Joel Moro Borrego

Innovative Statistical Frameworks for Enhanced V(D)J Repertoire Analysis

MASTER'S THESIS

supervised by Dr. Joan T. Matamalas and Dr. Masanori Aikawa

Master's Degree in Biomedical Data Science

Barcelona, 2025

Abstract

Single-cell immune repertoire analysis is essential for understanding adaptive immune responses, yet reconstructing single-cell V(D)J sequences from single-cell whole transcriptome RNA-sequencing (WTA) data presents significant challenges due to the short-read nature of 3′ sequencing and platform-specific limitations. In this study, we developed and evaluated a robust computational framework to reconstruct T-cell and B-cell receptor sequences using TRUST4, alongside a custom pipeline to address barcode-to-index translation for BD Rhapsody data. Despite the flexibility and error-tolerant capabilities of the framework, limitations inherent to BD Rhapsody, including the prevalence of multiplets and the incompatibility of standard tools optimized for 10x Genomics, reduced the statistical power of V(D)J reconstruction. While approximately 12–15% of cells were reconstructed by TRUST4, only 3–7% of total cells in the input datasets were fully recovered after accounting for barcode complexities.

We also compared BD Rhapsody's Sample Tag-based demultiplexing with genotype-based computational methods such as scSplit. The native BD pipeline demonstrated superior performance, attributable to its direct experimental barcoding approach, which outperformed scSplit's reliance on SNP inference from WTA data. Multiplets, likely stemming from technical errors during droplet encapsulation, further hindered data quality and highlight the need for improved experimental handling.

Our results demonstrate that while WTA can serve as a viable source for immune repertoire reconstruction, the 3′ sequencing approach is inherently limited for this purpose, with 5′ sequencing presenting a more suitable alternative for capturing full receptor diversity. Despite these challenges, the workflows developed here successfully integrate immune receptor data with transcriptomic profiles, offering a foundation for future improvements in single-cell immune profiling. This study underscores the critical need for optimizing both experimental protocols and computational tools to maximize the potential of single-cell immune analyses, particularly in non-standard platforms like BD Rhapsody.

Acknowledgements

Pursuing a full-time master's degree while working in research has been one of the most challenging journeys of my life. Balancing the demands of academia and a full-time research position required perseverance, sacrifice, and no small amount of support from those around me.

I would like to thank my supervisor Dr.Joan T. Matamalas for his guidance and for contributing to the completion of this Project. Thanks for offering me the opportunity of pursuing this collaboration.

To my family, thank you for always believing in me, even during moments when I struggled to believe in myself. Your constant encouragement and understanding have been a source of comfort and motivation throughout this intense period.

To my partner, thank you for your unwavering patience and love. Your support, whether through practical help or simply reminding me to step away from my desk for a moment, has been invaluable. You've made this difficult process far more bearable, and I am endlessly grateful.

To my flatmate Antón, thank you for bringing humour and levity to my days. Your knack for making me laugh, even during the most stressful moments, has been a much-needed reminder not to take things too seriously.

A heartfelt thanks to my previous and current lab mates and PIs, who showed empathy and understanding for the challenges of juggling a full-time job in research alongside a demanding master's program. Your flexibility and encouragement were crucial in helping me manage the complexities of this journey, and I am deeply appreciative of your support.

Finally, I extend my gratitude to everyone who stood by me during this period, whether through kind words, practical help, or simply showing patience as I worked through the demands of this dual commitment. Your support has meant more to me than I can say.

Dr. Joan T. Matamalas, certifies that the student Joel Moro Borrego has elaborated the work under his direction, and he authorizes the presentation of this master's thesis for its evaluation.

Advisor signature:

Table of contents

Motivation and Objectives

Understanding the immune system complexity requires advanced methodologies that can dissect immune responses at an unprecedented resolution. Immune receptor (VDJ) sequencing, when combined with single-cell transcriptomics, offers a transformative approach to study adaptive immunity by revealing the diversity and clonality of T-cell and B-cell receptors at the single-cell level. Despite its potential, significant challenges remain in applying VDJ sequencing, particularly in contexts where clinical samples are scarce, or data quality is compromised. This thesis aims to address these methodological challenges, focusing on improving immune receptor analysis and resolving technical artifacts in single-cell datasets.

One of the primary obstacles in single-cell multiomics is the presence of multiplets—instances where data from two or more cells are erroneously merged, leading to inaccuracies in assigning cells to their proper samples. Multiplets obscure the true biological signal, complicating the interpretation of immune responses. While some multiplets are related to the presence of two cells, some are related to the presence of two different barcodes for a single real cell, adding another layer of complexity to the identification of the cell. This project aims to develop and refine computational strategies to resolve these artifacts, thereby improving data quality and reliability.

Another major challenge is the incomplete coverage of T-cell receptor (TCR) and B-cell receptor (BCR) sequences. Traditional single-cell VDJ-seq datasets often contain a limited subset of cells with immune receptor information, restricting their utility in comprehensive analyses. By taking advantage of pre-existing transcriptomic data from 3’ single-cell RNA sequencing (scRNA-seq), it is possible to reconstruct TCR/BCR sequences for cells lacking explicit VDJ data. This methodological innovation could significantly expand the dataset scope, enabling a more detailed characterization of immune responses. Furthermore, using this approach also seeks to increase the volume of VDJ data available as it allows the usage of scRNA-seq data – which is spreadly popular -- in a retrospective manner.

Although this work leverages datasets from COVID-19 patients, the primary goal is to advance methodological approaches that have broad applicability in immunological research. The COVID-19 context provides a timely and clinically relevant case study, where understanding immune mechanisms has immediate implications for therapeutic interventions. The immune response to SARS-CoV-2 highlights the interplay between protective and pathogenic mechanisms, emphasizing the need for comprehensive immune receptor analysis to identify correlates of protection and drivers of severe disease.

Therefore, this thesis main objective relies on the development and refinement of methodologies to advance single-cell multiomics approaches, with a specific emphasis on improving immune receptor reconstruction and resolving technical challenges that hinder

data quality. To carry out this objective, several approaches arise such as improving receptor coverage by reconstructing TCR/BCR sequences from 3' scRNA-seq data, addressing the limitations of datasets with sparse VDJ information. To support this approach, resolving the multiplet problematic in single-cell datasets, ensuring the integrity of downstream analyses, is key. These methodological improvements converge toward a central objective: Applying the enhanced datasets to study immune responses with greater resolution and depth. As a proof of concept, the refined tools are applied to analyse immune cell subtypes, activation states, and clonal expansions in the context of COVID-19, shedding light on key mechanisms of adaptive immunity. While COVID-19 serves as the immediate focus, the innovations presented in this thesis hold broader implications for the study of adaptive immunity and immunopathology in diverse biological and clinical settings.

Chapter 1: Background and State-of-the-art

1. V(D)J Analysis

Certain TCR and BCR combinations are key in understanding immune response against the viral infection. The adaptive immune system relies on the ability of lymphocytes to recognize a wide range of foreign antigens. This diversity is achieved through the process of V(D)J recombination, which involves the rearrangement of gene segments encoding the variable (V), diversity (D), and joining (J) regions of TCR and BCR genes. During lymphocyte development, these gene segments are rearranged in a stochastic fashion, leading to the generation of unique antigen receptor sequences with distinct antigen-binding specificities. In that sense, V(D)J analysis is a powerful technique used to study the diversity and specificity of immune cell receptors, particularly TCR and BCR, at the genomic level. This technique provides insights into the processes of V(D)J recombination, which generate the vast array of antigen receptors required for adaptive immune responses.

1.1. Evolution of V(D)J Analysis

The evolution of V(D)J analysis has been a transformative journey in the field of immunology, revolutionizing our understanding of adaptive immunity and its role in health and disease. Initially proposed by Susumu Tonegawa in the 1970s, the concept of V(D)J recombination elucidated how the vast diversity of antigen receptors on B and T cells is generated through the rearrangement of variable (V), diversity (D), and joining (J) gene segments [64]. This groundbreaking discovery laid the foundation for studying immune responses at a molecular level and opened doors to exploring the intricacies of immune repertoire diversity. Traditional methods relied on Southern blotting and PCR amplification, enabling the identification of dominant V gene usage and limited clonal rearrangements [65]. These techniques, while groundbreaking, were limited by sensitivity and throughput.

The advent of high-throughput sequencing technologies revolutionized VDJ analysis in the late 2000s. Techniques like Illumina sequencing offered unprecedented scalability and sensitivity, allowing researchers to capture the vast diversity of the VDJ repertoire at a deeper level. This shift towards bulk VDJ analysis, where VDJ sequences from a population of B or T cells are analysed collectively, provided a comprehensive picture of the overall immune response.

Furthermore, alongside bulk VDJ sequencing, the progression of single-cell technologies has significantly enhanced our comprehension of the immune system. Single-cell VDJ sequencing allows for the concurrent analysis of antigen receptor sequences and transcriptomes at the cellular level, offering a more detailed perspective of the antibody and T cell repertoires. Through this method, valuable information regarding the clonal expansion, somatic

hypermutation, and functional attributes of SARS-CoV-2-specific lymphocytes has been unveiled, presenting novel opportunities for the development of precise immunotherapies and vaccines.

