at the sample level resulting in insufficient statistical power to capture differential abundance due to the small sample size in scRNA data. Further, they ignore scRNA-Seq specific confounding factors such as inefficient genetic material extraction, amplified sample-specific bias, and differences introduced by various sequencing techniques. Here we developed an in silico approach (scDiffPop) that performs a robust statistical analysis at the individual-cell-level to determine biologically meaningful cell type abundance difference. Comparing to other methods, the commonly adopted DESeq is relatively robust to outliers and computationally efficient when dealing with large samples. However, its false discovery

rate(FDR) control, like the other methods, is sensitive to sample size[1]. scDiffPop pools related cell types based on the hierarchical relationship and performs sample-level DESeq on the larger meta-groups to gain a stronger statistical power. After validated by several positive and negative tests, we applied scDiffPop on COVID -19 and immune checkpoint blockade (ICB) peripheral blood mononuclear cells (PBMC) datasets to explore which cell populations are most responsible to the pathological phenotypes. In the COVID-19 scenario, we identified that the $\gamma\delta$T cell, IgG Plasma Blast, and CD14+ Monocytes are the more crucial immune population that majorly respond to the viral infection. While applying scDiffPop to Yuen et al.[2] dataset composed of 5 responders and 5 non-responders to anti-PD-1 treatment, we find that CXCR4- NK cells and RUNX3+ NK cells are enriched in the responders, whereas monocyte populations are more abundant in non-responders.

Contents

Figure Captions

Figure 1.1: SARS-Cov-2 viral infection and innate immunity response flow chart. The SARS-Cov-2 virus enters the cells through ACE2 and TMPRSS2 and recognized by intracellular and endocytic receptors which induce production of pro-inflammatory cytokines and anti-viral immunity.

Figure 1.2:Single-cell multi-omic analysis reveals from Stephenson et.al. shows the differential immune cell composition of various severity [3].A:Participants included in the dataset. B: UMAP vidualization of the 781123 cells after QC with annotation. C: Cell composition bar plot. Cells from b are separated based on condition and severity. Quasi-likelihood F-test was performed to compare healthy and COVID-19 patients. Differential abundance is determined using a 10% FDR and are marked with an asterisk.

Figure 1.3: Cancer escape immune surveillance through multiple pathways [4]. A: NK cells induce tumor apoptosis by release of cytotoxic granules, binding of TRAIL, TNF and FasL, and antibody-dependent cellular cytotoxicity(ADCC). B: Mechanism to evade NK cell killing. C: Tumor cells developed to defunction NK and cytotoxic T cells. D: Tumor released cytokines such as TGFβ and IFNγ, downregulate NKG2D and IFNγ in NK cells. They also promotes conversion of CD4 T cells into regulatory T cells (Treg). E: Tumor cell overexpress MMP and ADAMS to hide the activating ligands, which cripples the NK cell response and promotes the conversion of CD4 T cells to Tregs.

Tables

Cell Type	Marker Gene
Pan Marker	"MS4A1", "GNLY", "CD3E", "CD14", "FCER1A", "FCGR3A", "LYZ", "PPBP", "CD8A","CD4","CD45"
CD4 T cell	"CXCR3","TBX21","CRTH2","GATA3","CCR6","RORC","IL17","PDCD1", "ICOS", "FOXP3"
CD8 T cell	"CCR7","CD62L","KLRG1","CD45RO","CD122"
Macrophage	"CCR7","CD62L","KLRG1","CD45RO","CD122"
NK Cells	"CD56","CD244","TRAV24","TRDV1", "TRDV2", "TRGV9", "IL18R1", "CCR5"
DC	"CD56","CD244","TRAV24","TRDV1", "TRDV2", "TRGV9", "IL18R1", "CCR5"

Figure 1: Gene marker for dot plot of differential immune cells.

Acknowledgements

I would like to take the chance to express my deep gratitude to my research mentors Professor X. Shirley Liu, Dr. Avinash Sahu, and my co-worker Phillip Nicol from the Department of Data Science, Center for Functional Cancer Epigenetics at the Dana-Farber Cancer Institute, for their instructive guidance and collaborative effort in the pursuit of this research. I would also like to acknowledge Professor Shiv Pillai and Professor Michael Carroll from the Harvard Medical School (HMS) Master of Medical Science of immunology program (MMS immunology) for giving me the chance to study and research at Harvard affiliated institute and for their guidance and suggestions along my entire master period.

Chapter 1

Background

1.1 COVID-19 Overview

Since the end of 2019, the highly transmissible coronavirus disease 2019 (COVID-19) caused by the SARS-COV-2 virus has infected more than 120-million people and caused 2.6 million deaths from all around the world[3]. This disease can be spread by asymptomatic, pre-symptomatic, and symptomatic carriers through airborne respiratory droplets and direct/indirect contact. The SARS-COV-2 virus can induce acute respiratory immune reactions which causes mild symptoms such as cough, fever and shortness of breath. It can also induce severe lung damage and even death. Depends on race, gender and age, patients response drastically differently to the SARS-COV-2 infection. For example, the elderly population would be more likely to develop severe symptoms and has a higher hospitalized and ventilated proportion[10]. It is shown that 80% of the death of COVID-19 is from the population of age 65 or older in the US. On the other hand, infected children show much milder symptoms that are more constrained to the upper respiratory tract and rarely exacerbate to the step of hospitalization. Even for the children admitted to the hospital, a small portion ($< 7\%$) of them would require invasive treatment

and mechanical ventilation[11].

Benefit from its efficient entry mechanism, SARS-COV-2 virus could spread rapidly and develop severe symptoms in infected people. When SARS-COV-2 virus infects a cell based on the ACE2 binding and TMPRSS2 cleavage of surface-anchored spike protein, the fusion peptide of spike inserts itself into host cell membrane allowing fusion of the virus envelope with host cell plasma membrane and the release of the N protein coated plus strand viral RNA into the cytosol[12]. Upon infection, the virus would be recognized by pattern recognition receptors(PRRs) such as Toll-like receptors (TLR), RIG-I and MDA5[13], which induce production of anti-viral cytokines. For example, recognition by TLR3 induces activates NLRP3 inflammasome and leads to caspase-1-dependent cleavage and pro-inflammatory interleukin-1β and IL-18 production, which tiggers Gasdermin D-mediated death of infected cells (Figure 1.1).

Although many COVID-19 cases showing mild symptoms, some patients experience sever tissue damage and even death. Given the varied severity in response to SARS-COV-2 infection, understanding the disease mechanisms and how it associates with the demographic factors became critical in systematically studying COVID-19 and alleviating the pandemic. Multiple papers report that the dysregulated immune-response is responsible for the sever symptoms [14][15][16][17][18][19]. One of the feature of COVID-19 is the delayed type I IFN response as shown in Figure 1.1. Type I INF response is very important in control viral-infection. By comparing the 659 COVID-19 patient with severe pneumonia and 534 healthy donors, Zhang.et.al found that inborn errors involving in the TLR3 and IRF7 induced type I IFN immunity contributes to life-threatening symptoms[18]. It is also shown by Casanova's lab that, within a dataset of 987 COVID patients showing severe symptom, 10.2% of them have pre-existing auto-reactive antibodies(auto-Ab) against type I IFNs before COVID-19 infection[20].This auto-Ab are common in people with autoimmune polyendocrinopathy syn-

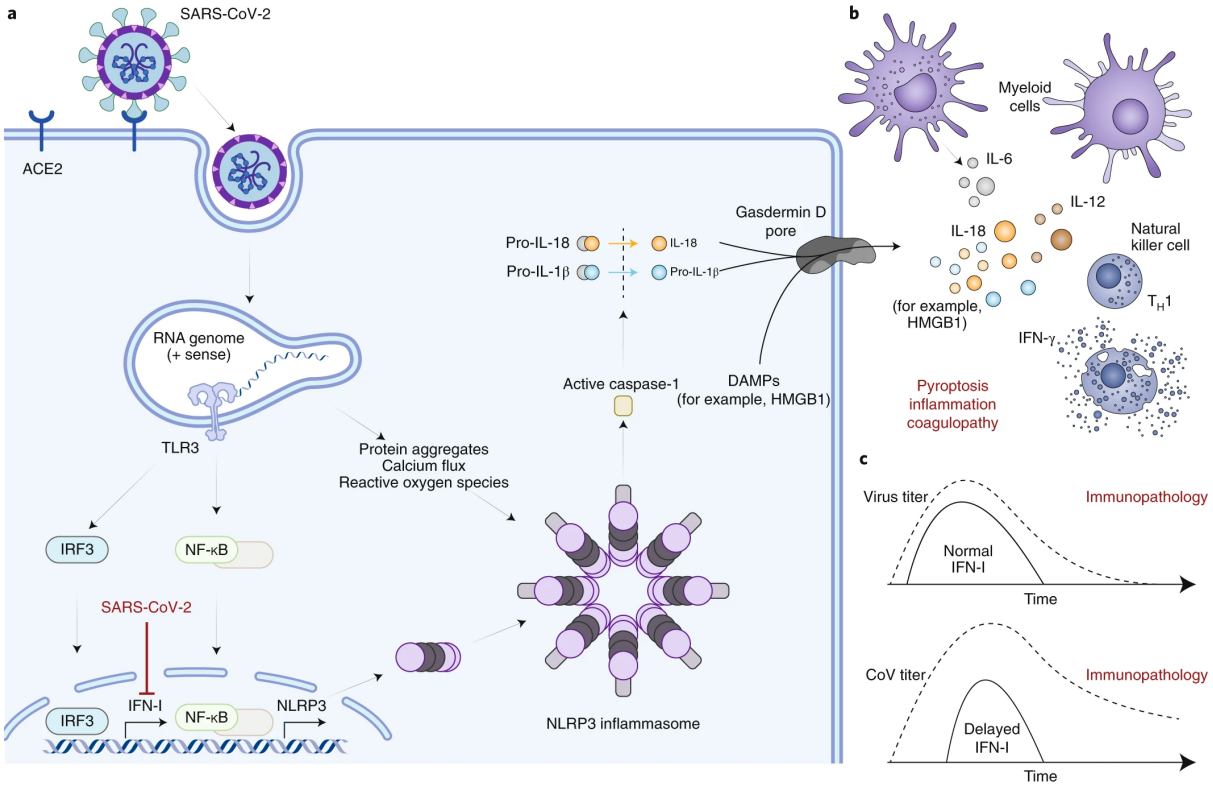

Figure 1.1: SARS-Cov-2 viral infection and innate immunity response flow chart.**A:**The SARS-Cov-2 virus enters the cells through ACE2 bindig and TMPRSS2 spike cleavage.Its recognition by PRRs activates NLRP3 inflammasomes. **B:** Anti-viral innate immunity production of pro-inflammatory cytokines. **C:** SARS-COV-2 infection shows delayed type I interferon response. [13]

drome type I (APS-1)[21] and lupus[22]. Moreover, its level is more pronounced in male than female [20](Figure 1.2). Thus, these studies not only suggest that type I IFN plays a crucial protective role against COVID-19, but also indicate its response could potentially explain how demographic difference and clinical history correlate with COVID-19 severity. The delayed and impaired type I IFN response spares time for viral replication and more tissue damage that triggers more exuberant immune response. As the immune system struggling to suppress the disease progression, escalated cytokines and chemokines are release, which attracts more pro-inflammatory cells to travel and to home at the lung that further exacerbate the

immunopathology leading to tissue damage and death. During the hyper-inflammatory situation, various immune cell types, including B cells, T cells, monocytes etc, contribute to the pro-inflammatory cytokines production[23]. Patients developing different severity have distinct immune cell composition in their peripheral blood[24]. Thus, in addition to the study of IFN response (which explains only about 10% of the severe case), more investigations had shifted their focus to decipher other pathways and to find which cell types are responsible for the unleashed immune response to COVID-19.

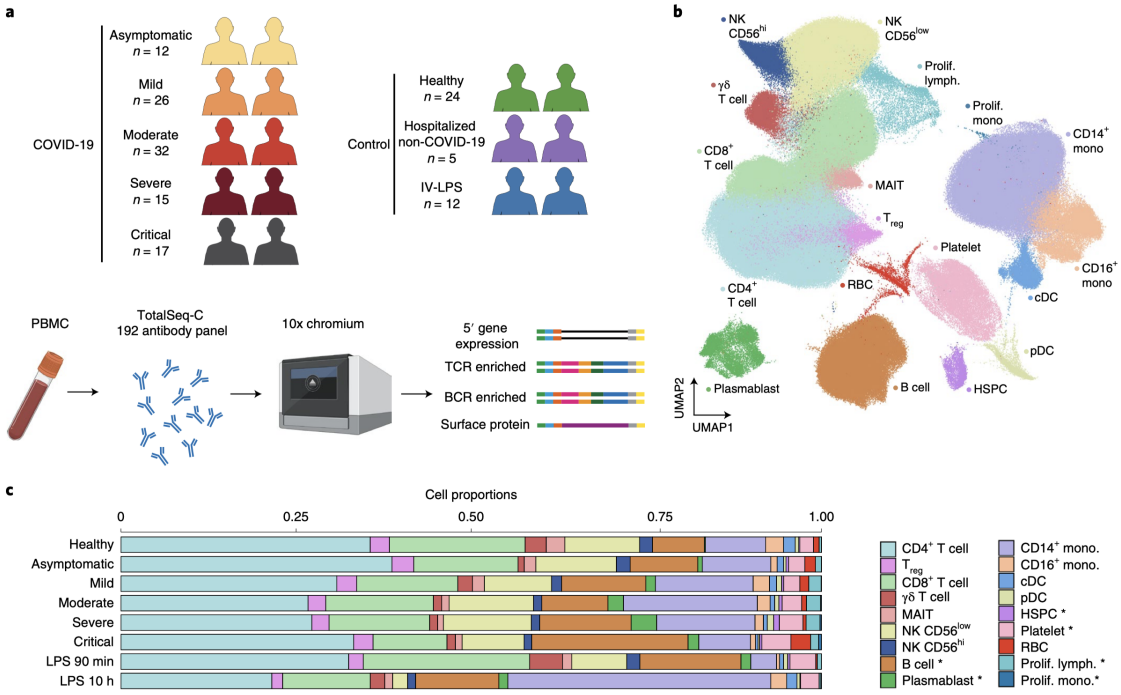

Figure 1.2: Single-cell multi-omic analysis reveals from Stephenson et.al. shows the differential immune cell composition of various severity [3].**A:**Participants included in the dataset. **B:** UMAP vidualization of the 781123 cells after QC with annotation. **C:** Cell composition bar plot. Cells from b are separated based on condition and severity. Quasi-likelihood F-test was performed to compare healthy and COVID-19 patients. Differential abundance is determined using a 10% FDR and are marked with an asterisk.

Immunological research on the single cell data of COVID-19 patients reports multiple im-

mune alterations that are predictive to disease severeness. For example, the impaired Type I and Type III IFN response, the innate immune cell dysregulation, exhaustion of infiltrated T and NK cells, and the excessively infiltrated neutrophils in the infected lung area are positively correlated with disease exacerbation[25]. As shown in Figure 1.2, Stephenson et.al conducted single cell experiments on 781,123 cells collected from healthy donors and COVID patients with different severity. Their analysis suggested that people experiencing wide range of clinical manifestation have different immune composition in their peripheral blood mononuclear cells (PBMCs) samples[3]. Both Figure 1.2 and other studies[26][27][28][29] report that persistent lymphopenia is common in COVID-19 patient. The lymphopenia could potentially affects T cell, B cell and NK cell lineages, but it shows greatest impact on T cell abundance. Some people hypothesised that the decreased lymphocytes at peripheral system reflects the immune cell recruitment to the infected respiratory track. From the bronchoalveolar fluid single-cell RNA sequencing data, however, no excessive lymphocytic infiltration is observed[30]. The researchers noticed that not only a lower counts of CD4+ and CD8+ T cell are observed in COVID-19 patients, but the infiltrated T cells also express an increased level of inhibitory receptors, which indicates functional exhaustion[31]. T cells show distinct reactivity among individuals, where some imply a reduced cytokine production [32], and others suggest an overaggressive T cell response[33]. Observed in many other immune cell-types, the distinct response can be attribute to both inter-individual difference and various disease severity. Thus, to systematically study the immunopathological mechanism, we would need a statistically robust method to to efficiently integrate large datasets and to extract biological meaningful information from the confounding patient-specificity.