1.3 V(D)J Reconstruction from 3’ Sequencing

Single-cell sequencing technologies have revolutionized transcriptomics and related fields by enabling the analysis of the obtained sequences at the resolution of individual cells. Unlike bulk-sequencing which averages the biological information captured across a heterogeneous population of cells, single-cell approaches capture the cell profiles independently. This allows us to uncover cellular heterogeneity, identify rare cell populations, and study dynamic biological processes such as differentiation, immune responses, and cellular interactions (Figure 1). The ability to analyse thousands to millions of cells simultaneously has made single-cell sequencing a powerful tool in fields ranging from developmental biology to immunology.

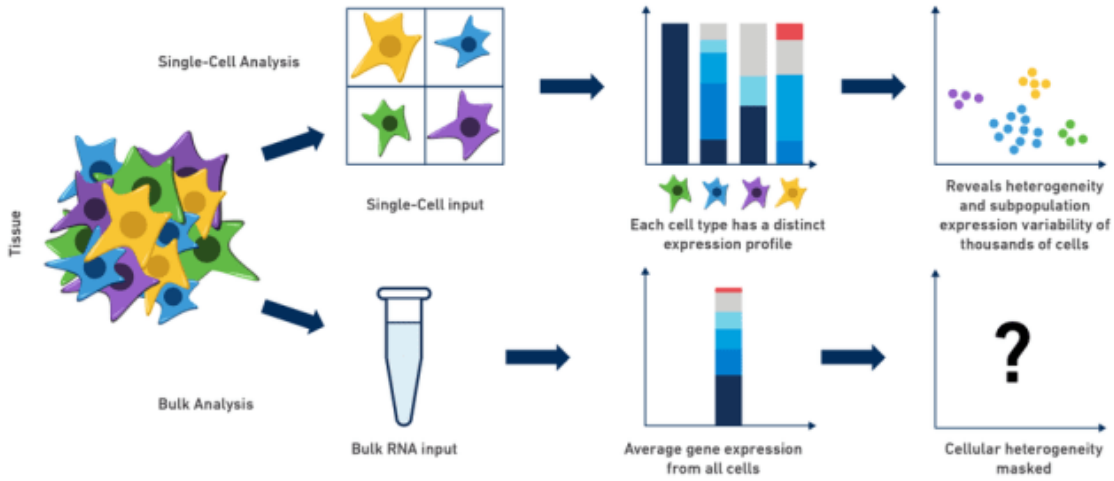

Figure 1. Extracted from 10x-genomics blog. *Single cell RNA-seq reveals cellular heterogeneity that is masked by bulk RNA-seq Methods.*

Single cell transcriptomic sequencing experiments are mostly 3’ based due to its cost-efficiency. Focusing on the 3' end of the transcript requires less sequencing depth compared to full-length sequencing, leading to lower costs and, also, the protocols for 3' based methods are generally simpler and less reagent-intensive than full-length methods.

3’ sequencing retrieves sufficient biological information for accurately quantifying gene expression levels, providing valuable insights since marker genes are typically located in the 3' UTR, therefore, allowing for effective cell type classification. Additionally, although limited compared to full-length sequencing, some isoform information can be obtained from 3' UTR variants. While 3' based scRNA-seq excels in these areas, it's important to note that full-length scRNA-seq is gaining traction for specific applications where comprehensive transcript information is essential, such as alternative splicing analysis, novel transcript discovery or our

main interest, V(D)J Analysis. 3' scRNA-seq primarily captures the 3' end of mRNA transcripts, which might not include the full V(D)J region, especially for longer TCR and BCR transcripts. This can lead to incomplete reconstruction and reduced accuracy. Also, due to the shorter read length, the detection of low-abundance TCR and BCR transcripts might be compromised, affecting the overall diversity captured in the repertoire. Additionally, shorter reads can increase the likelihood of sequencing errors, which can impact the accuracy of V(D)J reconstruction. For proper V(D)J reconstruction, then, 5' end sequencing is key. The 5' end of a transcript contains the crucial VDJ region for both TCR and BCR. This region is responsible for antigen recognition and is essential for understanding immune repertoire diversity and function. With the full VDJ sequence available, reconstruction algorithms can more accurately determine gene segments and junctions, and it is often compatible with other single-cell assays, such as CITE-seq, allowing for simultaneous profiling of gene expression and protein markers.

However, there are huge benefits hidden behind the limitations 3' end data. Viewed as a complement rather than the focus, trying to reconstruct V(D)J information from 3' data can be of crucial use. While traditional methods often rely on dedicated V(D)J sequencing, recent advancements have enabled the reconstruction of these sequences from 3' based scRNA-seq data, which primarily captures the 3' end of transcripts. Utilizing 3' scRNA-seq data allows for the analysis of a larger number of cells compared to dedicated V(D)J sequencing, providing a more comprehensive view of the immune repertoire. Most importantly, the amount of data available is massively large, giving place to interesting meta-analysis.

Several computational tools have been developed to facilitate the reconstruction of TCR and BCR sequences from 3' scRNA-seq data. For instance, TRUST4 is designed to reconstruct immune receptor repertoires from both bulk and single-cell RNA-seq data. TRUST4 is highly efficient and sensitive, capable of reconstructing full-length receptor sequences directly from FASTQ files or from BAM alignment files in either 3' or 5' scRNAseq sequencing experiments.

2. COVID

2.1. The Outbreak and symptoms

To understand the VDJ technique, it is also important to understand the biological background of the context under analysis. In the case of this project, the data is extracted from individuals that suffer from COVID infection.

The outbreak and subsequent global spread of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for the condition known as COVID-19, has prevailed in recent years [1]. COVID-19 was originally identified in late 2019 in Wuhan, China, promptly developing into a pandemic, as recognized by the World Health Organization (WHO) in March 2020[2]. Since then, the global community has witnessed an unprecedented human

cost, with millions of lives wasted and healthcare systems stretched beyond their limits. Aside from the acute health issues, the pandemic has had a significant impact on economies, social institutions, and the scientific community as a whole. The scientific community reacted to the pandemic with unparalleled level of effort and cooperation. In the face of an international crisis, the rapid sequencing of the SARS-CoV-2 genome, led to vaccine development and deployment, and advancements in treatment procedures constituted noteworthy scientific achievements [3].

COVID-19 can cause a wide range of symptoms, from asymptomatic or mild sickness to severe respiratory distress and multiorgan failure. In most of the cases it is described to cause fever, cough, shortness of breath, exhaustion, muscle or body aches, headache, sore throat, loss of taste or smell, nasal congestion, and gastrointestinal symptoms like nausea, vomiting, or diarrhea[4]. Most people with mild to moderate sickness recover without needing to be hospitalized, but some may have prolonged symptoms or develop complications. Severe cases may result in progressive respiratory failure, acute respiratory distress syndrome (ARDS), sepsis, septic shock, and multi-organ dysfunction, requiring immediate medical attention and mechanical ventilation. Certain risk factors, such as advanced age, underlying health disorders (e.g., cardiovascular disease, diabetes, chronic respiratory disease), and immunocompromised state, all raise the chance of serious illness and death [5].

2.2. SARS-CoV-2:

Severe acute respiratory syndrome. SARS-CoV-2 is an emerging coronavirus belonging to the Coronaviridae family. Coronaviruses are a large and diversified family of enclosed, positive-sense single-stranded RNA viruses. They infect a variety of hosts, including humans and other mammals, and cause diseases ranging from the common cold to severe respiratory infections [6]. SARS-CoV-2 has similarities to other well-known coronaviruses that have caused human epidemics, like SARS-CoV (which caused the 2003 severe acute respiratory syndrome outbreak) and MERS-CoV (which causes Middle East Respiratory Syndrome) [7].

Phylogenetic research suggests that SARS-CoV-2 belongs to the betacoronavirus genus, which is closely related to bat-derived coronaviruses, implying a zoonotic origin [8].

Viral Structure

The RNA genome of SARS-CoV-2 encodes multiple structural proteins, defined as the spike (S), envelope (E), membrane (M), and nucleocapsid (N), as well as non-structural proteins required for viral replication and transcription **(Fig 1**). The viral envelope emerges from the host cell membrane and is studded with spike glycoproteins, granting the virus a crown-like appearance under electron microscopy. The S protein is a trimeric glycoprotein made up of two functional subunits: S1 and S2. The S1 subunit contains the receptor-binding domain

(RBD), which interacts with the ACE2 receptor to allow viral attachment and entrance into host cells. The S2 component facilitates membrane fusion, which allows the viral genome to enter the host cell cytoplasm and begin replication. Understanding the structure and functions of these various components has been crucial for the development of vaccines and antiviral therapies aimed at disrupting different stages of the viral lifecycle, as well as identifying the gravity of the infections [9–12].

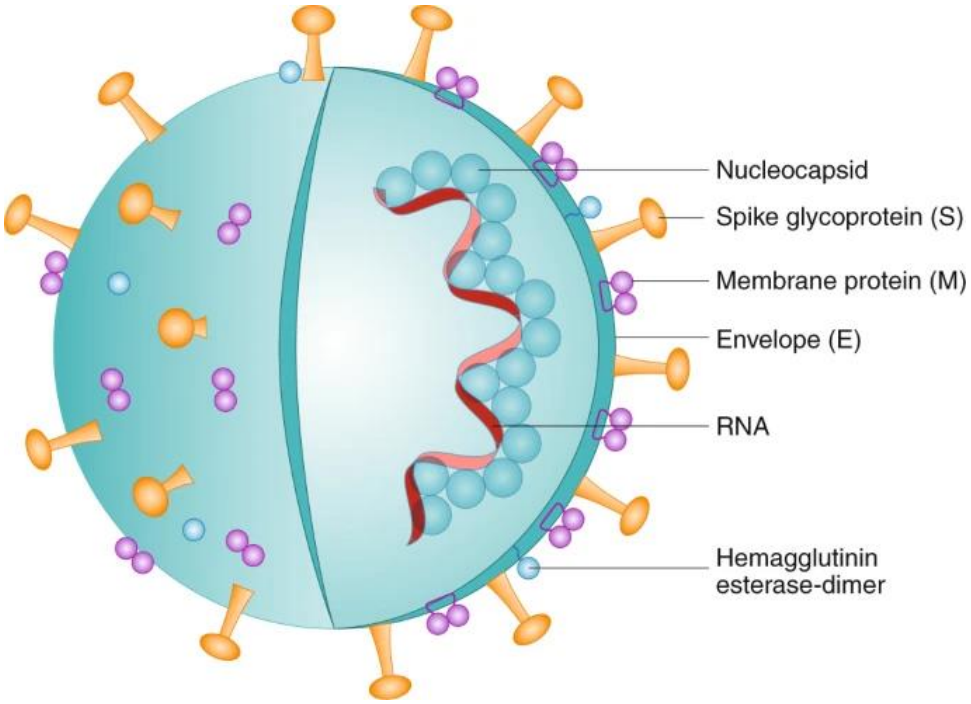

***Figure 2**: Schematic representation of SARS-CoV-2 structure. This is an enveloped, positive-sense RNA virus with four main structural proteins, including spike (S) and membrane (M) glycoproteins, as well as envelope (E) and nucleocapsid (N) proteins. Extracted from Florindo, H.F., Kleiner, R., Vaskovich-Koubi, D. et al. Immune-mediated approaches against COVID-19. Nat. Nanotechnol. **15**, 630–645 (2020). https://doi.org/10.1038/s41565-020-0732-3*

Mutations and Variants

Due to changes in its genome, SARS-CoV-2 exhibits genetic diversity, which may result in the formation of viral variants with unique behavioural traits. Natural mutations during viral replication can alter viral proteins, such as the spike protein, which is essential for viral entry and propagation. Therefore, the virus innate ability to adapt further complicates the continuing fight against the disease. Some changes can change the features of the virus, affecting its transmissibility, virulence (the severity of the sickness), and even immunological

escape, but most mutations have little effect. Variants, or unique lineages within the virus population with particular sets of mutations, may arise as a result of these major alterations[13].

Through genome sequencing techniques, which discloses the entire genetic composition of the virus, it has been possible to keep a careful eye on circulating variants. This makes it possible to find novel mutations and evaluate how they might affect general health. Many alterations in the SARS-CoV-2 genome have been found since the COVID-19 pandemic began, and some of these have caused the formation of viral variants of interest (VOIs) and of concern (VOCs)[14]. Public health efforts to prevent the spread of viruses are greatly hampered by VOCs, such as the Alpha (B.1.1.7) [15], Beta (B.1.351)[15], Gamma (P.1)[16], and Delta (B.1.617.2)[17] variations. These variants are linked to increased transmissibility, virulence, or resistance to neutralization by antibodies.

For instance, the S protein has undergone several changes, one of which is N501Y, which is linked to improved binding affinity to the ACE2 receptor and higher transmissibility [18]. Comparably, the S protein mutations found in the Beta variation, which was originally discovered in South Africa, as E484K, which may provide resistance to neutralizing antibodies produced by previous infection or immunization [19]. Furthermore, the greater transmissibility and propensity to elude immune responses of the Delta form, first discovered in India, have sparked worries. L452R and P681R, two changes in the S protein found in the Delta variation, may increase the infectiousness and transmission efficiency of the virus [20].

Therefore, mutations in SARS-CoV-2, particularly in areas encoding the spike protein, can affect the virus antigenicity and detection by the adaptive immune system. Changes in viral epitopes can alter the specificity of T and B cell responses, reducing the effectiveness of adaptive immunity against the virus. Furthermore, mutations that offer resistance to neutralizing antibodies or cytotoxic T cells can allow viral evasion of immune surveillance and hence immunological evasion. Thus, improving our understanding on how viral mutations interact with the innate and adaptive immune response is crucial for determining vaccination efficacy, as well as anticipating immune escape variations and directing vaccine design and development tactics.

2.3 V(D)J sequencing under infection

Under infection events, such the context of covid, V(D)J sequencing enables the comprehensive profiling of TCR and BCR repertoires, offering insights into the clonal diversity, specificity, and dynamics of immune responses in patients under infectious diseases. By analysing the TCR repertoire, we can identify antigen-specific T cell populations and track their expansion or contraction during infection. Similarly, BCR repertoire analysis allows for the characterization of virus-specific antibody responses and the assessment of their neutralizing capacity.

One of the most striking insights emerging from VDJ sequencing studies is the remarkable heterogeneity of the immune response to the virus, when the common sense would lead to expect a monoclonal response. This concept goes beyond simply acknowledging individual differences. It highlights the intricate variations that can occur within a single patient, across different immune cell populations, and even within the same cell type. The study by Wang et al. [66] exemplifies this heterogeneity. By employing single-cell sequencing, they analysed the immune cell profiles of COVID-19 patients in the recovery stage. Their findings revealed a surprisingly diverse landscape, with distinct populations of T cells, B cells, natural killer (NK) cells, and monocytes co-existing within the patients. This diversity extended to the functional characteristics of these cells, with some exhibiting strong activation markers and others displaying an exhausted or immunosuppressive phenotype.

In the inflammatory context, longitudinal analysis of the VDJ repertoire has shown that certain antibody lineages can persist and recognize conserved epitopes on the SARS-CoV-2 spike protein, even as the virus continues to evolve. This raises the possibility that these antibodies may also play a role in the long-term cardiovascular sequelae observed in some COVID-19 survivors.

Furthermore, single-cell VDJ sequencing has revealed COVID-19-specific clonal expansions of B cells and T cells, some of which are associated with the production of autoantibodies that can target host tissues, including the heart. This suggests that the adaptive immune response to SARS-CoV-2 infection may trigger cross-reactive antibodies or T cells that contribute to the development of cardiovascular complications in severe COVID-19 cases [67].

Finally, besides the TCR and BCR context, A recent study combined single-cell RNA sequencing (scRNA-seq) with TCR and immunoglobulin heavy chain (VDJ) sequencing to analyse peripheral blood mononuclear cells (PBMCs) from COVID-19 patients and healthy controls. The researchers found that the decreased fraction of non-classical monocytes (ncMono) in severe COVID-19 cases was associated with specific genetic risk factors identified through genome-wide association studies (GWAS). This suggests that dysregulation of the innate immune response, particularly in ncMono, may contribute to the development of severe COVID-19.

3. Immune System under Covid context

The immunological response to SARS-CoV-2, involves a complex interplay of the innate and adaptive immune systems. When the host immune system encounters the virus, it launches a series of responses targeted at confining and removing the infection. However, dysregulated immune responses can cause severe inflammation and tissue damage, adding to the pathophysiology of COVID-19 **(Fig 2**).