1.2 Cancer Immunology

1.2.1 Tumor Development and Immune Surveillance Evasion

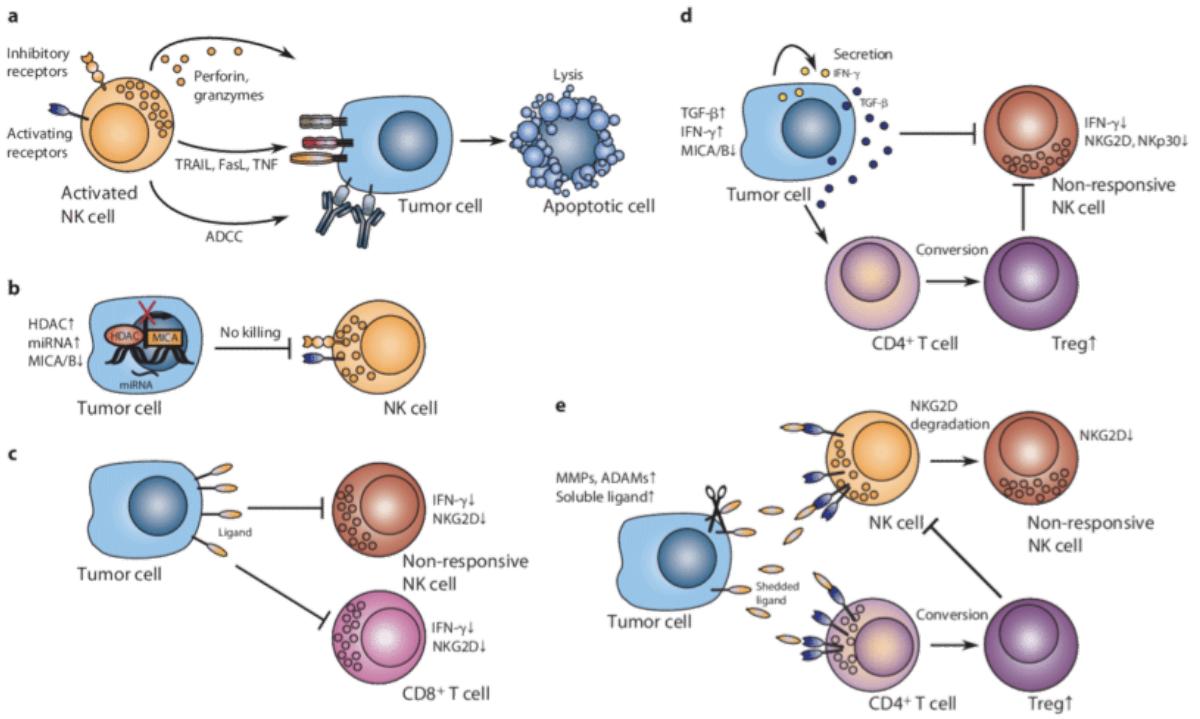

Figure 1.3: **Cancer escape immune surveillance through multiple pathways [4]**. **A:** NK cells induce tumor apoptosis by release of cytotoxic granules, binding of TRAIL, TNF and FasL, and antibody-dependent cellular cytotoxicity(ADCC). **B:** Mechanism to evade NK cell killing. **C:** Tumor cells developed to defunction NK and cytotoxic T cells. **D:** Tumor released cytokines such as TGFβ and IFNγ, downregulate NKG2D and IFNγ in NK cells. They also promotes conversion of CD4 T cells into regulatory T cells (Treg). **E:** Tumor cell overexpress MMP and ADAMS to hide the activating ligands, which cripples the NK cell response and promotes the conversion of CD4 T cells to Tregs.

Immune system dysfunction, including both overreacting and incapability of controlling and identifying invasion, could lead to a deadly consequence. A good example of immune surveillance failure would be the development of cancer. Since 2018, cancer has become the

second leading cause of death in the United States[34]. In the past decades, a great amount of effort has been invested in the field of cancer immunology to understand how tumor cells escape from immune surveillance. During the immune cell development, it goes through the process of identifying foreign from self so that it would not kill functional self-cells. Although tumor cells origins from a normal somatic cell, because of the genetic instability, they accumulate a large number of mutations and, therefore, express proteins absent in normal cells (neoantigens) that can be recognized by our immune systems. However, as more genes mutated, tumor cells evolve and acquire abilities to proliferate out of control, to induce angiogenesis, to metastasis through lymphatic and blood vessels, and to escape from immune surveillance. Many studies indicate the dysfunction of the immune system, such as T cell anergy, overly activated Treg, and defects in antigen presentation contribute to the immune surveillance evasion of cancer.

There are broadly three strategies that tumor cells adopted to escape immunity: avoid recognition, release/express inhibitory molecules, and activate suppressive immune populations. MHC class I expressed on the surface of every cell except for red blood cells is majorly responsible for presenting processed cellular protein to T cells for activation. Some tumor cells are able to downregulate or abolish MHC-related gene transcription. The loss of MHC not only limits the immune recognition of cancer cells but also impairs the anti-tumor T cell activation. However, the absence of MHC could be recognized by NK cells and trigger cytotoxic granule release. Oftentimes, instead of downregulating MHC, tumor cells would present selective loss of MHC haplotypes to reduce antigen-presentation. Some tumors evolved to upregulate the non-classical HLA-G that would compromise both NK and T cell response. In addition, to constrain self-presentation, tumor cells could send paracrine signaling to cripple the ability of antigen antigen presenting cells (APCs). Expressing both MHC class I and MHC class II receptors, these cells are specialized in presenting antigens to activate both CD8 and CD4 T

cells. Thus, the tumor cells could achieve less exposure to the immune system by violating the priming and maturation process of the APCs. For example, tumor cells could induce expression of immunosuppressive cytokines such as IL-10 and TGFβ to inhibit dendritic cell maturation and T cell cytotoxic function[35]. apoptosis[36].

T cells are the key mediator of anti-tumor immunity that can specifically target tumor cells expressing neoantigens. However, their function are compromised by tumor microenvironment which could suppress their reactivity and make them dysfunctional. As mentioned before, tumor could interfere DC maturation and inefficient antigen presentation, which lead to T cell anergy[37]. In addition, tumor released inhibitory molecules could also directly act on effector cells to induce T cell anergy and neutralize the Fas-ligands [36]. T cell anergy is also known as T-cell-induced tolerance that is caused by a constant lack of co-stimulatory molecules. Anergic T cell could identify the antigen from tumor cells but could not respond to induce activation and produce cytokines even when both the antigen and co-stimulatory molecules are presented. Except for anergy, tumor could also induce T cell exhaustion. Once T cell receptor (TCR) binds to high-affinity antigens, it triggers downstream cell-intrinsic pathways that activates naive T cells and guides differentiation into cytotoxic effector T cells. The effector T cells undergo expansion and acquire the ability to produce cytokines and to release granules containing granzyme and perforin that could lyse targeting cells. Following the proliferation peak, 90-95% of the effector cells die from apoptosis [38], and the survived T cells become memory T cells providing long-term protection [39]. To transform from effector to memory phase, an environment without antigen stimulation and persist inflammation is needed. However, chronic infection and cancer violate the requirement. Since tumor antigens are derived from self-proteins, they are less immunogenic. The tumor-specific T cells have TCR-antigen binding affinity, because the ones with high avidity are negative selected during development.

Moreover, antigen presentation is commonly crippled in tumor, which lead to insufficient T cell priming. Regulated by the immuno-suppressing tumor micro-environment, these T cell,that are chronically exposed to antigens and high-level of cytokines, become exhausted. Exhausted T cells express particular high level of inhibitory receptors such as PD-1, LAG-3, TIM-3, CTLA-4, BTLA, and TIGIT[40][41][42][43][44][45]. They lost the ability to produce cytokines (IL-2, IFNγ, TNFα), to generate granzyme B, and to induce cell-mediate apoptosis [46]. Although the exhausted T cells have undermined ability in controlling tumor progression, they are still functional and can be reinvigorated by blocking the inhibitory pathways.

Tumor immune tolerance can be induced by T cell depletion caused by regulatory T cells (Tregs)[47]. Characterized by the expression of FOXP3, Tregs are a specialized sub-population of CD4+ T cells, that can suppress immune response to maintain homeostasis and self-tolerance. There are two types of Tregs the tymus developed natural Tregs (nTreg) and the peripheral raised induced Tregs (iTreg). In addition to their protective role from auto-reactivity, Tregs could promote tumor progression by contributing to the immune suppressing tumor microenvironment. Various types of tumor acquire the ability to accumulate Treg by selectively recruiting Tregs, promoting Treg proliferation and converting infiltrated conventional CD4+ T cells in to Tregs. Tumor microenvironment is often hypoxia which induce expression of chemokine ligand CCL28 and recruit CCR10+ Tregs[48]. Except for CCL28, Tregs could also be recruited through other signaling pathways such as the CCL8/CCL5 axis [49]and the CCR5-dependent manner [50]. Moreover, tumor developed the ability to transform the infiltrated lymphocytes to tumor-specific Treg. Conversion of conventional CD4+ T cells to Tregs are primary reported in blood cancer induced by malignant cells and associated altered immune cells [51][52][53]. A more recent study claims that some tumor-associated Tregs in ovarian and colorectal cancer-bearing mouse are converted from IL-17A+FOXP- cells [54]. Tumor

microenvironment promotes Tregs activation and expansion, which further emphasizes its enrichment. For example, Treg cells adopt a combination of glycolytic and oxidative metabolism, which allows it to proliferate in the resources-scarce tumor environment [55]. Tregs function to establish the immunosuppressive environment through multiple mechanisms. For example, they release large amount to inhibitory cytokines such as IL-10, TGFβ, IL-35 and VEGF, which could inhibit effector T cell activity and DC differentiation [56][57][58]. It could also directly kill tumor-specific T cells and antigen-presenting DCs through Treg-induced apoptosis.

1.2.2 Cancer Immunotherapy

Immunotherapy has become an indispensable pillar in cancer treatment. The current approaches for cancer immunotherapy emphasized on advancing cytotoxic T cell response and converting the suppressive environment to an immune-hot niche. Upon activation, the checkpoint molecules are over-expressed in effector T cells to control for hyper-activation. Tumor cell hijacks this machinery to disable the anti-tumor response. Thus, the first antibody-based attempt, the immune-checkpoint blockade (ICB), functions by using monoclonal antibodies to block checkpoint molecules and ligands[59]. However, only 20% of patients would respond to immunotherapy across different cancer types[60].

To circumvent the limitation, immunotherapies aiming at other cell populations should be taken into consideration. Recent studies indicate that the presence of Tertiary lymphoid structure (TLS) found in tumor tissue is associated with a favorable prognosis and more effective local antitumor immune response in many cancer types[61]. To improve efficacy and to better predict and monitor patients' responses, efforts in studying mechanisms of how the check-point blockade functions to reinvigorate anti-tumor response and which populations are mainly affected become saliently important[62]. Resembling the structure of a germinal center, TLS is

densely populated by abundant B cells, T cells, and DCs. One study on patients with metastatic melanoma shows that the ICB reactivated T cell could provide long-term protection and prolonged survival. This study suggests that the durable response is associated with CD8+ T and CD20 + B cell co-occurrence in the TLS, and B-cell-rich tumors also have a higher level of naïve and memory T cell infiltration[63]. Not only does the TLS facilitate CCR7+ T cell and CXCR5+ B cell infiltrating to tumor site through HEV, but it also protects the immune cells from the immunosuppressive tumor microenvironment. This configuration also provides structural support for efficient DC antigen presentation and concentrated T-B crosstalk and mutual activation[64]. Thus, the presence of TSL at the tumor site suggests a more responsive and potent anti-tumor immunity that could stay reactive for long-term metastasis and correlate with an optimistic prognosis and better response to immunotherapy.

Dendritic cells are essential for T cell activation and, thus, are expected to play a defining role in response to ICB. Peng et al. show that the expression of PD-L1 on DC compromises T cell response[59]. Mediated by type II interferon, the PD-L1 expression upregulation in DCs upon antigen uptake is meant to protect them from cytotoxic T cell killing yet dampen antitumor T cell activation. To identify if the anti-PD-L1 inhibitor functions through enhancing T cell priming at secondary lymphoid structures or through reactivating the exhausted T cell population at the tumor site, the author applied FTY720 treatment to prevent lymphocyte egression so that on T cell is found in the peripheral system. No significant difference was observed when PD-L1 blockade therapy was conducted on FTY720 applied situation, whereas a significant elevation of IFNγ+ CD8+ T cell frequency was found at the tumor site. They also showed that DCs are dispensable for tumor-infiltrating T cell reactivation. Taken together, the PD-L1 blockade therapy works by blocking the PD-L1 on DC to grant potency for priming and invigorating exhausted cytotoxic T cells.

Rare T cell populations such as MAIT T and $\gamma\delta$ T did not get enough attention until a recent study revealing the tumor-specific cytotoxicity of Vγ9Vδ2 T. For therapeutic use, Vγ9Vδ2 T cell is expanded and activated ex vivo and adoptively transferred back. Although the method shows low toxicity, it also could only convey moderate result[65]. With knowing that $\gamma\delta$ T upregulate PD-1 expression 2 to 4 days after recognition, Hoeres et al. show that although PD-1 blockade did not promote cell lysis of the activated and expanded $\gamma\delta$ T, it induces elevated expression of IFN-γ, which is an essential pro-inflammatory and anti-tumor cytokines.

The importance of B cells and DCs in immunotherapeutic response sheds light on predicting patient response based on multiple immune cells- types other than cytotoxic T cell. Immune cells present in the tumor microenvironment and contribute to the immunosuppressive milieu might also be essential for dictating response. Moreover, the rare T cell populations such as MAIT T cell, NKT cells, and $\gamma\delta$ T cell would be particularly attractive targets since they share similarities with majority $\alpha\beta$ T cells but also shows innate immune properties. To understand which immune components regulate immunotherapeutic response, recent studies obtained genomic sequencing data from combinatory-immunotherapy trials to analyze the differences of the immune cell populations between responders and non-responders.

Tumor immune surveillance evasion is really complicate involving various cellular and molecular pathways. Although origins from different antigen class and pathological pathways, both COVID-19 and cancer can be concluded as immune dysfunction. To effectively protect us, the innate and adaptive immunity need to cooperate in a well-regulated manner. The delicate balance is so important to many pathological processes that intensive research has been focused on delineating its comprehensive interaction. Given the complex system, ICB clinical trail datasets are typically small and confounded by many factors such as treatments, patients status and tumor locations. Because of these confounding factors, development of robust an-

alytical methods are essential for identifying the real biologically meaningful from the noisy background.

1.3 Sequencing techniques

Since RNA-seq was developed, it has reshaped our understanding of almost all aspects of biology research and became an indispensable tool to obtain genomic information. The wild acceptance of RNA-seq promotes the differential gene expression (DGE) analysis between different biological conditions for a molecular-level understanding of the drug-resistance mechanism, the pathology pathways, and the difference between responder and non-responder. The RNA-Seq library preparation can be concluded into three steps: mRNA-extractoin, reverse transcription,and amplification [66]. The first step is RNA extraction and purification by poly(A) capturing to distinguish messenger RNA (mRNA) from the majority of transfer RNA (tRNA) and ribosomal RNA (rRNA). The obtained mRNAs are then processed by chemical or enzyme for truncation so that the RNA molecule segments would be optimal for sequencing. The RNA fragments are reverse transcribed to cDNAs that would b further modified by ligating adaptor sequence to both ends. Lastly, the cDNA flanked by the adaptor sequence is amplified through polymerase chain reaction with the adaptor sequence as a primer. Driven by the technological improvement, the sequencing technique could generate higher throughput at a lower expense and less starting RNA sample. Moreover, the increasing demand for accuracy in deeper sequencing stimulates the development of long-read RNA-Seq. The enormous amount of genomic information requires higher computational power and more efficient models for analysis.