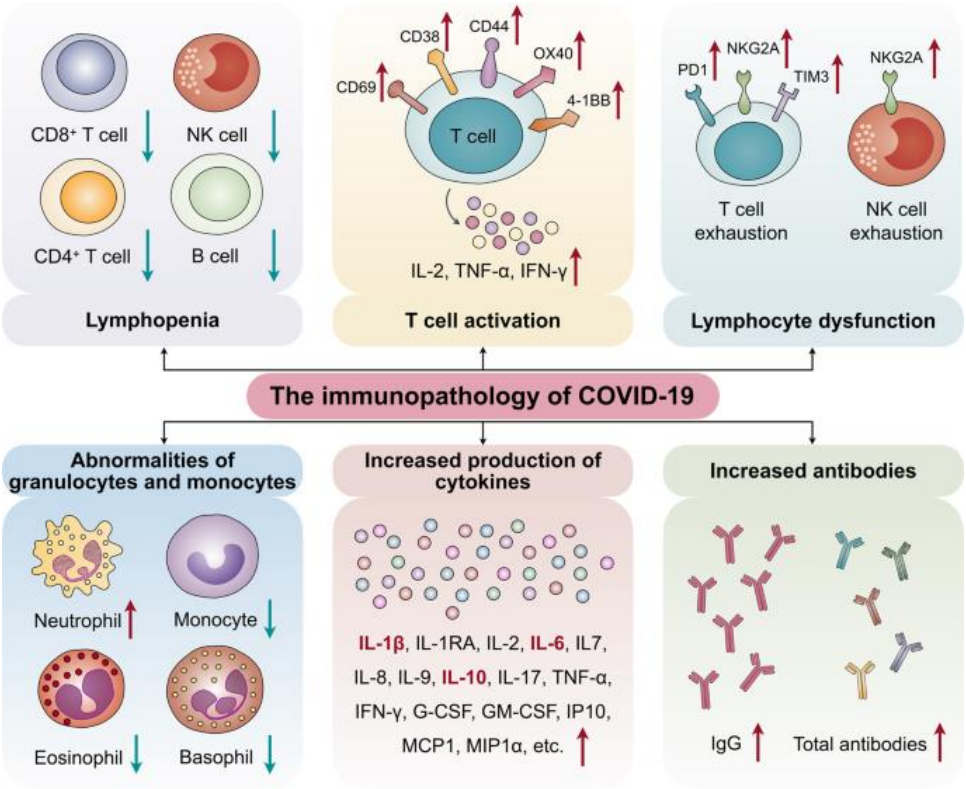

***Figure 3**: The immunopathology of COVID-19. The immune patterns of COVID-19 include lymphopenia, lymphocyte activation and dysfunction, abnormalities of granulocytes and monocytes, increased production of cytokines, and increased antibodies. Lapuente, D., Winkler, T.H. & Tenbusch, M. B-cell and antibody responses to SARS-CoV-2: infection, vaccination, and hybrid immunity. Cell Mol Immunol **21**, 144–158 (2024). https://doi.org/10.1038/s41423-023-01095-w*

3.1 Dysregulated inflammation

COVID-19, has a wide range of clinical manifestations, from mild respiratory symptoms to severe pneumonia, acute respiratory distress syndrome (ARDS), and multi-organ failure. The dysregulated immune response, characterized by an excessive inflammatory cascade, often referred to as a "cytokine storm", is central to the pathophysiology of the disease. This inflammatory cascade plays an important role in the progression of the disease and contributes to the development of severe complications. Acute lung injury, thrombosis, and systemic organ dysfunction are examples of this mentioned complications.[21]

When infected with SARS-CoV-2, viral particles engage with host cells, particularly epithelial cells lining the respiratory system, resulting in viral replication and the release of viral antigen. As pattern recognition receptors recognize viral components, such as Toll-like receptors and retinoic acid-inducible gene I-like receptors, innate immune cells, characterized as macrophages and dendritic cells, become active. Immune cells release pro-inflammatory cytokines, especially IL-6, TNF-α, and IL-1β, which trigger the inflammatory cascade. Activated immune cells release pro-inflammatory cytokines and chemokines, which recruit extra immune cells to the infection site and enhance the inflammatory response. Neutrophils,

monocytes, and T lymphocytes are drawn to the lungs and other afflicted tissues, where they cause tissue damage and inflammation by releasing cytotoxic chemicals, reactive oxygen species (ROS), and proteases. The activation of the complement system intensifies the inflammatory cascade, resulting in the production of anaphylatoxins and the recruitment of immune cells to the infection site [22].

In extreme cases of COVID-19, dysregulated immune activation and cytokine release result in the previously stated "cytokine storm", which causes systemic inflammation and widespread tissue destruction. Excessive pro-inflammatory cytokine production disturbs normal immunological homeostasis, contributing to the development of ARDS, sepsis, and multiorgan failure syndrome. The imbalance of pro-inflammatory and anti-inflammatory signals causes immunological dysregulation and immunopathology, worsening tissue damage and organ malfunction [21].

3.2 Cardiopathology related to COVID-19

Studies have highlighted the interesting relationship between the systemic inflammation induced by SARS-CoV-2 infection and the subsequent effects on the cardiovascular system.

According to research, COVID-19 can cause myocardial injury through both direct and indirect routes. Direct damage occurs when the virus infects cardiomyocytes, causing myocarditis and subsequent inflammation in the heart muscle. Tachycardia and hypotension are examples of indirect processes, as are arrhythmias and myocardial infarction. Additionally, vascular thrombogenic alterations and stress-induced cardiomyopathy can worsen heart damage in COVID-19 patients.[23]

The systemic inflammatory response generated by SARS-CoV-2 infection can result in endothelial damage and microvascular dysfunction, affecting heart function. Cytokine-mediated inflammation can cause structural damage to the heart, with rare cases developing to fulminant myocarditis even in the absence of respiratory involvement. Furthermore, coronary artery inflammation can cause myocardial ischemia, further jeopardizing cardiovascular function. The mismatch between oxygen supply and demand due to respiratory failure might eventually lead to cardiopulmonary failure in extreme situations[24]. Furthermore, patients have a higher chance of acquiring thrombotic problems such as deep vein thrombosis, ischemic stroke, and myocardial infarction. The virus-induced hyper-coagulopathy impairs microvascular perfusion and increases the risk of thromboembolic events. Arrhythmias are also common in COVID-19 individuals, with various dysrhythmias documented, including tachyarrhythmias, atrial fibrillation, and ventricular arrhythmias. These arrhythmias can be caused by variables such as cardiac damage, hypoxia, systemic inflammation, and particular drugs used in COVID-19 treatment [25].

3.3 Innate Immune Response

Upon encountering SARS-CoV-2, the innate immune system acts as the initial line of defence, orchestrating a rapid and nonspecific response to the invading virus. This early response is critical for detecting the presence of viral pathogens and initiating the activation of the adaptive immune system, while also directly limiting viral replication and spread within the host. This task is carried out by key innate immune cells strategically positioned throughout the body; macrophages, dendritic cells (DCs), and natural killer (NK) cells [26].

Macrophages are key players in the innate immune response to SARS-CoV-2. These versatile immune cells act as proactive defenders against invading pathogens, fulfilling crucial functions in viral detection, elimination, and immune regulation. In the context of covid, however, their multifaceted role, is not easily categorized as they exhibit both protective and detrimental functions, making them a double-edged sword in the fight against this viral infection [27]. Specifically, macrophages are equipped with an array of pattern recognition receptors that enable them to actively detect the presence of viral pathogens. Upon recognition, macrophages swiftly initiate a series of effector functions aimed at eliminating the virus. Central to their defence strategy is phagocytosis, whereby macrophages engulf and digest viral particles, effectively neutralizing the threat and preventing viral spread. Beyond the phagocytosis capabilities, macrophages are also active makers of proinflammatory cytokines, chemokines, and reactive oxygen species (ROS), as well as proinflammatory cytokines such as interleukin-1β [IL-1β], tumour necrosis factor-alpha [TNF-α], and others.

These chemicals are essential for attracting and stimulating additional immune cells to the infection site, boosting the immune system's response, and encouraging the removal of viruses. Furthermore, when viruses have been eliminated from the body, macrophages aid in tissue regeneration and immunological control. They have the ability to transition from proinflammatory M1 to anti-inflammatory M2 phenotypes, which contributes to the resolution of inflammation and the process of tissue healing [28–30].

In the context of covid, while macrophages offer a valuable line of defence during COVID-19 infection, their potent pro-inflammatory arsenal can become a double-edged sword. In severe cases, an excessive and dysregulated inflammatory response orchestrated by macrophages can lead to a life-threatening condition deriving in tissue damage[31]. This uncontrolled release of pro-inflammatory cytokines and chemokines wreaks havoc on lung tissue, causing significant damage and hindering gas exchange, ultimately leading to respiratory failure. Adding another layer of complexity to the narrative is the recent discovery that macrophages themselves can become infected by SARS-CoV-2 [28]. This ability to act as viral reservoirs poses a significant challenge. Infected macrophages can harbour the virus, hindering its complete eradication by the immune system and potentially contributing to chronic or relapsing infections [30,32,33].

Natural killer (NK) cells are innate lymphoid cells equipped with cytotoxic mechanisms that allow them to directly recognize and eliminate virus-infected cells. NK cells possess a unique

ability to recognize and eliminate virus-infected cells without prior antigen-specific activation, a hallmark of the innate immune response. They achieve this feat through an intricate interplay of activating inhibitory signals received from various cell surface receptors. When the balance of these signals' favours activation, NK cells unleash their cytotoxic arsenal, deploying perforin and granzyme molecules to induce apoptosis (programmed cell death) in infected cells, effectively limiting viral replication and spread.

The influence of NK cells extends beyond direct killing. They actively participate in shaping the adaptive immune response by secreting a diverse array of cytokines and chemokines. These signalling molecules function as crucial messengers, recruiting other immune cells like T lymphocytes and antigen-presenting cells to the infection site, fostering a coordinated immune assault against the virus. Additionally, NK cells can modulate the activity of dendritic cells, further influencing the development and direction of the adaptive immune response [34,35].

Dendritic cells (DCs) are specialized antigen-presenting cells (APCs) capable of capturing and processing viral antigens for presentation to T cells. DCs are also equipped with an array of PRRs that enable them to detect the presence of viral pathogens, including SARS-CoV-2. Upon encountering SARS-CoV-2, immature DCs residing in peripheral tissues, such as the respiratory tract, skin, and mucosal surfaces, undergo a process of maturation and migration to secondary lymphoid organs, such as lymph nodes. During migration, DCs upregulate the expression of co-stimulatory molecules and major histocompatibility complex (MHC) molecules, which are essential for T cell activation. Finally, in the lymph nodes, matured DCs present viral antigens to naive T cells, initiating the activation and differentiation of adaptive immune responses. DCs accomplish this through the formation of immunological synapses, specialized structures that facilitate the interaction between DCs and T cells. This crucial interaction hinges on the ability of DCs to process and present fragmented viral antigens (peptides) on their surface in conjunction with MHC molecules. The specific type of MHC molecule engaged, along with the constellation of co-stimulatory and inhibitory signals delivered by DCs, dictates the nature and magnitude of the T cell response. DCs can fine-tune the immune response by promoting the generation of different T cell subsets, such as effector T helper (Th) cells and cytotoxic T lymphocyte, crucial for eliminating virus-infected cells. Additionally, regulatory T cells (Tregs) can be induced by DCs, playing a vital role in preventing excessive immune activation and potential immunopathology [36–38].

However, the influence of DCs extends beyond T cell activation. They also interact with other immune cells, such as natural killer (NK) cells and B cells, shaping their function and contributing to the overall immune response. By secreting a diverse array of cytokines and chemokines, DCs serve as crucial conductors of the immune symphony, fine-tuning the communication and collaboration between various immune cells to effectively combat the viral threat. For example, DCs produce type I interferons (IFNs), such as IFN-α and IFN-β, which have potent antiviral properties and serve as key mediators of innate antiviral immunity [39,40].

3.4 Adaptative Immune Response

The adaptive immune response, building upon the foundation laid by the innate immune system, represents a sophisticated defence mechanism orchestrated by a diverse variety of immune cells. While the innate immune system provides immediate, nonspecific defence against pathogens, the adaptive immune response adds an additional layer of specificity and memory to the immune repertoire. In the context of SARS-CoV-2 infection, the adaptive immune response plays a critical role in recognizing and eliminating the virus, thereby conferring long-lasting immunity and protection against reinfection [41].

The adaptive immune response, in this context, is initiated upon the recognition of viral antigens by lymphocytes, like B cells and T cells. B cells, equipped with antigen-specific receptors known as B cell receptors (BCRs), recognize and bind to viral antigens leading to their activation and differentiation into plasma cells. These plasma cells produce vast quantities of antibodies, which specifically target and neutralize the virus, preventing its entry into host cells and facilitating its clearance by other immune cells. Similarly, T cells, divided in CD4+ helper T cells and CD8+ cytotoxic T cells, play critical roles in the adaptive immune response to SARS-CoV-2. Helper T cells recognize viral antigens presented by antigen-presenting cells, such as dendritic cells and macrophages, through their T cell receptors (TCRs). Upon activation, helper T cells orchestrate the immune response by releasing cytokines that stimulate B cells to produce antibodies and activate cytotoxic T cells. Cytotoxic T cells, in turn, directly target and kill virus-infected cells, preventing viral replication and spread within the host [42].

B Cells:

B cells are central players in the adaptive immune response, primarily responsible for producing antibodies against viral antigens. SARS-CoV-2 spike proteins, particularly the receptor-binding domain (RBD), serve as key targets for B cell recognition. Upon binding to viral antigens, B cells undergo activation, leading to their proliferation and differentiation into two primary effector cell types: plasma cells and memory B cells[43]. These plasma cells churn out copious amounts of antibodies, specifically targeting the viral antigens encountered during the infection. Antibodies produced by plasma cells bind to the spike proteins of SARS-CoV-2, interfering with viral entry into host cells and marking the virus for destruction by other immune cells, such as macrophages and neutrophils. In addition to generating a rapid antibody response during acute infection, B cells also give rise to memory B cells, which persist long-term in the body. Memory B cells remain poised to mount a swift and robust antibody response upon re-exposure to SARS-CoV-2. This immunological memory confers long-lasting immunity and protection against reinfection, providing a crucial defence mechanism against the ongoing threat of COVID-19[44,45].

It is important to remark, that the B cell response is not monolithic. B cells can generate a variety of antibodies, each with slightly different binding specificities. This phenomenon, known as antibody repertoire diversity, offers a crucial advantage in the fight against viral mutations. If a particular viral mutation arises, rendering the initial set of antibodies less

effective, the diverse repertoire of B cells ensures that there is a high probability of finding antibodies that can still bind to the mutated virus and provide continued protection.

On another layer, about specific limitations and constrains due to covid exposure, in severe cases of COVID-19, hospitalized patients have been observed to exhibit higher levels of RBD-specific IgG antibodies compared to non-hospitalized convalescents. This correlation between disease severity and antibody levels suggests a stronger immune response in more severely ill patients, potentially due to higher viral loads and prolonged viral replication. Furthermore, B cell responses in critically ill patients show pronounced extrafollicular (EF) B-cell responses that produce relatively high levels of neutralizing antibodies (nAbs), indicating a robust immune reaction in these individuals[46].

CD4+ Helper T Cells

The adaptive immune response to SARS-CoV-2 is initiated when CD4+ helper T cells recognize viral antigens presented on the surface of antigen-presenting cells, such as dendritic cells and macrophages. This recognition occurs through the interaction between the TCR on CD4+ T cells and the antigen-MHC class II complex on antigen-presenting cells, triggering the activation and differentiation of CD4+ T cells into distinct effector subsets [47].

One subset of activated CD4+ helper T cells, known as T follicular helper (Tfh) cells, migrates to germinal centres within secondary lymphoid organs. Within germinal centres, Tfh cells provide critical signals to B cells, for instance, co-stimulatory molecules and cytokines such as interleukin-21 (IL-21), which promote B cell proliferation, differentiation, and antibody production. This facilitates the maturation of B cells into antibody-secreting plasma cells and memory B cells [48].

Th1 cells, a different subgroup of activated CD4+ helper T cells, are essential for coordinating cellular response against SARS-CoV-2. Th1 cells generate cytokines, like interferon-gamma (IFN-γ), which promote macrophage and cytotoxic T cell activation and activity. Th1 cells aid in the removal of virus-infected cells and the resolution of the infection by boosting the phagocytic and cytotoxic capabilities of these effector cells [49,50]. Furthermore, Th1 cytokines contribute to the development of long-term immunological memory against SARS-CoV-2 by encouraging the generation of memory T cells. Additionally, CD4+ helper T cells are essential for controlling the immune system reaction the virus, which helps to avoid excessive inflammation and tissue damage. By preventing effector T cells from activating and functioning, regulatory T cells (Tregs), a subset of CD4+ T cells, dampen immunological responses. Tregs are identified by the expression of the transcription factor FOXP3 and aid in immunological homeostasis, suppress excessive inflammation and in the clearance of infections by preventing immunopathology [51].

CD8+ Cytotoxic T Cells

Activated CD8+ T cells possess potent cytotoxic capabilities, enabling them to directly target and eliminate virus-infected cells. Upon encountering infected cells, CD8+ T cells release

cytotoxic granules containing perforin and granzymes, which induce apoptosis, or programmed cell death, in the target cells.

Perforin forms pores in the cell membrane of infected cells, allowing granzymes to enter and trigger apoptotic pathways, effectively killing the virus-infected cells and limiting viral replication and spread within the host[52].

In addition to their role in acute viral clearance, CD8+ cytotoxic T cells contribute to the establishment of long-term immune memory. Following the resolution of the acute infection, a subset of activated CD8+ T cells differentiate into memory T cells, which persist long-term in the body. Memory CD8+ T cells exhibit enhanced effector functions and rapid recall responses upon re-exposure to the virus, providing a critical defence mechanism against reinfection and contributing to the long-lasting immunity conferred by vaccination or natural infection [53,54].

HLA and TCR variability

HLA polymorphisms influence the diversity of T cell responses observed in COVID-19 patients. Certain HLA alleles have been associated with enhanced or diminished CD8+ T cell recognition of SARS-CoV-2 antigens, affecting the effectiveness of anti-viral immunity [55]. HLA genes strongly influence vaccine-induced antibody responses and the risk of COVID-19 breakthrough infections. Studies have shown that HLA alleles, particularly HLA-DQB1*06, are associated with variations in antibody levels post-vaccination, suggesting a link between HLA types and vaccine responsiveness [56]. Additionally, HLA class I and class II alleles have been implicated in differential immunological responses to viral infections, including SARS-CoV-2. Variability in HLA genes can impact the specificity and magnitude of CD8+ T cell responses, affecting disease outcomes and vaccine efficacy [55,57,58].