Although the optimal composition of tools depends on specific biological hypotheses and computational resource availability, the analysis often follows the same pipeline comprising

four steps: the raw sequence alignment, the quantification of read counts, counts matrix normalization and scaling, and the differentially expressed gene analysis[67]. As more statistical and computational tools are developed, the RNA-Seq can be accommodated to multiple fields for application. For example, the gene expression profile can be used for cancer classification by detecting aberrant transcription. Comparing with the microarray technique, RNA-Seq could provide genomic information with higher resolution. The RNA-seq could not only detect the genomic mutation within the exon region but also quantify the expression level and shed light on genomic alteration on a whole-genome sequence level. From the RNA-seq data, one could easily identify the differentially expressed genes and gene isoforms that could guide experiments on molecular mechanism exploration. RNA-Seq enables mutation and germline variation detection that facilitates the allele-specific variants expression analysis for pathology mechanism study. The sequencing technique could also be applied on samples beyond mRNA, including long non-coding RNA (lncRNA), microRNA (miRNA), PIWI-interacting RNA (piRNA), etc. Conveying an important regulatory role in multiple disease-related gene transcription, the miRNA, for example, could be better characterize from the sequencing technique for clinical use.

RNA-seq also supports pathogen analysis. Given that RNA functions as important genetic material for the virus, the RNA-seq data could directly detect the viral identity and its related species from the genomic sequence. Although bacterium has DNA as their genetic material, they could release exogenous RNA. Studies indicate that RNA derived from gut microbiota is associate with gastrointestinal-track immune tolerance by affecting protein-protein interaction and regulating gene transcription[68]. RNA-seq technique powers the establishment of a comprehensive exogenous RNA reference database, which could guide future research on host-pathogen tolerance, food-sensitivity, and clinical infection[69]. Overall, RNA-seq based

technologies revolutionized molecular biology and enabled efficient explorations on previously inaccessible fields for scientific research, clinical diagnosis and prognosis application, and drug development purpose.

1.4 Single Cell RNA-Sequencing(scRNA-Seq)

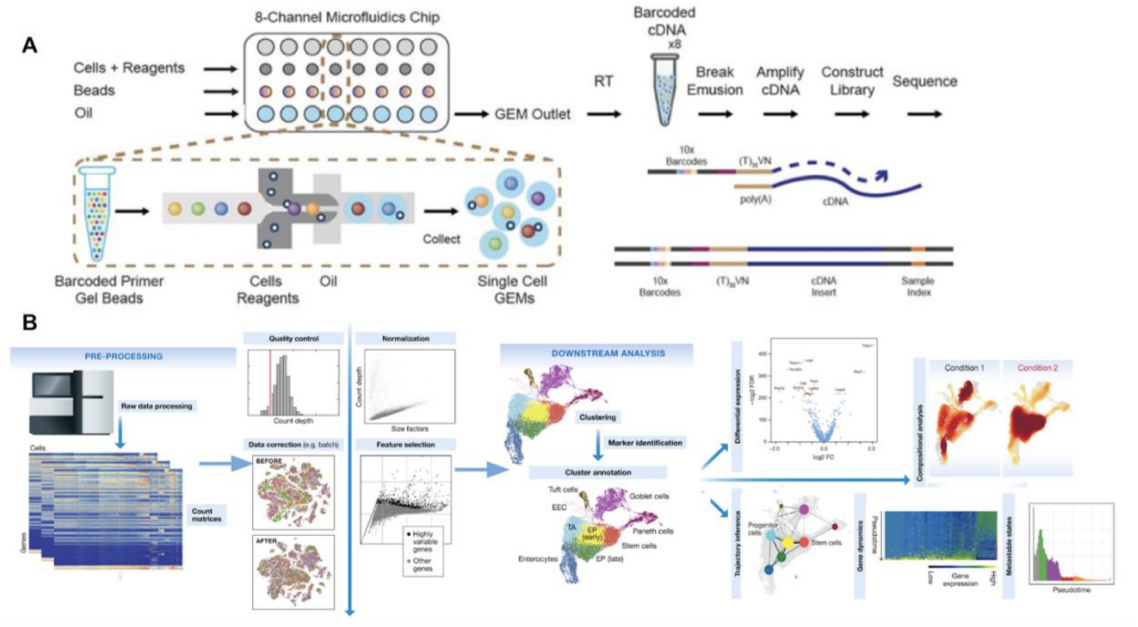

Figure 1.4: **scRNA-Seq Experiment and Analysis**: **A**: scRNA-Seq sample preparation flowchart [5]. **B**: Commonly adopted scRNA-Seq data analysis pipeline [6]

1.4.1 Power of Single Cell RNA-Seq Analysis

The development of the next-generation sequencing (NGS) fosters biologists to delineate the genomic, transcriptomic, and epigenomic landscape of their research interest, and the single-cell technique advances this understanding to a much higher resolution. Since the single-cell sequencing technique for DNA and RNA was recognized as the Method of the Year of 2013[70], it had been widely adopted by biological investigation in diverse fields. Providing the cell-level

resolution expression profile of transcriptome-wide genes, this technique enables cell-cell distance calculation that facilitates cell-type clustering and between clusters relationship establishment. More importantly, the detailed genomic data allows mapping of differential expression between conditions to specific cell populations, so that one can distinguish the cell type contributing most to a specific phenotype, study the interaction between populations, and identify the expression change along with cell differentiation and evolutionary states. Encouraged by the indisputable power of single-cell sequencing, a huge amount of effort had been invested in optimizing this technology, minimizing its cost, and standardizing the analysis pipeline, so that this tool would be more accessible and the precious data could be shared and tested by the scientific community. Along with the rapid development of this technology, sequencing hundreds of thousands of cells has been a routine and necessary practice for publication. To integrate these accumulating rich resources, many institutes have established purpose-oriented databases comprised of enormous datasets that have been filtered for quality and organized based on research focus, cell types, and organism. The grant quantity of data and its rising necessity for research emphasize the requirement for more efficient computational and statistical tools for exploring real biological relevant findings from technological and experimental noise that came from the amplification process of the limited amount of genomic material available from each cell.

1.4.2 scRNA-Seq Experimental Challenges and Data Analysis

Stimulated by the widespread of single-cell RNA Sequencing (scRNA-Seq), various experimental protocols and computational pipelines had been developed to serve different purposes such as cell clustering, differential gene expression analysis, chronological and evolutionary trajectory delineation, etc. However, despite how powerful scRNA-seq, there are some inher-

ent cofounding factors associate with sample preparation, sequencing, and alignment process. Because of the limited amount of mRNA present in a single cell, series of modification and amplification is needed for library preparation to recover most of the gene expression profile. Although paired-end reverse transcription with unique molecular identifiers (UMIs) to ensure cell specificity ameliorates strand-specificity and PCR bias, only 10 -20% of the transcripts can be captured by the poly(dT) primer after cell lysis[71]. With knowing all the technical limitations, one way to minimize the noise is sequencing on a deeper level for a more accurate estimation of the transcriptional state. Once the scRNA-seq data is obtained, meticulous quality control for both the reads and sequence-alignment should be performed. After the data is pre-processed and aligned, a common practice is to generate an expression matrix with rows of genes and columns of cells. This matrix is required for sharing on public databases and used as input for future analysis. In spite of the previous careful practice, there are still many biases that could deviate true biological difference, for example, the batch effect, sequence-depth, and library size, which I am going to elaborate later in this chapter. Thus, normalization is necessary. Once obtained the normalized expression matrix, we can finally implement experiments to solve biological questions and feel comfortable trusting the results. A starting point of investigating the expression profile is to cluster the cells and based on the well-studied marker genes and additional antibody staining from other experiments such as flow cytometry to annotate cell types of interest. Taking the annotated clusters, one can also calculate the distance between cell populations to build a hierarchical tree which could help predict the evolutionary relationship and provide insight on cell differentiation. Differentially expressed gene analysis between two biological conditions could be performed on the annotated expression matrix. After acquired the differentially expressed gene list, we could map the genes back to the cell-type level based on abundance so that we could find the cell populations that contribute to either one of

the biological phenotypes. Combining the cell-type level significance with known-knowledge of the between cell interaction and cell-type-specific function offers intuitive guidance for experimental design on a cell level. However, it is extremely challenging to extract biologically meaningful information from the scRNA-Seq data, especially on the cell-subtype level because of the chaotic sample-specific factors, the limited number of biological replicates, and the small cluster for some cell types.

Many published single-cell RNA-Seq experiments lack replicates even in the high impact journals. Unlike the bulk RNA-Seq that more replicates compensate for shallower sequencing and promote the statistical power and accuracy for differential expression analysis35, scRNA-Seq experiments are not divisible due to the extremely laborsome and expensive sample preparation step. Not to mention the batch variation is confounded with variation between samples since every batch is just one sample, and there is no real replicate in scRNA experiments. However, given the unavoidable technical variations introduced along the sample preparation process, it is hard to extract the biological difference. To handle this problem, many integrating methods are developed to combine datasets with a similar setting to expand the number of cells in each condition by adjusting the gene count matrix from different experiments to mitigate the batch effect. This practice is very common in recent publications, especially for those using human samples. Because of the scarcity and ethical reasons, scRNA-Seq data of the human sample contains much more confounding factors than data generated from the mouse experiment. In addition to batch effect and technical variations, patient factors such as age, sex, life habit as well as disease progression in human scRNA data required much more careful normalization and adjustment. One of the integrating methods proposed by Stuart et al. applies multivariate methods to identify conservative gene patterns across datasets to anchor cells in a lower dimension to adjust the expression accordingly[72]. However, this method focuses

on integration between relatively large datasets, which is not efficient in providing correction for sample-specific bias within the dataset. Thus, instead of modifying the expression matrix before differential expression analysis, for a small dataset, one could also manipulate the significance calculation while conducting the differential expression analysis to mitigate the sample-specific difference.

Accurate cell-type annotation is crucial for downstream differential expression analysis. Several publicly available scRNA-Seq data annotation methods are developed based on reference gene expression profiles. For example, SingleR[73] utilizes the bulk transcriptomes pure sample to improve scRNA-seq clusters annotation. Unlike SingleR, scMatch[73] does not require a pre-defined cluster for annotation. It annotates the scRNA data by matching each sample to the closest cell type in a large reference dataset. Another annotation tool named Garnet[74] can rapidly annotate cell types based on the hierarchical markup language of cell-type-specific genes. Garnet takes advantage of the well-studied gene expression profiles for different cell-types to generate the markup language and maps to scRNA-Seq data to identify the cells that express definitive gene markers for each of the cell types. Then Garnett uses these anchors to train a classifier to annotate the other cells based on similarity. Although these methods are computationally efficient and have been proven relatively accurate, they require a well-annotated dataset which is not available for some biological conditions such as tumor microenvironment.

1.4.3 Published Cell-type Abundance Analytical Tools

Visual Examination on Reduced Dimension

With a comprehensive understanding of the challenges involved in identifying the abundant differential cell-type, we can proceed to the recently published methods regarding this topic.

Without any complicated statistical test, the most intuitive way of identifying cell abundance difference is to look at the plot of the cell on a reduced dimension. Uniform Manifold Approximation and Project (UMAP)[75] and t-Distributed Stochastic Neighbor Embedding (t-SNE)[76] are the two most popular dimension reduction embedding for a two-dimensions visualization. The cell population differential abundance between two biological conditions could be found by visual examination of the low dimension plot. This method is inherited from the inspection of the old-school cell-marker gated flow-cytometry plot, where only limited markers are stained with antibodies for the cell to be identified. Although it is obvious that visual inspection could not put the full power of RNA-Seq data into use, this method is still adopted in recent publications. For example, in the paper from Chua et al.[77] and the paper from Liao et al.[30] published in 2020, they claim that some cell-types show a higher proportion in one condition based on the UMAP observation and the violin plots of the associated signature genes. Oftentimes, scientists would also perform cell-fraction statistical analysis to verify the conclusion from visualization by performing a simple student t-test, Wilcoxon-Mann-Whitney test, and Chi-squared test. However, the analysis on cell-level is affected by the dimension reduction process and visualization limitation, which fails to capture the gene features stored in multidimensional space. Thus, here I am going to introduce you to two methods that perform analysis on the original high-dimension matrix to predict the majorly differential abundance cell type between biological phenotypes.

Augur: Trained Random Forest Classifier

Augur (Skinnider et al.)[78] is a machine learning method that trains a random forest classifier to classify the cells into either one of the two conditions. For each of the cell-types, the classifier is trained the gene profile and associated true label from a subset of cells. The classifier

is later used to generate predictions on a test dataset. The test statistic of cross-validated Augur is the area under the receiver operating characteristic curve (cvAUC). AUC score equals the accumulated probability of the fraction of the true positive rate (TPR) over the false positive rate (FPR) over a range of classification thresholds. It is commonly used to access the performance of a binary classifier. Combining with cross-validation, the cvAUC would also allow evaluation of the classifier's generalizability to a new dataset. The authors believe that the more responsive the cell-type to a perturbation of biological condition, the more separable it should be comparing to the unaffected ones. Since classifier performance heavily depends on the training sample size, the predicted AUC is positively correlated with the number of cells in the sample. However, because of the previously mentioned technical difficulty and the fact of uneven presence of different cell types, this dependency accounts for inaccurate identification of the differentially abundant cell-type. Augur solves this problem by picking a smaller dataset from the sample for cross-validation and report average cvAUC for further analysis. In this case, when they applied Augur to their simulated data, it can correctly predict the differential abundant cell types.

They further compare Augur with the methods using differential gene expression significance cutoff to identify the cell-type contributing to a specific biological condition. They mention that the differential gene counts highly depend on the sample size of the cell-type, which makes cell-type with larger sample size but the smaller transcriptomic difference to be recognized while the rare but significantly differential population being overlooked. Augur could avoid the gene counts different due to sampling size and generate predictions that could best capture the real transcriptional perturbations between conditions. To achieve a more computationally efficient training, Augur implemented feature filtering steps during which two steps are performed: 1) the genes with small inter-cell-type variation are eliminated, and 2) during

each training iteration, randomly selected genes are dropped to improve memory and computational efficiency. The second idea is similar to cross-validation in that only a portion of the data is used for training to avoid overfitting. These feature selection steps allow Augur to adapt to perform analysis on larger datasets with thousands of hundreds and even millions of cells. Because of its inherited machine learning property and the careful controlling on overfitting, Augur could generate robust results regardless of the sequencing depth.

Despite all the advantages it gains from machine learning, Augur also suffers from the classifier training process. Although it would not be affected by the sample size difference between cell-types, it requires that each sub-population contains a decent number of cells for efficient training. Even though the scRNA-seq normally contains a sample of thousands of cells, after clustering and annotation, the cells belong to each of the cell-type clusters would be limited. The small sample size could lead to two major problems that could impair the statistical power and compromise prediction accuracy. First of all, because there are too few training cases, the model would not only generate inaccurate predictions but also become really unstable. A small training set made it hard to capture the general feature of the data. Instead, the model would be overfitting to the training data points, which leads to low generalizability. On the other hand, a small dataset also indicates fewer test cases. The test result would be highly perturbated by unrepresentative outliers, and the test variance is going to be huge. Moreover, the test statistics (such as sensitivity, specificity, positive/ negative predictive value, etc.) are important in demonstrating the model performance. Almost all of these scores are calculated from fractions with the number of the test sample at the denominator. The test sample size is essential in controlling the randomness of observed model performance. Although one might observe an acceptable performance from a model trained on a small dataset, the learning curve is actually concealed by the random testing uncertainty[79]. Thus, regardless of how compu-

tational efficient and how well it controls for the other factors, Augur fails to perform analysis on subpopulations with small sample size.

DA-Seq: Separability of Cells based on the Projected Coordinates in Cell Enrichment Space

DA-Seq (Zhao et al.) is another approach to detect the differential abundant cell population[80]. Unlike Augur, which requires predefined cell annotation and clusters, DA-Seq is not constrained by the predefined sub-groups but performs analysis on a single-cell level so that the identified differential abundant subpopulation could contain cells distributing across various pre-defined clusters. DA-Seq computes a differential abundance enrichment score for each of the cells on a multiscale level. These scores are calculated based on the relative enrichment of the k-nearest neighbors obtained from the transcriptional space between two biological conditions. The prevalence score of each cell is obtained from a range of k so that a sequence of statistics (multiscale measure) is obtained based on different neighborhood sizes. These multiscale measures are considered as a projection from gene-expression space to a newly defined relative-distance space with a lower dimension. Cells from the same differential abundance subgroup contributing to one of the biological conditions are expected to cluster closely in this distance space.

After acquired the projected coordinates, a logistical regression classifier is applied to cells in the new space. Similar to Augur, when training the logistic regression classifier, cross-validation is implemented to ensure generalizability. Moreover, ridge regularization is applied for more stable model performance. The predicted probability of each cell from logistic regression is collected and used for unsupervised clustering. The cluster obtains from the prediction probabilities describes the relationship of cells in terms of differential abundance between

phenotypes. After obtained the list of Differential Abundance sub-population, differential expression analysis is performed for cell-type identification. For a more computational efficient DEA, selecting a subset of the most representative genes that captures a majority of the variation within the sub-sample is often performed to shrink the matrix dimension. The author applied the feature selection method based on stochastic gates to obtain the minimum characteristic genes that differentiate the DA subpopulation. However, a method like DA-Seq that is based on the separability in projection space is sensitive to technical noise and sequencing depth. Additionally, intra-condition heterogeneity due to patient-specific confounding factors are not normalized in the DA-Seq, and, due to cost efficiency, the current scRNA-Seq experiments often lack biological replicates that exacerbate this effect. Thus, due to the cohort and sample difference, although the cells are separable in gene expression and modified DA space, they might not represent the true biological meaningful difference between conditions.

Chapter 2

Data, Methods, and Result

2.1 Introduction

In this paper, we introduce scDiffPop, a robust statistical model for identifying the differential abundant cell types between two biological conditions. We will first describe the technical details of scDiffPop algorithm. We will compare its prediction accuracy with Augur and validate its efficacy on several scRNA-Seq datasets. We apply Augur and scDiffPop to the dataset of matched blood and tumor single-cell data and check if the obtained differential abundant cell types match the known knowledge. We then demonstrate the scalability of scDiffPop by applying the method on a large mouse embryo dataset with 2 million cells from different developmental stages for a negative test. After acquired confidence in accuracy and efficiency, we will apply scDiffPop on ICB and COVID-19 PBMC datasets to identify the cell types that are mostly affected by the biological conditions.

scDiffPop is designed to solve two problems of scRNA-seq data: 1) the method-inherited small sample size and 2) the amplified patient specific bias, which are not well handled in the existed methods. Many published papers are using the cell number fraction to determine dif-

ferential abundance significance. However, the cell fraction method collapse a huge amount of single cell information into just the fraction on patient-level leading to weak statistical power. Thus, we want to leverage the cell-level information into the model to gain stronger power for determining the differentially abundant cell types. To overcome the limited subtype population, scDiffPop builds a hierarchical tree that summarizes the gene-expression similarity and relationships between cell types. By combining several cell types into metapopulation, we could perform differential gene expression analysis on the metapopulation and gain higher statistical power. On the other hand, to solve the sample-specific difference, scDiffPop performs a robust permutation test on a patient-level psuedobulk to determine if the biomarkers of the given cell subpopulation are significantly overexpressed in either of the two phenotypes.

2.2 Results

2.2.1 Overview of scDiffPop Algorithm

We believe that if a cell population is differentially abundant between two conditions, then the meta-population containing the cell population will be molecularly different, which can be identified by DE analysis. Starting with outlining the important steps of the scDiffPop algorithm, I will explain what is expected from each step and how these statistical methods solve scRNA-Seq analysis challenges. A more detailed description can be found in the material and method section. Figure 2.1 A a graphical flowchart, could provide some intuitive understanding of our algorithm. First of all, scDiffPop requires scRNA-Seq data from two distinct biological conditions with clear annotation. For example, in Figure 2.1, based on the tissue type, the datasets are divided into cells from blood and from matched tumor samples with proper cell-type annotated. When the cell type annotation is too broad or unavailable, we assign the cells

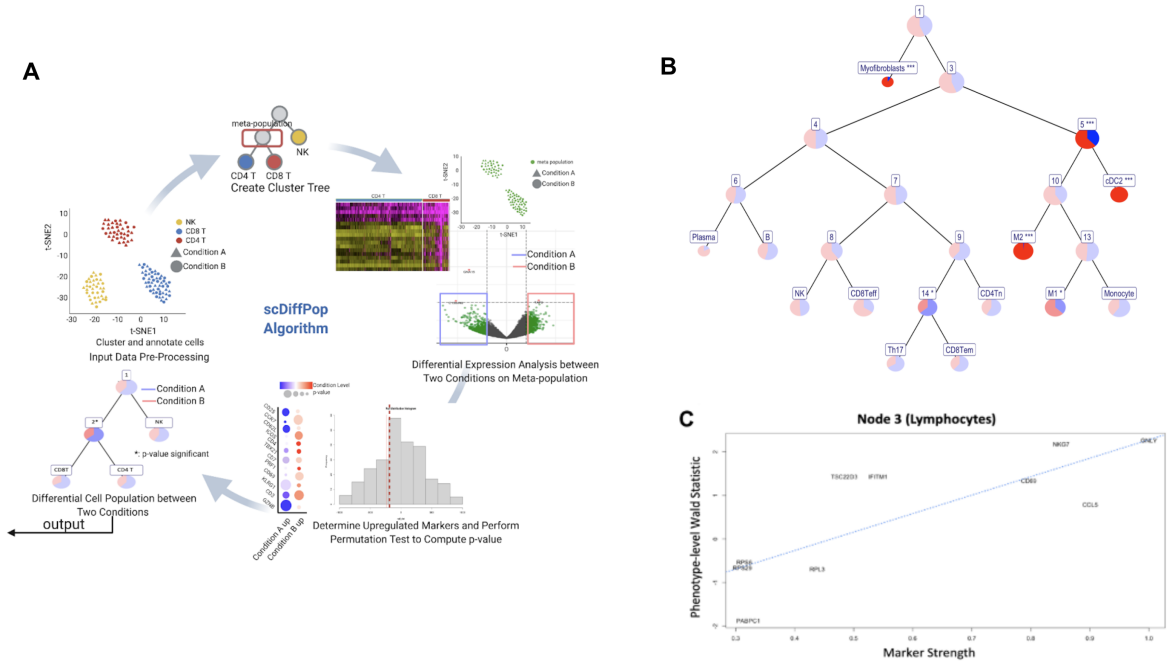

Figure 2.1: **A:** scDiffPop algorithm starting with the scRNA-Seq of two conditions with annotation. **B:** applying the scDiffPop on the matched blood/tumor sample of four patients. B is the pie chart where red denoted the cell portion from the tumor and blue denoted cells from the blood sample. The star (*) is the significance, one star is one order magnitude (e.g. ** is less than FDR<=0.01). **C:** Gene marker plot with marker strength on the x-axis and Wald statistic on the y-axis.

into clusters and combine the reference-based label transfer method with the well-known gene marker expression analysis to provide the most accurate annotation. This method integrated the idea behind the previously mentioned SingleR and scMatch with manual adjustment (refer to method session) for an accurate label prediction that provides a hierarchical relationship that makes biological sense. Once the cells are properly labeled, scDiffPop is going to generate a cluster tree that delineates the hierarchical relationship of all subpopulations based on the Euclidian distance in gene expression space. The relationship tree is built from a binary divisive clustering algorithm, where all cell subpopulations are initially collected to the same group and

recursively split into smaller sections based on expression profiles so that the constructed tree has all annotation on the leaves. We also provide the flexibility for users dealing with unusual samples or have specific modifications on expression profile to input customized tree (e.g., Supplement Figure 3.1).However, we would recommend using the generated tree if annotation is not reliable or if the one cell-type is a mixture of multiple subpopulations. Our method is constructed based on the euclidean distance between every cell types, therefore, could better capture the real between cell-type relationships. Shown in Figure 2.1B is the cluster tree for the blood/tumor dataset comprised of leukocytes that could accurately represent the distance among different immune subtypes. The cells are firstly branched into lymphocytes and myeloid cells and keep dividing into finer classes until all cell types are assigned to the leaves. After constructed the cluster tree, scDiffPop proceeds to the most crucial differential expression analysis step. scDiffPop performs DEA on each node of the tree based on the expression of the marker gene. We expected to observe more significantly differentially expressed biomarkers of the cell type that shows different abundance between the two conditions. We used the widely accepted DESeq2 [81] on a sample-level pseudo-bulk to avoid inter-cell variation-related confounding factors. DESeq2 is going to generate a list of genes with associated Wald, and BH adjusted statistics explaining in which phenotype and how significant the gene differentially expressed. scDiffPop then takes the fold change weighted Wald statistics to determine the most definitive gene markers of each sub-population. By checking the marker gene expression in different conditions, scDiffPop determines the abundance difference. Figure 2.1C is a plot with the y-axis being the phenotypes and the Wald statistics on the x-axis. From this plot of Wald statistic of 10 marker genes of the lymphocyte population, we observed that most of the markers are overexpressed in the responders. Moreover, since the Wald weighted average does not convey parametric meaning, we performed a robust permutation test to assess the statistical signifi-

cance. A permutation test is a statistical significance test that randomly shuffles the labels to calculate all possible values of the test statistic. By default, scDiffPop will perform 250 permutations to calculate the p-value, which is adjusted for multiple testing through false discovery rate (FDR) [82]. As a data exploratory tool, scDiffPop provides freedom on picking visualization methods that best serve the users' research interests. Users could choose the tree depth and coloring scheme for pie tree and gradience tree to show the significance of differential abundance.

2.2.2 scDiffPop Outperforms Augur in the Matched Blood-Tumor scRNA datasets

As previously introduced, Augur determines differential abundance by looking at the separability of cell types between conditions. To compare with Augur, we applied both methods on a scRNA-seq dataset from Yuen et al.[2] containing 25459 leukocytes from four matched blood and tumor samples. scDiffPop identified five cell types that are significantly enriched (indicate by the *) in tumor samples: the M1 macrophage (FDR <0.1), the M2 macrophage (FDR <0.001), the conventional Dendritic Cell (cDC) (FDR <0.001), and the myofibroblast (FDR <0.001) (Figure 2.2B). It also shows that the Th17 and CD8 effector memory T (CD8 Tem) metapopulation is enriched in the blood samples with a significance of FDR less than 0.1. Our finding could be confirmed by other studies and are biologically accurate. For example, macrophages are tissue-resident cells that are normally rare in blood. In particular, M2 macrophages inducing immune suppression for wound healing could promote tumor progression[83]. Expressing a higher level of CCL3, tumor cells could recruit pre-cDCs from the bloodstream which are activated and proliferate at the tumor site to form tumor-associated cDCs with a lower level of CD11 and MHC class II expression [84]. Myofibroblasts are proven

to promote epithelial tumor evolution by mediating the epithelial-mesenchymal crosstalk [85]. Thus, it is not surprising to see these five subpopulations overexpressed in the tumor samples. Although Th17 cells can be found in some tumor samples and promote tumor growth, its enrichment depends on the tumor type, malignancy, and therapeutic intervention[86]. Thus, although showing significance on its metapopulation level, scDiffPop successfully managed the difference between tumor types and generated a conservative conclusion on the Th17 abundance with no significance on the specific subpopulation.

We then compared our result with Augur's prediction. Since Augur assumes the cell types that could be clearly classed into either condition are the differentially abundant populations, it reports the populations with AUC closer to 1 as thoes differ the most between two conditions. When applying Augur to the same dataset, only three differential abundant cell types are

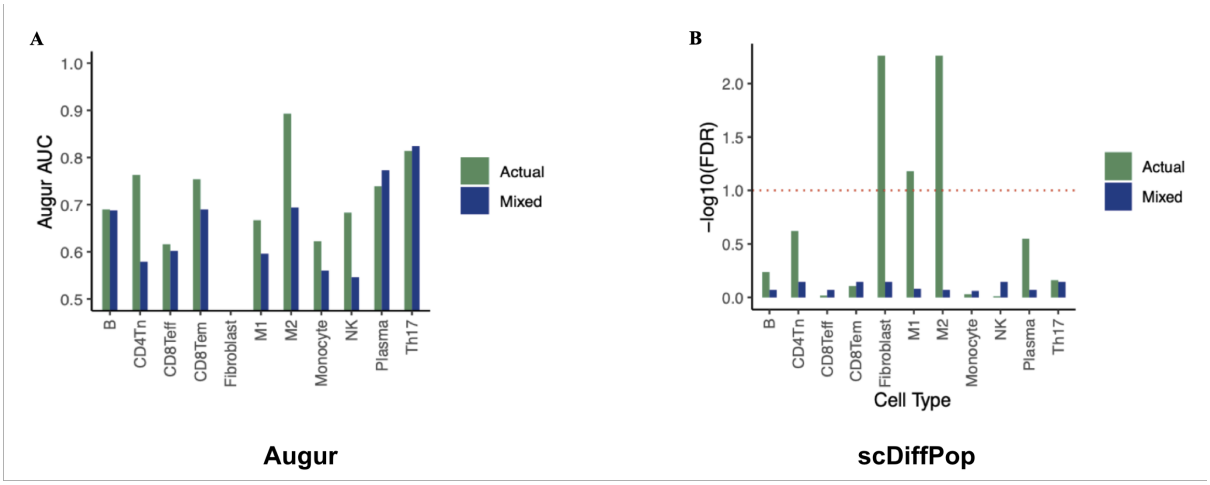

Figure 2.2: **Augur and scDiffPop comparison on the blood/tumor matched sample.** To compare the AUC with FDR, we take the negative log 10 on FDR (e.g -log10FDR =1, FDR= 0.1). The green bar is the significance obtained from the original dataset, and the blue bar is the statistics obtained on a randomly labeled dataset as a negative test. **A** is Augur performance and the closer to 1 the more significant. **B** is scDiffPop significance: ones above the red dotted line (FDR<0.1) are considered differential abundant cell-types.

identified: the M2 macrophage (AUC = 0.89), Th 17 (AUC = 0.81), and CD8+ T cells (AUC =0.75) (Figure 2.2A). Since Augur trains the random forest classifier using cross-validation, it fails to obtain accurate features of the subpopulation with very low cell counts and could not generate a reliable prediction. This explains why cDC and myofibroblasts were not shown in its prediction. However, Augur found significance in cell types other than those reported by scDiffPop, such as the T cell populations. We then implemented DA-Seq for confirmation. DA-Seq found a similar population as scDiffPop and could not verify the significance of the T cell population proposed by Augur. Thus, we concluded that scDiffPop, a very robust and conservative tool, can successfully identify the differential abundant cell types between conditions and outperforms Augur on the small cluster analysis.