On another layer, it has already been mentioned that the specificity of T cell responses to viral antigens presented by HLA molecules is governed by the interaction between the TCR and the peptide-MHC complex **(Fig 3)**. The TCR recognizes specific antigenic peptides derived from viral proteins that are presented by HLA molecules on the surface of infected cells or antigen-presenting cells. This recognition occurs through the complementary pairing of the TCR with the peptide-MHC complex, akin to a lock and key mechanism, ensuring precise targeting of virus-infected cells [59]. The diversity of the TCR repertoire enables T cells to recognize a wide range of viral antigens presented by diverse HLA molecules. TCR genes undergo extensive somatic recombination and mutation, generating a vast array of TCR variants with unique antigen-binding specificities [60]. This diversity allows T cells to recognize and respond to a wide variety of viral peptides presented by different HLA alleles, ensuring broad coverage and effectiveness of the adaptive immune response to SARS-CoV-2.

In the context of covid, research has identified specific clonal expansions of CD4+ and CD8+ T cells following infection [61]. These clonotypes exhibit preferential recombination usage of V(D)J gene segments, indicating a targeted immune response against SARS-CoV-2 antigens. Notably, certain TCR combinations, such as TRBV6-5-TRBD2-TRBJ2-7, are shared between CD4+ and

CD8+ T cells in different COVID-19 patients, suggesting a coordinated immune response across T cell subsets [62,63].

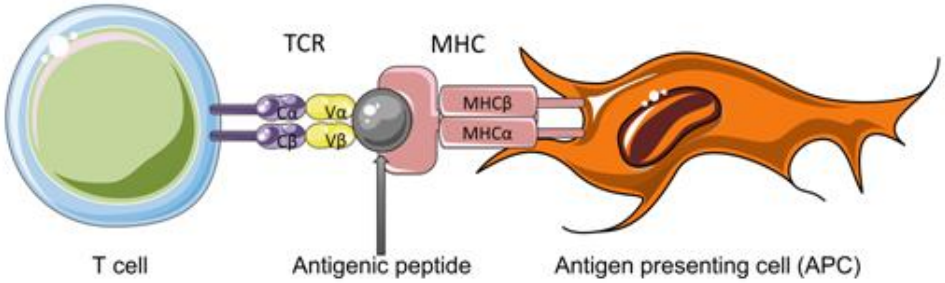

Figure 4: *Trimolecular complex, comprised of a T cell bearing a T cell receptor (TCR), an antigenic peptide, and an antigen presenting cell (APC) expressing a major histocompatibility complex (MHC) molecule. The TCR consists of two chains, α and β, each bearing a constant (Cα, Cβ) region and a variable (Vα, Vβ) region. The variable regions are in direct contact with the antigenic peptide and surface of the MHC molecule. JoLS, J Life Sci, Vol. 2, No. 4, December 2020:38-58, https://doi.org/fmcz PMID: 33364626 PMCID: PMC7757640*

Chapter 2: Human Dataset Exploration

1. Introduction

1.1. Patients and Experiment

The dataset used for this project was composed by samples from human origin. Specifically, the dataset comprises samples collected from 20 COVID-19 patients at the Brigham and Women's hospital in Boston, evenly split between 10 patients classified as Thrombocytopenic (positive) and 10 as Thrombocytopenic (negative). These samples were gathered during the period between April 2020 and September 2020, with data available from up to two time points per patient.

Peripheral blood mononuclear cells (PBMCs) were analysed through various advanced techniques. PBMCs are preferred in sequencing and immunology studies, such as ours, because they provide a focused view of the immune system's cellular components, which is our focus. Using whole blood would add another layer of complexity and noise to our study since it contains a high proportion of components like red blood cells and neutrophils, Also, plasma, while rich in proteins and cytokines, lacks cellular information and cannot capture the functional dynamics of immune cells.

Additionally, PBMCs are also more stable during processing, compatible with cryopreservation for long-term studies, and ideal for single cell sequencing and immune repertoire profiling.

Taking this into account, Whole transcriptome analysis (WTA) was conducted using single-cell RNA sequencing (scRNAseq), alongside a targeted scRNAseq approach focusing on 499 genes and an additional set of 68 custom genes. T-cell and B-cell receptor repertoires were profiled using VDJ sequencing, while surface protein data were obtained using antibody sequencing (AbSeq). Also, plasma samples were subjected to targeted aptamer-based proteomic analysis, covering approximately 7,000 proteins. However, this latter experiment was not included in this analysis.

About patient demographics, as mentioned, the dataset includes a total of 20 individuals which are divided into those with and without cardiovascular outcomes (Fig 5). Among the patients without cardiovascular outcomes, there are 9 individuals, while 11 patients fall into the category of those with cardiovascular outcomes. Interestingly, the sex distribution shows that males dominate the cardiovascular outcome group, with 7 males compared to 4 females, while the reverse is true for the group without cardiovascular outcomes, where there are more females (5) than males (4).

We must mention as well that the age profile differs notably between the two groups. Patients without cardiovascular outcomes tend to be older, with an average age of 62.9 years and a standard deviation of 3.8, whereas those with cardiovascular outcomes are younger, with an average age of 54.5 years and a slightly higher variation of 5.4 years. This suggests a possible relationship between age and the presence or absence of cardiovascular complications in the cohort.

Regarding racial demographics, both groups show a predominance of white patients, with 6 individuals in each group. Black patients represent a smaller proportion in both groups, with 2 in each category. Notably, Hispanic and Asian patients are only present in the cardiovascular outcome group, with one individual each, while the "Other" racial category includes one patient in both groups. This distribution indicates a majority representation of white patients, with other racial groups appearing in much smaller proportions.

1.2. Experiment Design

The experiment was designed to split the samples into four different batches, each batch representing a patient timepoint with 2 control reference conditions (Fig 5). This control conditions attain to unstimulated and IFNγ-stimulated controls, two in each batch as mentioned. Specifically, batch 1 was represented by 9 COVID-19 individuals, while batch 2, 3 and 4 include a total number of 10 patients with timepoint conditions besides the two mentioned controls.

The idea of including IFNγ-stimulated controls relies based on COVID-19 action in the immune system. IFNγ is a critical cytokine involved in orchestrating immune responses. It plays a central role in activating macrophages, enhancing antigen presentation, and driving the adaptive immune response by promoting T cell differentiation and B cell class switching. Thereby, since COVID-19 causes immune dysregulation and cytokine storms, as mentioned in previous sections, having that kind of control contributes to understand the functional state of immune cells under response to evens of infection and inflammation.

Hence, from each of this patient peripheral blood is extracted and collected flow cytometry (FACS) is used to assess four panels, covering 24 markers across T cells, B cells, and monocytes. Barcoded sample multiplexing kits (SMKs) are utilized to tag cells, allowing the simultaneous profiling of various parameters. In each batch, cells are divided into two experimental sets, labelled as Set A and Set B, with cells processed for whole transcriptome analysis and surface protein profiling. The Set A samples focus on combined transcriptomic and AbSeq surface marker analysis, while Set B includes targeted sequencing of immune pathways and receptor profiling. With this design we ensure a balanced and reproducible cell distribution per condition, enabling multi-omics analysis with a robust focus on immune responses.

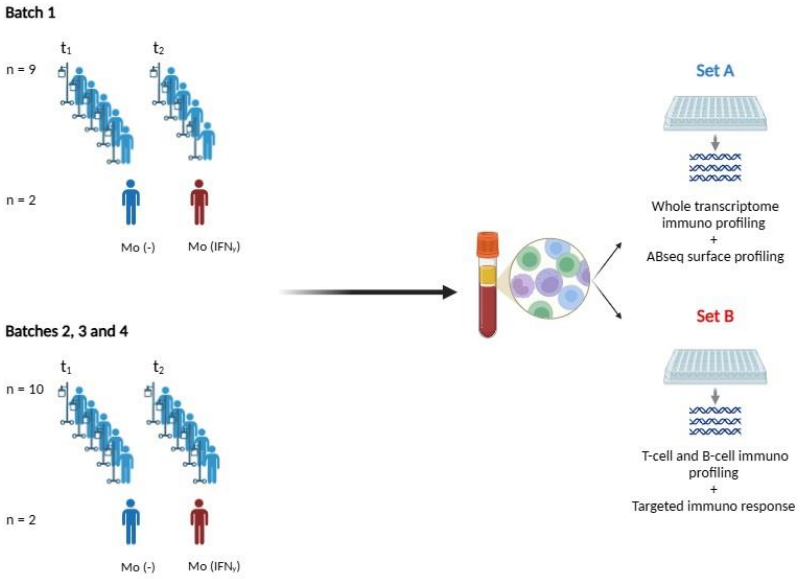

Figure 5: *Organization of the sequencing plan, including the number of batches, patient per batch, information on the two controls per batch and the different sets for which we have sequenced.*

1.3. Ethical Considerations

This study adheres to the highest ethical standards to ensure the protection of participant rights, privacy, and safety. The data was obtained through collaboration with Brigham and Women's Hospital (BWH) at Harvard Medical School, with all protocols approved by the appropriate institutional review boards (IRBs) in compliance with the Declaration of Helsinki. Prior to sample collection, informed consent was obtained from all participants or their legal representatives, ensuring they were fully informed about the purpose, methods, and potential implications of the research.

As a European researcher working with American data, special care has been taken to comply with both European data protection regulations, such as the General Data Protection Regulation (GDPR), and American ethical standards. All patient data is fully anonymized and de-identified at the source, ensuring that no personally identifiable information is accessible. Data analysis is conducted remotely through secure servers hosted in the United States, which provide access to sequences and analytical tools but strictly prohibit downloading or external storage of sensitive information. This ensures compliance with international privacy laws and maintains the confidentiality of the participants.

Moreover, the study design further reflects a commitment to ethical inclusivity and equity by incorporating participants from diverse racial and demographic backgrounds. Additionally, strict biosafety protocols were observed throughout the handling and analysis of biological samples, particularly in the context of COVID-19. These measures underscore a broader

commitment to responsible research practices, prioritizing participant safety, respect, and data integrity at every step.

2. Whole transcriptome analysis

2.1 Upstream analysis

BD Rhapsody provides their customers with a software for upstream analysis. Therefore, this process begins with the raw sequencing data, which is uploaded to the secure Seven Bridges environment, ensuring compliance with data privacy and security standards.

The pipeline initiates with FastQC [68], which evaluates the quality of raw sequencing reads, providing detailed metrics on base quality scores, GC content, and the presence of overrepresented sequences. Based on this quality assessment, trimomatic [69] is performed to remove adapter sequences and low-quality bases, ensuring cleaner data for downstream steps. Once the reads are trimmed, they are aligned to the reference genome using STAR [70], a highly efficient algorithm for spliced read alignment. STAR ensures accurate mapping of reads to both exonic and intronic regions, which is essential for comprehensive transcriptome analysis. After alignment, UMI-tools [71] is used to deduplicate reads by collapsing those with identical unique molecular identifiers (UMIs). This step prevents overestimation of gene expression levels caused by PCR amplification.

The resulting gene expression counts are quantified using BD Rhapsody's workflow, which generates count matrices directly compatible with downstream analyses. Seven Bridges also provides the output already as a Seurat [72] object in *.Rds* format for usage in the downstream analysis.

All software run in by Seven Bridges was set up under standard default configuration.

2.2. Frequencies by Sample

For the analysis I started by creating the Seurat objects from the count matrices obtained in the upstream analysis. Given that setA (WTA) contains ABseq information the cells were already annotated into the different immune cell types, and it was possible to extract the frequencies of each patient in each of the different batches. For the sake of making the analysis less complicated in the first stages, even if the patients are the same individual at different timepoints in the different batches, it was assumed that each is an independent individual. This would help to perform the quality control and extract the cell type frequencies.

Frequencies of the number of cells per patient were extracted (Table 1). To extract these frequencies, we made use of the metadata data frame stored in the S4 Seurat object. Using tools from Tidyverse [73] package data wrangling was performed to extract the final tables (see GitHub *explorationA.Rmd*).

At first glance we can start to notice striking factors that make our data look quite complicated to analyse. For example, the number of Multiplets is concerningly high. Multiplets occur when more than one cell is captured in a single droplet or capture event during the scRNA-seq process, leading to mixed or combined gene expression profiles that do not represent a single, distinct cell type. These false-positive data points can distort the interpretation of the results, leading to misleading conclusions about gene expression patterns, cell type identification, and cellular heterogeneity. There are also other kind of multiplets that arise from mistakes in the demultiplex process or the barcode itself, were a cell is given two different barcodes. In that case, the information of the cell is totally valid, but it is not possible to trace back which was the sample of origin.

In our case, the numbers of multiplets observed across the batches vary significantly, with some batches displaying particularly high counts. For example, Batch 4 shows a notably high number of multiplets (5242), which is concerning because it suggests a possible issue with the technical process during cell capture or processing. High multiplet rates could indicate insufficient quality control during sample preparation, such as suboptimal droplet separation or excessive cell concentration, which can result in the co-capture of multiple cells.

Moreover, the impact of multiplets extends beyond the mere increase in noise within the data. If not properly addressed, these erroneous data points could confound downstream analyses, such as clustering, differential expression analysis, or the identification of cell type markers. High multiplet rates could significantly alter the observed cell type frequencies, which in turn could bias conclusions regarding the composition of the tissue or biological system being studied. Additionally, the presence of multiplets might lead to inaccurate cell-type assignments and hamper efforts to delineate subpopulations, which are critical in many biological and disease-related investigations.

Besides the multiplet affair, another number that stands out is the relatively low number of cells in certain sample tags, such as SampleTag05_hs across several batches. In Batch 1, for instance, SampleTag05_hs has only 51 cells, which is notably smaller compared to other sample tags within the same batch. This low cell count might indicate several potential issues, including a problem with sample collection, cell viability, or processing efficiency. A low number of cells in a sample could lead to a reduced representation of certain cell types or biological conditions in the dataset, potentially skewing downstream analyses, particularly those related to the identification of rare cell populations. In our case, we might miss out on important heart-disease-related patients.

Additionally, SampleTag04_hs and SampleTag10_hs in Batch 2 also show relatively low counts (40 and 51, respectively), which might suggest that these samples were underrepresented or perhaps did not fare as well during the sample preparation phase. Low cell counts in certain samples could be of concern in cases where the biological diversity or functional state of those samples needs to be explored in depth. Such samples may fail to provide sufficient statistical power for meaningful comparison, making it harder to detect subtle differences or trends between experimental conditions.

	Batch 1	Batch 2	Batch 3	Batch 4
Multiplet	1688	2916	2091	5242
SampleTag01_hs	13	1772	1504	1482
SampleTag02_hs	201	2574	1179	1038
SampleTag03_hs	780	1622	1401	3582
SampleTag04_hs	301	40	1102	1590
SampleTag05_hs	51	1218	247	1384
SampleTag06_hs	1084	1770	1340	1466
SampleTag07_hs	236	1845	2570	1324
SampleTag08_hs	1155	2149	1259	1632
SampleTag09_hs	2017	354	634	1984
SampleTag10_hs	3978	51	413	1222
SampleTag11_hs	1858	2679	4039	2473
SampleTag12_hs	-	1083	3803	6515
Undetermined	122	120	353	91

Table 1: *Cell frequencies of each sample contained in each of the batches of the WTA experiment (setA).*

Finally, it is noticed the undetermined category. According to BD rhapsody reference those cells stand for putative cells might not have enough Sample Tag counts to definitively call their sample of origin, and those are labelled as Undetermined. These cells are typically the result of technical issues during the sample preparation or capture stages, such as low-quality tags or insufficient amplification, which hinder accurate assignment. The number of "Undetermined" cells is noteworthy because it reflects the efficiency and accuracy of the sample tagging and cell capture processes. In this dataset, the numbers of "Undetermined" cells vary across batches, with Batch 1 showing 122 "Undetermined" cells, Batch 2 showing 120, Batch 3 showing 353, and Batch 4 displaying 91. While these numbers are not overwhelmingly high, they could still present a potential issue, particularly if they represent a significant proportion of certain sample tags or batches. If a substantial fraction of cells from samples are labelled as "Undetermined," this could suggest technical difficulties with tagging or issues with cell capture efficiency for those samples.

2.3. Quality control

In single-cell transcriptomic analysis, the quality control plays a crucial role in the posterior analysis. For reproducibility, downstream analysis, together with any biological interpretation, demand stringency in quality control. There are several key aspects that are considered through this step. Firstly, the number of total genes detected by cells is fundamental. Cells with abnormally low gene detection rates may result from a low RNA content or inefficient capture and amplification during library preparation. These cells are often considered "low-quality" and excluded from downstream analyses because they contribute limited biological information. On the other hand, cells with a very high number of detected genes may indicate doublets or multiplets, cases where more than one cell has ºbeen captured together and sequenced as a single cell, thereby artificially increasing the

gene counts and distorting cell-specific profiles. Furthermore, the number of genes per cell and number of counts per cell also give insights on the cell type under study. Samples usually sequenced at a certain time have cells with a similar number of total genes per cell defined as complexity. Finding samples of same origin with different complexity may give an insight into a contamination from unexpected cell types

Another key feature that we must consider is the mitochondrial percentage. It gives insight into cell health and possible apoptosis or stress responses. In cases of scRNA-seq, high mitochondrial transcript proportions generally show that the cell's cytoplasmic RNA has degraded, leaving behind the more stable RNA. That could be due to stress, apoptosis, or poor sample handling, in which case data would be biased or uninformative. By filtering out cells with mitochondrial percentages that are too high—above some threshold, such as 5–20%, depending on the experiment—researchers can make sure that their dataset is mostly composed of healthy, viable cells. The high mitochondrial percentage cannot be simply ignored, since doing so might lead to spurious conclusions since apoptotic cells often show distinct transcriptional profiles that faithfully do not mirror the real biological state of the sample.

On the same line, ribosomal percentage of sequencing reads is also quite important. While ribosomal genes play a fundamental role in protein synthesis, it also a fact that its overrepresentation in sequencing libraries can add some noise that unable the detection of low-abundance transcripts. Moreover, a high ribosomal percentage is usually caused by inefficient depletion of rRNA during library preparation or even contamination with cytoplasmic ribosomes. On the contrary, an abnormally low amount may also mean that the cell is in an apoptotic state.