As mentioned in the introduction, patient-specific random effect could contribute to false separability in gene expression space that does not have biological meaning. Thus, to test if this bias is truly involved in Augur and whether scDiffPop could mitigate it, we performed the test on the previous ICB data and swapped the label of two randomly picked samples from each condition (mixed experiment). If the model captures the real difference between conditions, we expected to see no differentially abundant cell-types when some samples are mislabeled for there are only four matching samples, and after switching labels, the distance between blood and tumor samples is diminished. However, we observed in Figure 2.2A that although the general AUC value decreases, among the top four ranked cell types from Augur prediction, two of them (CD4 T naïve and CD 8 T effector) reported AUC scores from mixed data similar to the results generated from original data. In addition, the AUC for Th17 was even higher when trained with a mislabeled dataset. This indicated that the classifier performance did not necessarily reflect the biological difference between conditions, and its prediction is influenced by patient-specific effects. On the other hand, the FDR values of the cell-types based on the

mislabeled data were all far below the cutoff of 0.1 in the scDiffPop result (Figure 2.2B). Contrary to the Augur results, the top-ranked cell-types on the scDiffPop list are only significant in the properly labeled dataset, but this significance vanishes when the difference is mitigated by the two mislabeled samples. Thus, we can conclude that highly depending on the training dataset, Augur prediction is swung by sample quality and could not properly handle the sample specificity. However, by performing a simple permutation test, scDiffPop controls for the confounding sample-specific effects and returns robust predictions that represent the real differential abundant cell types between two biological conditions.

2.2.3 scDiffPop Performs Efficiently and Accurately on Larger Dataset

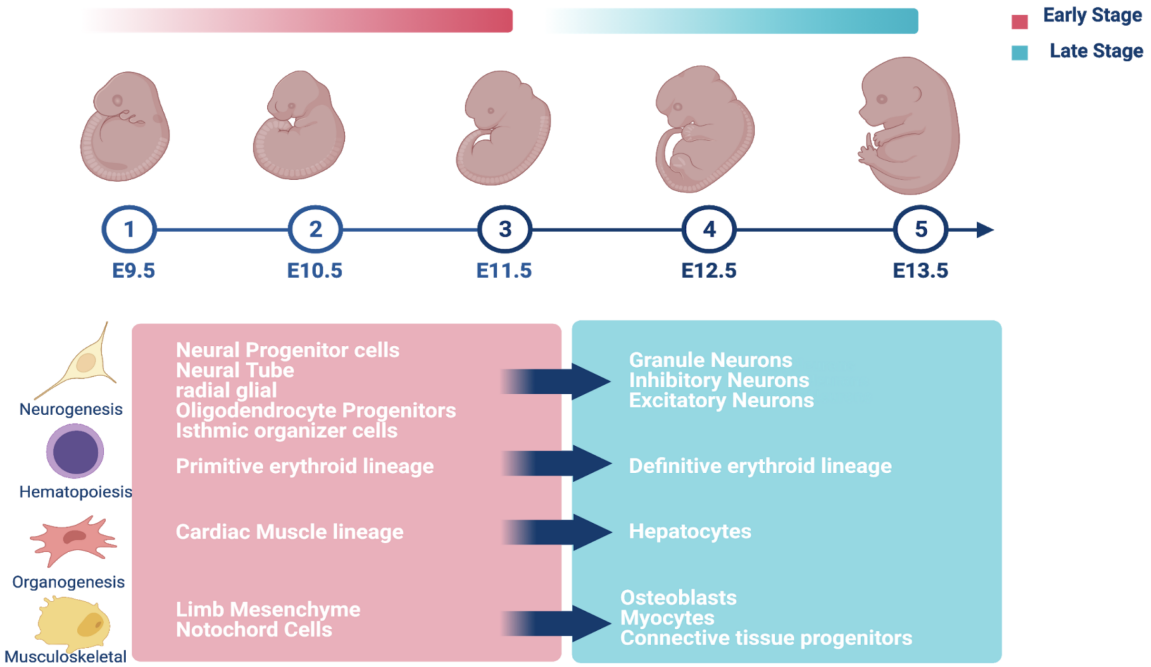

Figure 2.3: **Apply scDiffPop on the MOCA[7] dataset**. From the enriched cell types of both early and late stage, we can see the process of CNS development, hematopoesis and organogenesis.

To test the scalability of scDiffPop, we applied it to the Mouse Organogenesis Cell Atlas

(MOCA)[7]. MOCA proposed by Cao et al. contains scRNA-seq data of 2 million cells derived from 61 mouse embryos at various stages of development between 9.5 and 13.5 days of gestation (E9.5, E10.5, E11.5, E12.5, E13.5). This dataset has 37 distinct cell types identified and annotated. We manually divided the dataset into early-stage (including E9.5-E11.5) and late-stage (E12.5 and E13.5). This division is based on the pseudotime trajectory of the gene expression profiles on the pseudobulk of mouse embryos from different stages. We performed scDiffPop based on the previous setting on the MOCA dataset. Because of the large size and fine annotation, the extensive gradient plot (Figure 2.6) is attached to the end of this section. Figure 2.3 summarizes the finding from the hierarchical tree plot. We can see that scDiffPop reports multiple differentially abundant cell-types between the early and late developmental stages. Among the 37 annotated populations, 25 of them are significantly different with FDR <0.01. The primitive euthyroid lineage, neural tube, and early mesenchyme are the most enriched cell types in the early stage of development, whereas granule neuron, inhibitory neuron, excitatory neuron, connective tissue progenitor, and osteoblasts are most abundant in late-stage pseudobulk. Comparing the list of cell types, we could observe many cell types enriched in the early development stage are actually the precursor or preliminary state of the differential subpopulations abundant in the late stage. For example, the highly enriched primitive erythroid cells in the early stage are the precursors of definitive erythroid cells that mature in the extravascular space and supersede the primitive cells along development[87]. Moreover, a clear neuron development is also clearly described from the cell-type abundance that the neural progenitor cell and neural tube are found in the early-stage whereas the more specialized granule neuron and the excitatory neurons are enriched in the late stage. To take a closer look, we found that the metapopulation node 20 branches out to hematopoiesis and CNS organogenesis-related cells. These two processes are known to happen at the early stage of development[88] [89]. More-

over, the analysis of scDiffPop on the huge dataset is fairly computational efficient that takes around half an hour (may vary with the computational power of the device). Thus, it is fair to say that scDiffPop could efficiently generate a hierarchical gradient tree that captures the relationship among cell-types and propose the truly differentially abundance subpopulations.

Moreover, we also used the same dataset for the negative control test, where we took the development stage with most cells (E10.5) and randomly assign the cells to the two groups with similar sizes. Since these two conditions are arbitrarily assigned with no biological difference, we expected to see no significance. Although there is no meaningful biological relevant difference, the variations introduced by the sample-specific factors and batch effect could still lead to false-positive significance. However, scDiffPop returns no enriched population with all FDR above 0.78, which proves the scDiffPop's capability of removing effects from these confounding factors. Then, we introduced intra-condition variations by randomly label cells from the entire MOCA dataset without any information from the predefined developmental stage. Because of the unevenness of each stage, we would still expect to see some difference. Indeed, scDiffPop reports 9 differential cell-types with FDR less than 0.1. However, the effect sizes of these 9 populations are on average 1.2, which is much smaller than that on the actual dataset with an effect size of 10.39. From both the true negative test and the randomly assigned test, we were more confident that our finding within the original dataset has a large effect size and significance is biologically relevant.

2.2.4 scDiffPop Identified the NK cells, Naïve CD4 T cells and Effector Memory T cells to be Enriched in ICB Responders

After validating scDiffPop on both datasets, we proceed to apply this method to find interesting biological differences on the COVID-19 and ICB datasets. First, we adopt the PBMC data

from the immune checkpoint blockade trial (ICB) containing 26609 cells from 10 patients on $\alpha - PD1$ immunotherapy[2]. This dataset is comprised of comparable number of cells from responder and non-responder, and half of the patients are responder and half are non-responders.

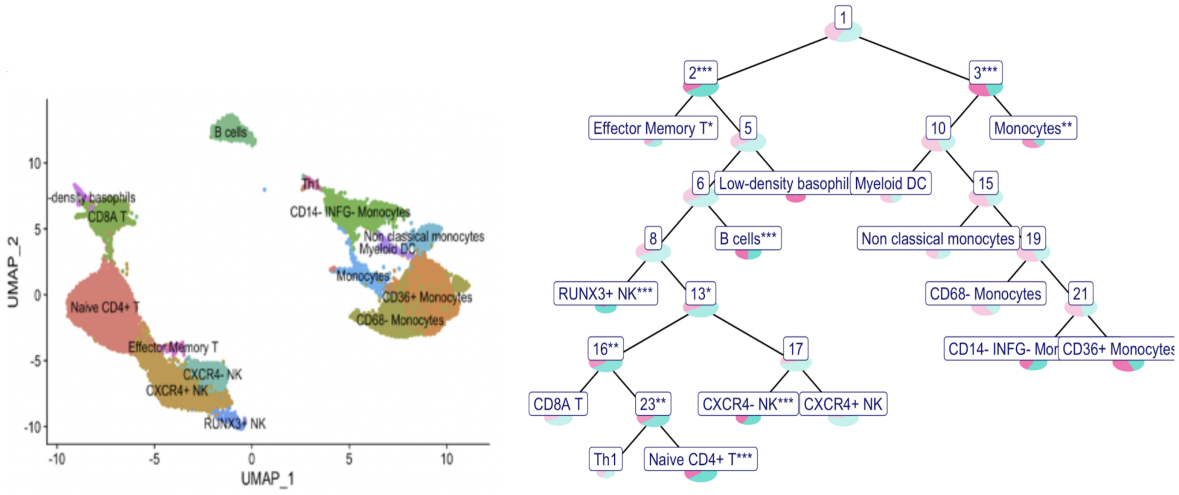

Figure 2.4: **Applying scDiffPop to immune checkpoint blockade data from Yuen et al.[2]**. Consisting of 26609 cells from 10 patients treated with α-PD-1 immunotherapy, the dataset has half responder and half non-responders. The dataset is properly annotated. **A:** UMAP plots of the responder and non-responder with 7 cell-types annotated. **B:** The gradience tree generated from scDiffPop summarized the relationship between cell types and how enriched each subpopulation is between two conditions. The red represents enrichment in non-responder, and the green represents enrichment in responder, and the color gradience shows the significance.

The UMAP plot in Figure 2.4A allows us to check if the annotation is validated. We can see that the monocytes, B cells, and other lymphocytes are all clustered together with the proper between cluster difference. Since scDiffPop depends heavily on the annotation accuracy, we could trust our prediction only after we validated the annotation. scDiffPop identified 6 significantly different populations, and among them, RUNX3+ NK cells, CXCR4- NK cells, naïve CD4 T cells, Effector Memory T cells, and B cells are significantly enriched in responders,

whereas monocytes are abundant in the non-responder. Since CD8 T cells are the primary effectors in conveying anti-tumor immunity, we would expect to see a significant enrichment of this population in the responders. However, although CD 8 T cells are potent in inducing tumor death, it is more responsive to anti-CTLA-4 immunotherapy. For anti-PD1 therapy, NK cells are more directly related to patients' response[90], which is captured by scDiffPop in the responder population. CXCR4 expression level on NK cells and T cells vary according to the maturation and differentiation stage [91]. The higher CXCR4 expression on NK cells indicates more cells homing to bone marrow compartment. Therefore, high portion of CXCR4+NK cells is associated with less NK infiltration in the tumor which lead to poor prognosis[92]. Thus, it is expected to see the enrichment of CXCR4- NK cells in responders. The significance of RUNX3+ NK cell abundance in responders can also be confirmed by recent studies. RUNX3 is a marker for matured NK cells[93]. Highly expressed in tumor infiltrating NK cells and cytotoxic lymphocytes, RUNX3 is important for immune cell activation and proliferation[94]. In addition to priming CD8 T cells, Naïve CD4 T cells could generate anti-tumor immunity directly. Moreover, CD4 T cells are the major cell-types that induce immunity aginst self-derived epitopes[34]. Similarly, multiple studies indicate that memory T cells, including effector memory, central memory and stem memory T cells, are directly related to persistent anti-tumor immunity[95]. Thus, the enriched cell-types in responders meet our expectations and can be confirmed by published studies. The enrichment of monocyte in non-responders also makes biological sense. Although monocyte population is very heterogeneous composed of many subtypes, in general, the lower the monocyte within blood proportion indicates better response to immunotherapy[96]. Since the Yuen et al. dataset is composed of samples from PBMC of melanoma, there is less sample-associated confounding factors. In most scenarios, ICB data analysis is going to be challenging because the samples are oftentimes mixture of tumor from

multiple organs, that poses difficulty to both annotation and integration. The tissue-difference would also compromise statistical power when conduct DE. Thus, further modification of scDiffPop is needed for future ICB analysis.

2.2.5 Application to PBMCs from COVID-19 patients

Finally, we applied scDiffPop to the timely research on COVID-19. First, we obtained the dataset from Wilk et al.[8] with PBMC data of 7 patients hospitalized with COVID-19 and 6 healthy donors. Based on the publicly available Seurat object with proper annotation and adjusted normalized counts, we directly applied scDiffPop and created the gradient tree (Figure 2.5A). Before zooming into cell-types, we first assessed the hierarchical tree structure, and we found that the tree could generally describe the relationship between cells in PBMC, but it fails for some cell-types. For example, we expected to see CD8 effector T (CD8eff T) be similar to the other T cells, but it branched out two levels earlier than the CD8m T cells. More importantly, there is a decent portion of cells labeled red blood cells (RBCs), which is inaccurate. Since the majority of RBCs lack a nucleus and do not express typical transcriptomes, we would expect a smaller portion. Even there exist nucleated red blood cells, it is also rare. Therefore, it would be reasonable to assume that the inaccurate annotations might account for the wrongly assigned relationship. Given that the majority of the tree overlaps with expectation, we could still trust the differential abundance. As shown in Figure 2.5B, scDiffPop recognized 7 differential abundance cell types between patients hospitalized with COVID-19 and healthy donors. CD14+ monocytes, granulocytes, CD8+ effector T cells, and plasma B cells are identified as enriched with FDR less than 0.05 in COVID-19 patients. Especially, the metapopulation node 18 comprised of the class-switched plasma B cell is significantly differentially abundant in the hospitalized COVID patients, whereas the subpopulation annotated as “B cell” shows no sig-

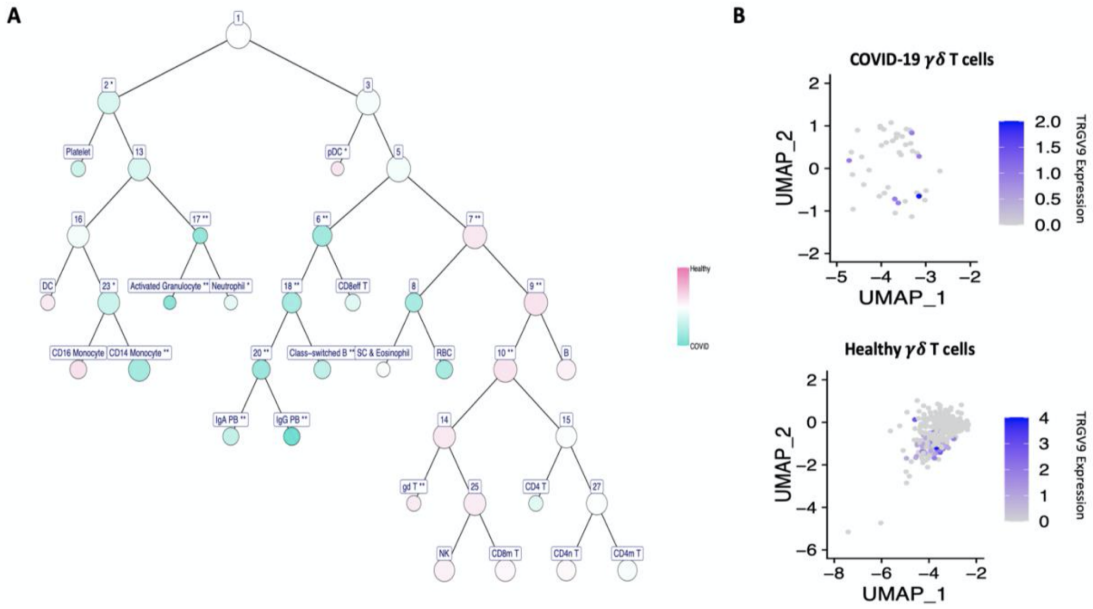

Figure 2.5: **A:** scDiffPop generated gradient tree represents the differences in PBMCs between healthy controls and patients hospitalized with COVID-19 from Wilk et al.[8]. A. The cluster tree inferred by scDiffPop could capture the between cell type relationship, and its prediction can be confirmed by other studies. **B:** $\gamma\delta$T cells (below node 14) are significantly enriched in healthy controls. After a literature search, we analyzed the $\gamma\delta$ T cell subtypes by Plotting the expression of TRGV9. TRGV9 is a biomarker of Vγ9Vδ2 T cells which were predicted to serve a protective role against the virus causing the 2003 SARS epidemic[].

nificance. This differential distribution of B cell sub-population between the two conditions could serve as a diagnostic target. Sosa-Hernandez et. al.[97] confirmed this finding states that the frequency of antibody-secreting B cells increases accordingly with symptom severity. They further proposed that other B cell subsets such as the transitional B cells and memory B cells decrease in server patients compares to mild patients with mild symptoms, which indicates that with more detailed annotation for the "B cell" cluster, we might see some significance. Another exciting observation from the gradience tree is that the T cell population is generally enriched in the samples from healthy donors. Among the T cell subsets, $\gamma\delta$ T cells are found to be significantly abundant in healthy controls with FDR less than 0.05. This finding overlaps with a recent study which is also conducted using PBMC data from 18 healthy and 38 COVID-19 patients, suggesting that $\gamma\delta$ T cells are more abundant in healthy controls and overexpress CD4 upon activation[98]. A study of the SARS-CoV infection during 2003 shows that people who survived this disease shows a population expansion of effector memory $V\gamma 9V\delta 2$ T, which is not found in the more common $\alpha\beta$ T cell population[9]. Moreover, the expansion of the $V\gamma 9V\delta 2$ T cell population shows its protective role against SARS-CoV infection by inducing IFN-dependent anti-SARS-CoV response and involving in the direct infected-cell killing process. Since the virus causing the outbreak in 2003 belongs to the same class as SARS-CoV2, scientists had emphasized on investigating if $\gamma\delta$ T also contributes to anti-COVID-19 immunity. scDiffPop reports T cell receptor gamma variable (TRGV9) as a marker for the if $\gamma\delta$ T population. TRGV9 is the gamma chain of $V\gamma 9V\delta 2$ T cells. As shown on the feature plot of TRGV9, we could see that there are more cells expressing TRGV9 in healthy controls compares to the cells from the COVID-19 samples (Figure 2.5B). Rijkers et al. show that at the time of hospitalization, most of the patients shows less $V\gamma 9V\delta 2$ T cell compares to healthy controls [99]. However, comparing with the general T cell population that has only 8% of the cells

responding, 26% of the $V\gamma 9V\delta 2$ T cell population proliferates and differentiate into effector state after 2 weeks of hospitalization, which indicates Vγ9Vδ2 T cells are most responsive and essential in controlling SARS-CoV2 infection. However, it seems our finding only captures the initially hospitalized state but not represents the recovery situation. Thus, after looking into patients' metadata, we found that among the 7 hospitalized patients, 3 of them requires ventilators, 3 of them did not, and 1 patient does not have a related record. We then applied scDiffPop on the "ventilated" and "non-ventilated" conditions and obtained a few T cell subsets relatively differentially abundant in non-ventilated patients, such as $\gamma\delta$ T cells. This finding could come from the conservative property of scDiffPop, but it could also reflect the homogeneity of the COVID-19 patients for all of these patients are hospitalized less or equal to 10 days, and none of them showed recovery signs.

2.3 Discussion

scDiffPop can extract biologically relevant differential abundant cell types from scRNAseq datasets while controlling for patient specific effects. On the dataset comprised of matched tumor and PBMC samples from four patients, scDiffPop could generate an accurate hierarchical tree to describe the relationships among cell types and report cell types reflecting actual tissue specificity. Moreover, when we randomly shuffled the sample labels of two patients, scDiffPop passed the negative test and identified no cell population is significantly enriched. In comparison with currently available tools that serve the same purpose, we apply both Augur and scDiffPop on this dataset. Although both augur and scDiffPop are able to extract cell subpopulations that are tissue-specific, Augur shows lower statistical power and could not identify the differentially abundant cell-types that have a small sample size. Moreover, when randomly assign the labels of two patients, Augur reports similar or higher AUC as they predicted for

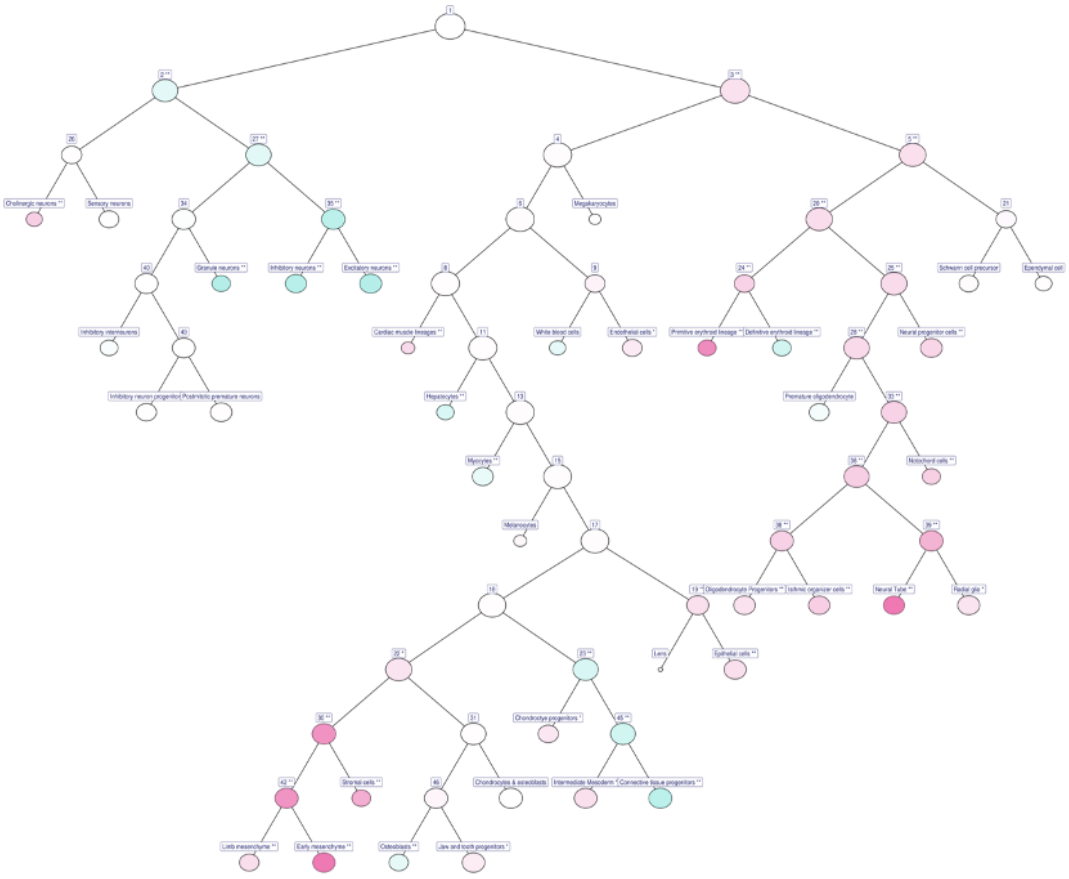

Figure 2.6: The cluster tree generated by scDiffPop to quantify differences in cell populations between mouse embryos at early and late developmental stages. The red represents enrichment in the early stage, and the green represents enrichment in the late stage.

the actual dataset. Thus,the performance of the Augur classifier is affected by sample-specific factors, and the separability might not represent the real biological difference.

To test the scalability, we perform scDiffPop on the MOCA dataset with 2 million cells. When dividing the cells based on development stages, scDiffPop was able to find the signature cell types of different stages to be significantly differentially abundant. When applying scDiffPop to cells from the same stage that are arbitrarily assigned to two groups, no cell population is reported to be significantly different. This test shows that well-controlled on sample-specific effects, scDiffPop could differentiate the real and false-positive abundance. The hierarchical tree generated by scDiffPop based on the annotation and gene profile could reflect biological relevance, and the differential abundance analysis on the metapopulation node could handle cell-type with a limited number of cells. Moreover, scDiffPop is very scalable to the datasets with a large number of cells, for the run time of scDiffPop is only associated with the number of cell types included in the annotation, the number of samples, and the number of genes in the library. Since it performs differential expression analysis on a sample pseudobulk level on each node, the cell type that determines the tree depth and size and the number of samples would influence how many DEA steps are required. Although the permutation test takes a longer time than the parametric test, scDiffPop can be run in a parallelized way to solve the lag. It takes scDiffPop fifteen minutes to run on a small dataset such as blood/tumor sample with 12 cell-types and three hours on a large dataset such as the MOCA dataset with 37 cell-types (on 64 CPUs with 300 GB).

Then, we want to use scDiffPop to explore biological problems and applied it to COVID19 data and ICB datasets. We find that the abundant differential cell-types overlap with the current understanding of COVID-19, which gives us confidence in our model and result. By taking a closer look, we find $\gamma\delta$ T cell over-expressed in healthy donors is most interesting. This

finding echos the precious knowledge about Vγ9Vδ2 T cell showing a protective role against SARS-CoV in 2003. Thus, we further checked the marker gene of $\gamma\delta$ T cell population and confirmed that Vγ9Vδ2 T cells are the dominantly different population that contributes to $\gamma\delta$ T cell differentiation. To proceed with investigating how V γ 9Vδ2 T cells serve as a protective barrier against SARS-CoV-2, we applied the method to larger COVID-19 patients and checked whether we could see other $\gamma\delta$ T cell sub-types enriched in different disease stage and recovery stage.

2.4 Methods

2.4.1 Data Pre-processing

When analyzing the publicly available data, using the already processed dataset with annotation confirmed by flow cytometry is preferred. If performed on the original dataset, careful sequencing and alignment quality control are required to obtain the gene expression profile. After acquired the expression counts matrix, sample barcode, and features list, we recommend using the R package Seurat[100] and SeuratDisk[101] . We follow the commonly applied processing pipeline for scRNA-Seq data pre-processing. Before creating Seurat objects, we filtered out features expressed less than 3 cells and cells with high mitochondrial gene portion. After creating the Seurat object, we normalize and scale the data and perform the PCA on 2000 variable features. Then we use Harmony to correct PC scores for multiple experimental and biological confounding factors[102] . We then run UMAP on the corrected PCA projection and use the DimPlot function to plot the UMAP for annotation accuracy check.

Since an accurate annotation is essential for scDiffPop performance, we perform careful annotation before scDiffPop application. We use the scRNA-Seq PBMC dataset[103] with

92,000 of 10 cell types properly annotated as the reference dataset and use the method developed in the Seurat v3 integration a label-transferring to generate rough annotation first[72] . Based on the primary annotation, we then create a hierarchical tree and UMAP to confirms the annotation. Once we have confidence about the original annotation, we then check the annotation proportion within UMAP-clusters. If one cluster has over 70% cell population of one predicted-annotation, we directly give these clusters the corresponding cell-types. For those mixture clusters, we plot the marker genes for each subtype using the FeaturePlot function (Table 1). We also use the FindMarkers function to obtain a list of differentially expressed genes and compares it with the gene profile of each cell-type provided by Human Cell Atalas[104] . Integrating all the cell markers information, we are then able to make a more confident annotation to the mixture cluster.

2.4.2 Building a cluster tree

We constructed the cluster tree using divisive clustering. We first assign all cell-types to the same meta-population and recursively split the large group into smaller binary subpopulations. The algorithm keeps the recursive step until all cell annotations are assigned to the leave of the cluster tree. The following procedure is used to split a group of k $\geq$ 2 cell types into two subgroups: 1. Use harmonization method to normalize the expression matrix for all cells in the metagroup and perform principal component analysis (PCA) on the normalized matrix. Keep the top 50 PC scores. 2. For 1$\leq$ i $\leq$ k, define $\bar{y}_i \in \mathbb{R}^{50}$ to be the average PC score for cell type *i*. 3. Perform 2-means clustering on $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$. The cell type to cluster assignment generates the two distinct subgroups.

The third step implies that our current tree is binary. However, this binary tree is not practical for large and complex datasets. Thus, we also provide flexibility for cluster trees with

more nodes on the same level. Moreover, our method is heavily dependent on the PCA projection, which is inefficient on a large dataset. Thus, we sample 100 cells from each annotated subgroup to perform Step1.

2.4.3 Differential expression analysis

At sample-level pseudobulk, we performed the differential expression analysis at each metapopulation in the cluster tree denoted as T. Since the Annotated cell types will always be on the leaf nodes in the cluster tree, we define the cell population at an internal node v $\in$T as the leaves in the tree rooted at v. To begin with, we use the FindMarker function from Seurat to identify the marker genes of the cell population. By default, we are going to keep the top N = 25 overexpressed genes with positive LFC an padj <0.01. We then compute the average of the Log2-fold-change value of these 25 marker genes denoted as x_i . The strength of marker i is calculated by $S(x_i) = \frac{x_i}{Max_{1<j<N} x_j}$, where the denominator is the maximum value of the 25 log2-fold-change average. In other words, the strength of each marker is defined as the fraction of LFC of that marker to the aptitude of the LFC of all markers so that the top marker always has a strength score of 1.

As noted, we first need to find the sample-level pseudobulk to proceed for differential expression analysis. We add up the counts from all cells labeled with the same sample to obtain the number of reads for each gene in the sample-pseudobulk. The acquired count matrix has rows of genes and columns of samples in the experiment.

We then apply DESeq2 for differential expression analysis on the pseudobulk count matrix. Although the only important genes we are looking at in further analysis are the differentially expressed marker genes, since DESeq2 would apply partial pooling to shrink sample difference, all genes must be kept. DESeq2 returns a Wald statistic to quantify the differen-

tial expression strength of each gene. Let t_i be the Wald statistic for i-th marker of one cell population. Then the test statistic for this sample from scDiffPop can be written as

$$\frac{1}{N}\sum i = 1^N t_i S(x_i) \tag{2.1}$$

If the scDiffPop statistics is large in magnitude, we state that this associated cell subpopulation is overexpressed in one of the two conditions, and the sign of 2.1 determines which condition this cell-type is enriched in.

Then, we compute the probability of observing a more extreme value than the scDiffPop statistic under the null hypothesis that sample labels are randomly assigned among the cell type. However, since the Wald statistics of marker genes are dependent on each other, the scDiffPop statistics are not parametrically interpretable. Thus, we applied the permutation test for p-value estimation. In this test, we would be able to randomly assign the label of two conditions and obtain the test statistic by calculating all possible label arrangements of observation under the null hypothesis. Let t_{actual} be the actual test statistic and let $t_1, t2, \ldots, t_p$ be P simulated test statistics obtained by recomputing the scDiffPop statistic using permutation test, where by default, we simulate $n = 250$ iterations. Let $P' = |t_i : |t_i| > |t_{actual}||$ be the number of simulated test statistics that are more extreme than t_{actual}. As shown on the paper published by Phipson and Smyth in 201070 , $\frac{P'}{P}$ is an unbiased estimator for p value, but does not control for type I error. Thus, we use the following formula estimator:

$$\frac{P'+1}{N+1} \tag{2.2}$$

This modified estimator for p-value correctly controls for the type I error rate. Since by default 250 iterations of permutations are performed, P is a vector of length 250. However, we encourage more permutation iterations for a better estimation. Also, since our estimation is based on multiple cell types and multiple marker genes, we would need to further adjust the p

value for multiple testing use the Benjamini-Hochberg (BH adjusted p-value)[82] which is also known as the False discovery rate (FDR). We set the threshold of 0.05, so that cell populations showing FDR less than the cutoff are significantly enriched.

2.4.4 Visualizing the results

We utilize the CRAN packages igraph[105] and ggraph[106] to plot the gradient and pie cluster tree. Let m_i denote the size of the cell population corresponding to node i, then the node size is proportional to $\frac{log_2(m_i)}{log_2(m)}$, where m is the total number of cells sequenced.

2.4.5 Installing scDiffPop

Although the scDiffPop is still under development, we have an available version on Github: https://github.com/phillipnicol/scDiffPop One can install this to R using the devtool command: *devtools::install_github("phillipnicol/scDiffPop")*

Chapter 3

Discussion and Perspective

3.1 Discussion

In this paper, we provide a powerful yet straightforward tool for integrative scRNA-Seq analysis called scDiffPop. This method is based on a robust statistical model for differential abundance analysis. The current method on identifying the abundant cell-types determines the statistical significance by comparing cell proportion difference on sample-level. Since this method ignored the cell-level information, it does not have strong statistical power. Our method aims to utilize more cell-level information to solve two challenges in RNA-Seq analysis:1) the small sample size of some cell type compromises statistical power, and 2) sample specific-factors (such as sex and age) would confound finding. To solve the first issue, we use the gene-expression data to generate a cluster tree to delineate the relationship among annotated cell populations and perform analysis on the metapopulation instead to gain significance. To mitigate the sample specificity, we conduct differential expression analysis on patient-level and conduct an n=250 permutation test to validate the statistics' significance. Moreover, compares to other published methods such as Augur that determine the significance based on separabil-

ity, scDiffPop implement the well-accepted DESeq2 as the foundation to find the differentially expressed gene markers, which allows correction for batch effect by passing a design matrix parameter into the DESeq function.

To validate the accuracy and efficiency of scDiffPop, we applied both Augur and scDiffPop on the matching tumor/blood sample from four patients. Although both Augur and scDiffPop could find significantly enriched cell types and the majority overlaps, they report drastically different results for the mixed model. In the mixed experiment with the label of the cells from two patients swapped, we expect to see no significant difference for the miss-labeling significantly diminished the biological difference between the tumor and blood tissue specificity. However, Augur shows similar or higher significance comparing to the prediction from training on the original dataset. On the other hand, scDiffPop shows no significant population in the mixed experiment. Thus, from the semi-simulated negative test, we conclude that the classifier trained by Augur capturing the sample-specific factors failed to extract the real biological relevance from background noise. scDiffPop shows robustness against these confounding effects and provides many reliable predictions.

While applying scDiffPop to the MOCA dataset with 2 million cells, we demonstrate the scalability of scDiffPop. Independent of how many cells are sequenced, the running time of scDiffPop is associated with how many cell-types, feature numbers, and sample numbers. Moreover, to run a large dataset more efficiently, scDiffPop only takes 100 cells from each celltype for gene marker identification. After applied scDiffPop to MOCA dataset, we identified 25 cell populations being significant. We found the cell subpopulations enriched in the early stage are related to early central nerve system (CNS) development and hematopoiesis. Then, we perform a negative test on the cells from the same developmental stage. As expected, scDiffPop reports no cell type being significant. Moreover, we conducted the mixed labeling

test on the MOCA dataset and observed that although Augur and scDiffPop report overlapping cell-type list, Augur shows significance for many cell-types (Cholinergic neurons, stromal cells, excitation neurons, etc.) under the miss-labeled situation, whereas scDiffPop, once again, identified no significantly enriched subpopulation (Supplement 2). Another interesting finding from this additional test is that the Augur predicts the cell types with much lower significance for the original dataset. Thus, many cell populations that are not identified by Augur could be captured by scDiffPop with much higher significance.

This method is particularly designed to identify cell populations that differ between two conditions. We applied the method to the COVID-19 dataset and ICB datasets. From the differential abundant cell populations enriched in the healthy donor comparing with the hospitalized patients, we are particularly interested in the $\gamma\delta$T cells. The $\gamma\delta$T cells were identified as playing a protective role against SARS-CoV infection in the 2003 outbreak. After sub-setting the γT cells based on the T cell receptor isoforms, we find the significance of the $\gamma\delta$T cell population is mainly coming from the Vγ9Vδ2 T cell subset. This subpopulation is also identified to be essential to anti-COVID infection and expanded after COVID-19 infection. However, since patients of our dataset were hospitalized less than 2 weeks and did not show recovery, the expansion is not shown in our analysis.

The application of scDiffPop to the ICB dataset is challenging. Although scDiffPop was able to provide enriched cell types that are known to be essential for anti-PD-1 immunotherapy response, such as the NK and CD68- monocyte. The enrichment in the non-responder condition looks counterintuitive. For example, dendritic cell-mediated antigen presentation and T cell activation that has been used as responsive markers are enriched in non-responders. These wrongly predicted cell types are probably because of the large variation of the ICB dataset (different tumor type), which indicates that our model still needs more control for this large

between-sample variation.

scDiffPop is still in its developmental stage with many imperfections, but we have provided enough validation tests to show its power in this preliminary state. We believe it could be a powerful tool for integrative analysis for scRNA-Seq analysis. Although scRNA-Seq has been widely adopted for conditions comparison experiments, because of the technical noise, batch effects, and sample specificity, researchers cannot perform analysis directly on the dataset. Taking advantage of DESeq2, scDiffPop provides the flexibility to design a model to control for batch effects and any other cohort-specific effects. Moreover, scDiffPop perform DEA on the pooled metapopulation that circumvents Augur's limitation on training classifier at small subpopulation. Currently, we are attempting to perform integrative analysis on all publicly available COVID-19 scRNA-seq datasets.

Extensive tests are still needed to validate scDiffPop in various settings. For example, it is important to test the sensitivity of scDiffPop to variation of input parameters such as the number of biomarkers used and the number of permutations used in p-value estimation.Extensive tests with simulated or more detailed datasets are necessary so that we could better understand the strengths and weaknesses of scDiffPop. Moreover, it is also important to provide more method comparison to justify the efficiency and accuracy of scDiffPop. In this paper, we only include the comparison test of scDiffPop to Augur, but as mentioned in the introduction, there are many other available tools. Thus, we should include at least 3 other methods to conduct a similar analysis.

Moreover, as mentioned previously, with the guidance from our current analysis, we are still having a hard to give a self-contained conclusion. Thus, we need to acquire more COVID-19 datasets with patients from different disease stages so that we could delineate the timely trajectory of the immune cell populations. In that, we could capture which cell population

is experiencing expansion at which disease stage. This study is going to be especially more interesting on recovered patients since they could provide matched PBMC samples of all time. By identifying the cell types associated with recovery, this analysis could provide valuable guidance for advancing vaccine development during the global pandemic situation. We also hypothesized that the lower Vγ9Vδ2 T cell frequency in the PBMC sample is because the majority of the $\gamma\delta$ T cells are recruited at the infected lung tissue site. Thus, we would need to obtain matched PBMC and lung tissue samples to prove our hypothesis.

Since the scDiffPop is especially dependent on the annotation accuracy to deal with the inaccurate annotation problem, we are planning on integrating the annotation method to provide reannotated labels for tree construction. We are going to combine the label transfer method adopted in our analysis to scDiffPop with freedom on manual adjustment. The annotation method is based on the Seurat integration and labels transferring vignette, where Seurat object from different datasets could be adjusted for batch effect and merged. Transferring the label in the provided reference dataset, the method will find anchors and give annotation at the cell-level. Based on the cell level annotation, we assign the UMAP-clusters with 70% cells labeled with the same prediction as that cell-type. For the cluster with mixture labels, we plot the marker gene dot plots for manual adjustment.

Moreover, the tree that we are using now is binary, which would compromise scDiffPop's ability to capture the real among cell relationship. Thus, we would expand our tree model to more flexible structures. With the scaffold well established, we are trying to integrate clusterProfiler[107] for pathway analysis in each of the nodes. Furthermore, we are exploring a better visualization strategy and give users the ability to adjust at what level of the hierarchical tree they would like to plot.

Bibliography

1. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14,** 91. ISSN: 14712105 (1 Mar. 2013).

2. Yuen, K. C. *et al.* High systemic and tumor-associated IL-8 correlates with reduced clinical benefit of PD-L1 blockade. *Nature Medicine* **26,** 693–698. ISSN: 1546170X (5 May 2020).

3. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nature medicine,* 1–13. ISSN: 1546-170X (Apr. 2021).

4. *Strategies of tumor immune escape from NK cell-dependent... | Download Scientific Diagram* https://www.researchgate.net/figure/Strategies-of-tumor-immune-escape-from-NK-cell-dependent-immunosurveillance-NK-cell_fig1_51131243.

5. et.al., G.-A. A. Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biology* **18** (1 2017).

6. Luecke Malte D, T. F. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15** (6 2019).

7. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566,** 496–502. ISSN: 14764687 (7745 Feb. 2019).

8. Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nature Medicine* **26,** 1070–1076. ISSN: 1546170X (7 July 2020).

9. Poccia, F. *et al.* Anti-severe acute respiratory syndrome coronavirus immune responses: The role played by Vγ9Vδ2 T cells. *Journal of Infectious Diseases* **193,** 1244–1249. ISSN: 00221899 (9 May 2006).

10. Muniyappa, R. & Gubbi, S. COVID-19 pandemic, coronaviruses, and diabetes mellitus. *American Journal of Physiology - Endocrinology and Metabolism* **318,** E736–E741. ISSN: 15221555 (5 May 2020).

11. Hu, B., Guo, H., Zhou, P. & Shi, Z. L. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology* **19,** 141–154. ISSN: 17401534 (3 Oct. 2020).

12. Shang, J. *et al.* Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 11727–11734. ISSN: 10916490 (21 May 2020).

13. Brodin, P. Immune determinants of COVID-19 disease presentation and severity. *Nature Medicine* **27,** 28–33. ISSN: 1546170X (1 Jan. 2021).

14. Guo, C. *et al.* Single-cell analysis of two severe COVID-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm. *Nature Communications* **11.** ISSN: 20411723.

15. Schulte-Schrepping, J. *et al.* Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell* **182,** 1419–1440.e23. ISSN: 10974172 (6 Sept. 2020).

16. Silvin, A., Chapuis, N., Dunsmore, G., Cell, A. G. & undefined 2020. Elevated calprotectin and abnormal myeloid cell subsets discriminate severe from mild COVID-19. *Elsevier.*

17. Wen, W. *et al.* Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. *nature.com.*

18. Zhang, F. *et al.* Adaptive immune responses to SARS-CoV-2 infection in severe versus mild individuals. *nature.com.*

19. Zhang, J. *et al.* Single-cell landscape of immunological responses in patients with COVID-19. *nature.com.*

20. Bastard, P. *et al.* Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* **370.** ISSN: 10959203 (6515 Oct. 2020).

21. Meager, A. *et al.* Anti-interferon autoantibodies in autoimmune polyendocrinopathy syndrome type 1. *PLoS Medicine* **3,** 1152–1164. ISSN: 15491277 (7 2006).

22. Stertz, S. & Hale, B. G. Interferon system deficiencies exacerbating severe pandemic virus infections. *Trends in microbiology.* ISSN: 1878-4380 (Mar. 2021).

23. Grifoni, A. *et al.* Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. *Cell* **181,** 1489–1501.e15. ISSN: 10974172 (7 June 2020).

24. Ren, X. *et al.* Large-scale single-cell analysis reveals critical immune characteristics of COVID-19 patients. *bioRxiv,* 2020.10.29.360479. ISSN: 26928205 (Oct. 2020).

25. Laing, A. G. *et al.* A consensus Covid-19 immune signature combines immuno-protection with discrete sepsis-like traits associated with poor prognosis. *medRxiv,* 2020.06.08.20125112 (June 2020).

26. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Elsevier.*

27. Giamarellos-Bourboulis, E., Netea, M., host & . . ., N. R. C. & undefined 2020. Complex immune dysregulation in COVID-19 patients with severe respiratory failure. *Elsevier.*

28. Tan, L. *et al.* Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. *nature.com.*

29. Chen, G., Wu, D., Guo, W., of . . ., Y. C. T. J. & undefined 2020. Clinical and immunological features of severe and moderate coronavirus disease 2019. *Am Soc Clin Investig.*

30. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature Medicine* **26,** 842–844. ISSN: 1546170X (6 June 2020).

31. Diao, B. *et al.* Reduction and Functional Exhaustion of T Cells in Patients With Coronavirus Disease 2019 (COVID-19). *Frontiers in Immunology* **11,** 827. ISSN: 16643224 (May 2020).

32. Zheng, M. *et al.* Functional exhaustion of antiviral lymphocytes in COVID-19 patients. *Cellular and Molecular Immunology* **17,** 533–535. ISSN: 20420226 (5 May 2020).

33. Yu, K. *et al.* Thymosin alpha-1 Protected T Cells from Excessive Activation in Severe COVID-19.

34. Henley, S. J. *et al.* Annual report to the nation on the status of cancer, part I: National cancer statistics. *Cancer* **126,** 2225–2249. ISSN: 10970142 (10 May 2020).

35. Berger, C. L. *et al.* Cutaneous T-cell lymphoma: Malignant proliferation of T-regulatory cells. *Blood* **105,** 1640–1647. ISSN: 00064971 (4 Feb. 2005).

36. Seliger, B. Strategies of tumor immune evasion. *BioDrugs* **19,** 347–354. ISSN: 11738804 (6 Aug. 2005).

37. Abe, B. T. & Macian, F. Uncovering the mechanisms that regulate tumor-induced T-cell anergy. *OncoImmunology* **2.** ISSN: 21624011 (2 Feb. 2013).

38. Janeway, J. C. A., Travers, P., Walport, M. & Shlomchik, M. J. T cell-mediated cytotoxicity (2001).

39. Chang, J. T., Wherry, E. J. & Goldrath, A. W. Molecular regulation of effector and memory T cell differentiation. *Nature Immunology* **15,** 1104–1115. ISSN: 15292916 (12 Nov. 2014).

40. Joller, N. *et al.* Cutting Edge: TIGIT Has T Cell-Intrinsic Inhibitory Functions. *The Journal of Immunology* **186,** 1338–1342. ISSN: 0022-1767 (3 Feb. 2011).

41. Fourcade, J. *et al.* CD8 + T cells specific for tumor antigens can be rendered dysfunctional by the tumor microenvironment through upregulation of the inhibitory receptors BTLA and PD-1. *Cancer Research* **72,** 887–896. ISSN: 00085472 (4 Feb. 2012).

42. Blackburn, S. D. *et al.* Coregulation of CD8+ T cell exhaustion by multiple inhibitory receptors during chronic viral infection. *Nature Immunology* **10,** 29–37. ISSN: 15292908 (1 2009).

43. Crawford, A. & Wherry, E. J. The diversity of costimulatory and inhibitory receptor pathways and the regulation of antiviral T cell responses. *Current Opinion in Immunology* **21,** 179–186. ISSN: 09527915 (2 Apr. 2009).

44. Jin, H. T. *et al.* Cooperation of Tim-3 and PD-1 in CD8 T-cell exhaustion during chronic viral infection. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 14733–14738. ISSN: 00278424 (33 Aug. 2010).

45. Barber, D. L. *et al.* Restoring function in exhausted CD8 T cells during chronic viral infection. *Nature* **439,** 682–687. ISSN: 00280836 (7077 Feb. 2006).

46. Jiang, Y., Li, Y. & Zhu, B. T-cell exhaustion in the tumor microenvironment. *Cell Death and Disease* **6,** e1792–e1792. ISSN: 20414889 (6 June 2015).

47. Sakaguchi, S. *et al.* Immunologic tolerance maintained by CD25+ CD4+ regulatory T cells: Their common role in controlling autoimmunity, tumor immunity, and transplantation tolerance. *Immunological Reviews* **182,** 18–32. ISSN: 01052896 (2001).

48. Deng, G. Tumor-infiltrating regulatory T cells: origins and features. *American journal of clinical and experimental immunology* **7,** 81–87. ISSN: 2164-7712 (5 2018).

49. Halvorsen, E. C. *et al.* Maraviroc decreases CCL8-mediated migration of CCR5+ regulatory T cells and reduces metastatic tumor growth in the lungs. *OncoImmunology* **5.** ISSN: 2162402X (6 June 2016).

50. Ward, S. T. *et al.* The effects of CCR5 inhibition on regulatory T-cell recruitment to colorectal cancer. *British Journal of Cancer* **112,** 319–328. ISSN: 15321827 (2 Jan. 2015).

51. Mittal, S. *et al.* Local and systemic induction of CD4 +CD25 + regulatory T-cell population by non-Hodgkin lymphoma. *Blood* **111,** 5359–5370. ISSN: 00064971 (11 June 2008).

52. Han, Y. *et al.* Malignant B cells induce the conversion of CD4 +CD25 - T cells to regulatory T cells in B-cell non-Hodgkin lymphoma. *PLoS ONE* **6.** ISSN: 19326203 (12 Dec. 2011).

53. Zheng, S. G. *et al.* TGF-β Requires CTLA-4 Early after T Cell Activation to Induce FoxP3 and Generate Adaptive CD4 + CD25 + Regulatory Cells. *The Journal of Immunology* **176,** 3321–3329. ISSN: 0022-1767 (6 Mar. 2006).

54. Downs-Canner, S. *et al.* Suppressive IL-17A+ Foxp3+ and ex-Th17 IL-17Aneg Foxp3+ Treg cells are a source of tumour-associated Treg cells. *Nature Communications* **8.** ISSN: 20411723 (Mar. 2017).

55. Pacella, I. *et al.* Fatty acid metabolism complements glycolysis in th selective regulatory t cell expansion during tumor growth. *Proceedings of the National Academy of Sciences of the United States of America* **115,** E6546–E6555. ISSN: 10916490 (28 July 2018).

56. Loser, K. *et al.* IL-10 Controls Ultraviolet-Induced Carcinogenesis in Mice. *The Journal of Immunology* **179,** 365–371. ISSN: 0022-1767 (1 July 2007).

57. Strauss, L. *et al.* A unique subset of CD4+CD25highFoxp3+ T cells secreting interleukin-10 and transforming growth factor-β1 mediates suppression in the tumor microenvironment. *Clinical Cancer Research* **13,** 4345–4354. ISSN: 10780432 (15 Aug. 2007).

58. Facciabene, A. *et al.* Tumour hypoxia promotes tolerance and angiogenesis via CCL28 and T reg cells. *Nature* **475,** 226–230. ISSN: 00280836 (7355 July 2011).

59. Peng, Q. *et al.* PD-L1 on dendritic cells attenuates T cell activation and regulates response to immune checkpoint blockade. *Nature Communications* **11,** 1–8. ISSN: 20411723 (1 Dec. 2020).

60. *B cells to the forefront of immunotherapy* `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7523515/`.

61. Sautès-Fridman, C. *et al.* Tertiary lymphoid structures in cancers: Prognostic value, regulation, and manipulation for therapeutic intervention. *Frontiers in Immunology* **7,** 407. ISSN: 16643224 (OCT Oct. 2016).

62. Marshall, J. S., Warrington, R., Watson, W. & Kim, H. L. An introduction to immunology and immunopathology. *Allergy, Asthma and Clinical Immunology* **14,** 49. ISSN: 17101492 (Suppl 2 Sept. 2018).

63. Cabrita, R. *et al.* Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature* **577,** 561–565. ISSN: 14764687 (7791 Jan. 2020).

64. Kwun, J. *et al.* Crosstalk between T and B Cells in the Germinal Center after Transplantation. *Transplantation* **101,** 704–712. ISSN: 00411337 (4 Apr. 2017).

65. Hoeres, T., Smetak, M., Pretscher, D. & Wilhelm, M. Improving the efficiency of Vγ9Vδ2 T-cell immunotherapy in cancer. *Frontiers in Immunology* **9.** ISSN: 16643224 (APR Apr. 2018).

66. Den Berge, K. V. *et al.* RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. *Annual Review of Biomedical Data Science* **2,** 139–173. ISSN: 2574-3414 (1 July 2019).

67. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics* **20,** 631–656. ISSN: 14710064 (11 Nov. 2019).

68. Ahmed, W., Zheng, K. & Liu, Z. F. Small non-coding RNAs: New insights in modulation of host immune response by intracellular bacterial pathogens. *Frontiers in Immunology* **7.** ISSN: 16643224 (OCT Oct. 2016).

69. Byron, S. A., Keuren-Jensen, K. R. V., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nature Reviews Genetics* **17,** 257–271. ISSN: 14710064 (5 May 2016).

70. Method of the Year 2013. *Nature Methods* **11,** 1. ISSN: 15487105 (1 Dec. 2014).

71. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine* **50,** 96. ISSN: 20926413 (8 Aug. 2018).

72. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data Resource Comprehensive Integration of Single-Cell Data. *Cell* **177** (2019).

73. Hou, R., Denisenko, E. & Forrest, A. R. R. scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* **35** (ed Kelso, J.) 4688–4695. ISSN: 1367-4803 (22 Nov. 2019).

74. Rho, K. *et al.* Garnet - gene set analysis with exploration of annotation relations. *BMC Bioinformatics* **12,** S25. ISSN: 14712105 (SUPPL. 1 Feb. 2011).

75. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* (Feb. 2018).

76. *t-SNE – Laurens van der Maaten* `https://lvdmaaten.github.io/tsne/`.

77. Chua, R. L. *et al.* COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nature Biotechnology* **38,** 970–979. ISSN: 15461696 (8 Aug. 2020).

78. Skinnider, M. A. *et al.* Cell type prioritization in single-cell data. *bioRxiv,* 2019.12.20.884916 (Dec. 2019).

79. Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C. & Popp, J. Sample size planning for classification models. *Analytica Chimica Acta* **760,** 25–33. ISSN: 00032670 (Jan. 2013).

80. Zhao, J., Jaffe, A., Cheng, X., Flavell, R. & Kluger, Y. Detecting regions of differential abundance between scRNA-Seq datasets. *bioRxiv,* 711929 (July 2019).

81. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550. ISSN: 1474760X (12 Dec. 2014).

82. Haynes, W. *Benjamini–Hochberg Method* 78–78 (Springer New York, 2013).

83. Lin, Y., Xu, J. & Lan, H. Tumor-associated macrophages in tumor metastasis: Biological roles and clinical therapeutic applications. *Journal of Hematology and Oncology* **12,** 1–16. ISSN: 17568722 (1 July 2019).

84. Diao, J., Zhao, J., Winter, E. & Cattral, M. S. Tumors Conventional Dendritic Cell Precursors in Recruitment and Differentiation of. *J Immunol References* **184,** 1261–1267 (2021).

85. Otranto, M. *et al.* The role of the myofibroblast in tumor stroma remodeling. *Cell Adhesion and Migration* **6,** 203–219. ISSN: 19336926 (3 2012).

86. Bailey, S. R. *et al.* Th17 cells in cancer: The ultimate identity crisis. *Frontiers in Immunology* **5.** ISSN: 16643224 (JUN 2014).

87. Palis, J. Primitive and Definitive Erythropoiesis. *Blood* **120,** SCI-37-SCI–37. ISSN: 0006-4971 (21 Nov. 2012).

88. Elshazzly, M. & Caban, O. *Embryology, Central Nervous System* (StatPearls Publishing, Apr. 2019).

89. Kauts, M. L., Vink, C. S. & Dzierzak, E. Hematopoietic (stem) cell development — how divergent are the roads taken? *FEBS Letters* **590,** 3975–3986. ISSN: 18733468 (22 Nov. 2016).

90. Subrahmanyam, P. B. *et al.* Distinct predictive biomarker candidates for response to anti-CTLA-4 and anti-PD-1 immunotherapy in melanoma patients. *Journal for ImmunoTherapy of Cancer* **6,** 18. ISSN: 20511426 (1 Mar. 2018).

91. Susek, K. H., Karvouni, M., Alici, E. & Lundqvist, A. The role of CXC chemokine receptors 1–4 on immune cells in the tumor microenvironment. *Frontiers in Immunology* **9,** 2159. ISSN: 16643224 (Sept. 2018).

92. Levy, E. *et al.* Enhanced Bone Marrow Homing of Natural Killer Cells Following mRNA Transfection With Gain-of-Function Variant CXCR4R334X. *Frontiers in Immunology* **10,** 1262. ISSN: 1664-3224 (JUN June 2019).

93. Wang, D. & Malarkannan, S. *Transcriptional regulation of natural killer cell development and functions* June 2020.

94. Manandhar, S. & Lee, Y. M. Emerging role of RUNX3 in the regulation of tumor microenvironment. *BMB Reports* **51,** 174–181. ISSN: 1976670X (4 Apr. 2018).

95. Liu, Q., Sun, Z. & Chen, L. Memory T cells: strategies for optimizing tumor immunotherapy. *Protein and Cell* **11,** 549–564. ISSN: 16748018 (8 Aug. 2020).

96. Nixon, A. B. *et al.* Peripheral immune-based biomarkers in cancer immunotherapy: can we realize their predictive potential? *Journal for ImmunoTherapy of Cancer* **7.** ISSN: 20511426 (1 Nov. 2019).

97. Sosa-Hernández, V. A. *et al.* B Cell Subsets as Severity-Associated Signatures in COVID-19 Patients. *Frontiers in Immunology* **11,** 1. ISSN: 1664-3224 (Dec. 2020).

98. Lei, L. *et al.* The phenotypic changes of $\gamma\delta$ T cells in COVID-19 patients. *medRxiv,* 2020.04.05.20046433. ISSN: 1582-4934 (Apr. 2020).

99. Rijkers, G., Vervenne, T. & van der Pol, P. More bricks in the wall against SARS-CoV-2 infection: involvement of $\gamma 9\delta 2$ T cells. *Cellular and Molecular Immunology* **17,** 771–772. ISSN: 20420226 (7 July 2020).

100. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36,** 411–420. ISSN: 15461696 (5 June 2018).

101. *SeuratDisk: Interfaces for HDF5-Based Single Cell File Formats — SeuratDisk-package* • *SeuratDisk* https://mojaveazure.github.io/seurat-disk/reference/SeuratDisk-package.html.

102. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16,** 1289–1296. ISSN: 15487105 (12 Dec. 2019).

103. Ding, J. *et al.* Systematic comparative analysis of single cell RNA-sequencing methods. *bioRxiv,* 632216 (May 2019).

104. Regev, A. *et al.* The human cell atlas. *eLife* **6.** ISSN: 2050084X (Dec. 2017).

105. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems,** 1695. https://igraph.org (2006).

106. *CRAN - Package ggraph* https://cran.r-project.org/web/packages/ggraph/index.html.

107. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology* **16,** 284–287. ISSN: 15362310 (5 May 2012).

Supplementary Figures

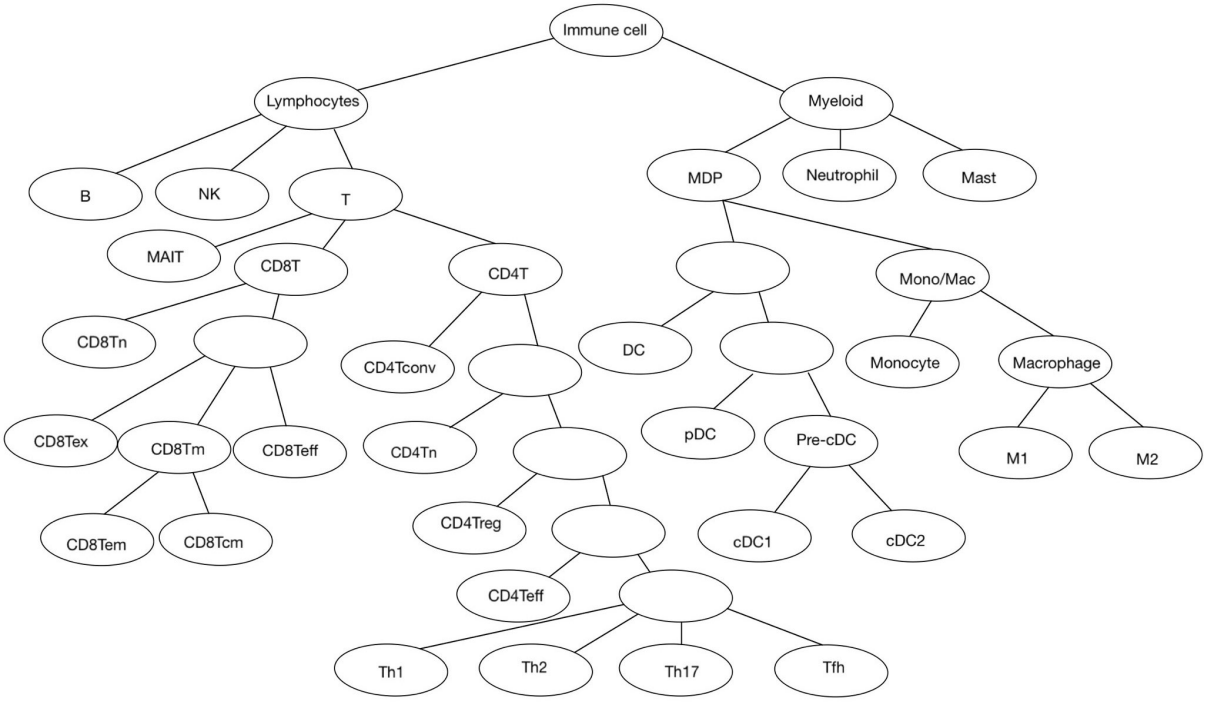

Figure 3.1: Customized immune cell hierarchical tree.

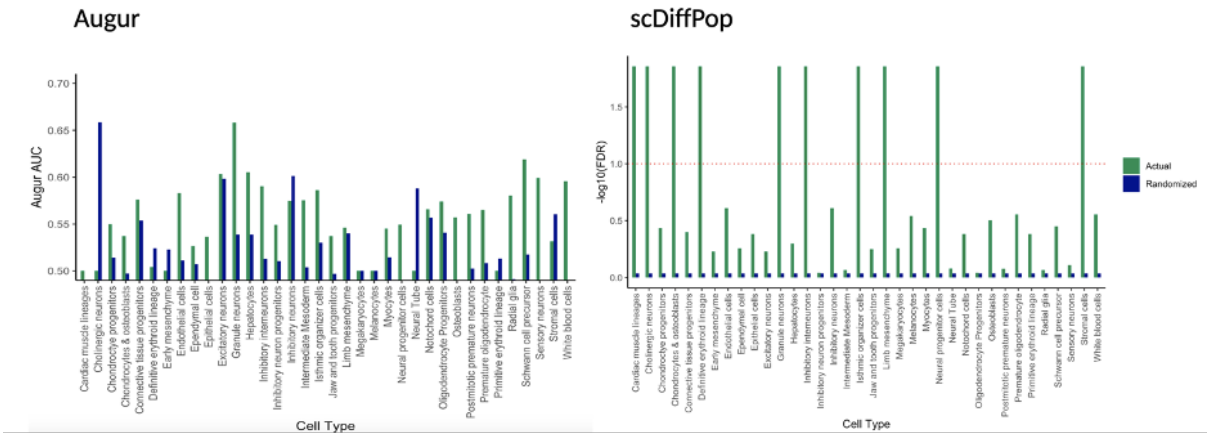

Figure 3.2: MOCA comparison test between scDiffPop and Augur