Also, it is also worthwhile to assess the haemoglobin percentage, especially in blood samples from human origin. Specifically, we are working with PBMCs. Mature red blood cells are absent in PBMC preparations and the presence of significant expression of such in our data is an indicator of a contamination from, maybe, red blood cell precursors or cells arising from unproper insolation. In addition, since we work with PBMCs we want to check specifically on immune cells. High haemoglobin expression in PBMCs will interfere in downstream analysis and will mask the expression of this cell types of interest for us.

Finally, it is quite important to understand how all these insights relate to each other. For instance, cells that have undergone high stress naturally use to exhibit higher mitochondrial percentages which also, usually, correlates with the number of genes detected per cell. Additionally, the number of genes per cell must correlate positively with the number of total counts per cell. In the context of our data, quality controls did not arise almost any no concerning factors were identified on any of the samples after inspection of all the batches (See Annex on GitHub QC). However, it is interesting to note that in batch 4, sample 03 contains an abnormally high level of haemoglobin expression. Specifically, expression ranged from 60% to 80% of the total expression counts of all genes in the cell. Most likely it is indicative of an issue in isolation process which could derive from an issue with the density

gradient in the centrifugation step. Considering that it is a single sample, sometimes it is better to proceed with a standard filtering and then backtrack and redo the filtering process if results seem to contain too much noise. From that premise, it was decided to use cells that had more than 300 RNA molecules detected and less that 4000, mitochondrial percentage below 10% and ribosomal percentage above 5%.

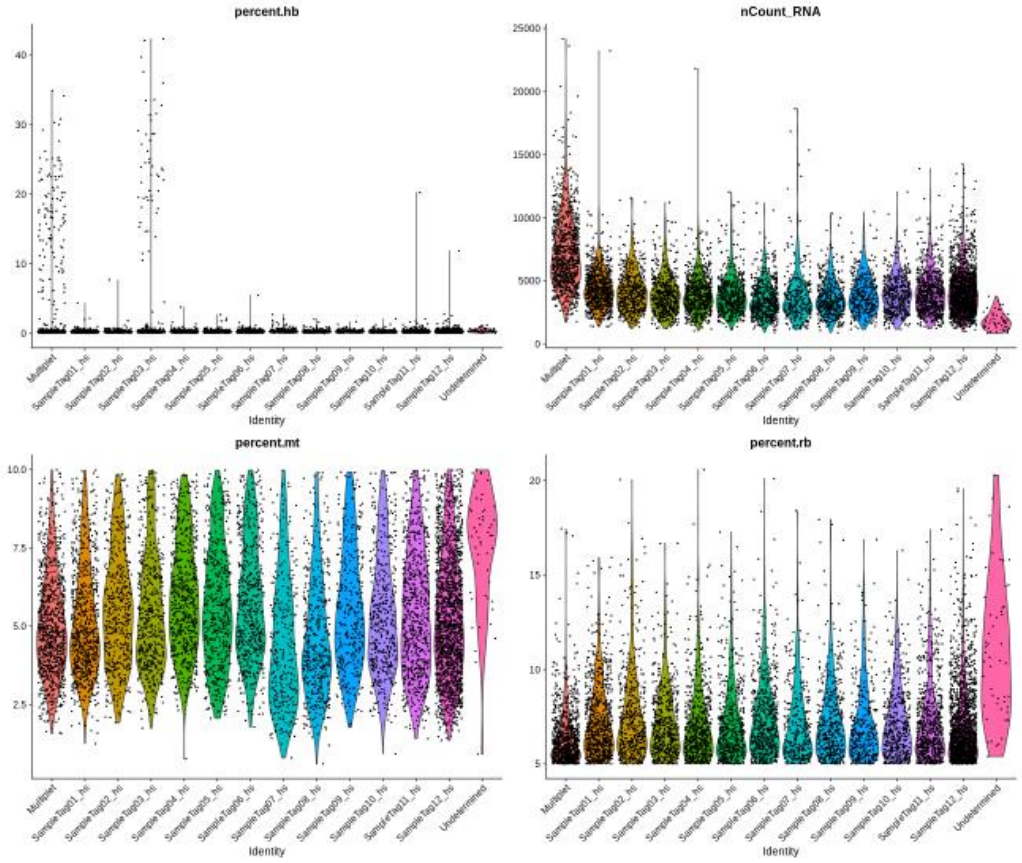

Figure 6: *Quality control of batch 4 samples in WTA showing violin plots on percentage of haemoglobin, number of RNA counts per cell, mitochondrial percentage a ribosomal percentage. Strikingly high haemoglobin percentage in sample03.*

Upon performing the filtering, in all batches the number of usable sets was reduced as follows (Table 2). It is important to notice that mostly multiplets were removed since probably those were cell multiplets rather than label multiplets.

Sample	Batch 1	Batch 2	Batch 3	Batch 4
Multiplet	296	395	480	898
SampleTag01_hs	2	631	550	733
SampleTag02_hs	49	698	488	468
SampleTag03_hs	261	514	475	507
SampleTag04_hs	88	1	296	611
SampleTag05_hs	5	181	55	501
SampleTag06_hs	835	1073	489	559
SampleTag07_hs	64	656	1184	294
SampleTag08_hs	675	858	683	425
SampleTag09_hs	1128	65	172	320
SampleTag10_hs	2176	4	30	435
SampleTag11_hs	818	620	1662	703
SampleTag12_hs		262	1607	1785
Undetermined	94	66	177	51

Table 2: *Cell frequencies of each sample contained in each of the batches of the WTA experiment (setA) after the QC filtering step.*

2.4. Data Processing

It was decided to process the datasets with a standard Seurat procedure by normalizing, scaling, running a PCA and then performing a UMAP reduction.

Up to this stage, we have successfully removed low-quality cells, ambient RNA contamination, and doublets from the dataset. The refined data is now presented as a count matrix, structured as a numeric matrix with dimensions of cells by genes. These counts reflect the processes of molecule capture, reverse transcription, and sequencing performed during the scRNA-seq experiment. However, each of these steps introduces a degree of variability in the measured count depth, even for identical cells. Consequently, differences in gene expression observed between cells within the count data could simply result from sampling effects. As a result, the dataset—represented by the count matrix—still contains variance terms that fluctuate significantly. This variability often poses challenges for analysis, as many statistical methods assume a uniform variance structure in the data.

Therefore, the next step consists of a normalisation of the data which seeks to modify the raw counts within a dataset to account for varying sampling effects by scaling the observable variance to a defined range. In practice, several normalization methods are employed, differing in their levels of complexity. These techniques are primarily crafted to ensure the applicability of subsequent analytical tasks and their associated statistical methodologies. As mentioned in the single cell best practices [74] -- document provided by Theis's lab –author of the well-known tool Scanpy for single-cell analysis in python language-- A recent evaluation conducted by Ahlmann-Eltze and Huber in 2023 [75] assessed 22 distinct transformation methods

for single-cell data. This study focused on comparing the effectiveness of various normalization techniques by examining the overlap between cell graphs and the ground truth. It is important to note, however, that a comprehensive assessment evaluating the influence of normalization on a broad range of downstream analytical tasks has yet to be undertaken.

Having said that, it was decided to apply Seurat's global scaling normalization method based on log normalization. This normalised the feature expression measurements for each cell by the total expression and multiplies this by the scale factor in our case default of 10000. Finally, the function log-transform the result. It is imperative to comment that this method relies on an assumption. This assumes that each cell originally contains the same or a similar number of RNA molecules. That is why it is important to perform a good quality control.

Next step focuses on identifying the highly variable features, which is indeed a feature selection step. This subset of features is the one that exhibits the highest cell-to-cell variation in the dataset meaning that represents genes that are highly expressed in some cells while lowly expressed in others. This step is important as well, maybe even more than normalization. Single-cell datasets are sparse and contain data that can range from the 30000 genes upwards in many cases. Many of these genes are not informative and contain mostly zero counts. Therefore, excluding those uninformative genes might help to represent meaningfully the biological variation in our samples. For our analysis, it was decided to pick 2000 most variable genes.

Following it was proceeded with performing a linear transformation. Seurat scale function shifts expression of each gene with mean expression by 0 and scales it so that the variance across cells is 1, giving equal weight in downstream analysis so that highly expressed genes do not dominate over systematically down-regulated ones. Only the most variable ones are scaled.

After having completed the previous processing, it is time to jump into the dimensional reduction techniques (Fig7). As previously mentioned, single-cell data is sparse due to the high number of genes and cell involved generating matrices of incredibly high dimensions with very large number of columns and rows. As also previously mentioned, not all genes contained in this kind of data are informative so we can try to reduce even more the complexity of our dataset by using dimensionality reduction techniques. Firstly, we perform a Principal Component Analysis (PCA). PCA facilitates the creation of an alternative set of uncorrelated variables, which are refers as principal components, through an orthogonal transformation of the original dataset [76]. Each principal component is derived as a linear combination of the features present in the original data and, these components, are arranged in a descending order based on their variance, which is the factor that defines the nature of the transformation. The foremost principal component, by virtue of the methodology, corresponds to the maximum variance identified in the dataset, while components explaining the least variance are typically disregarded.

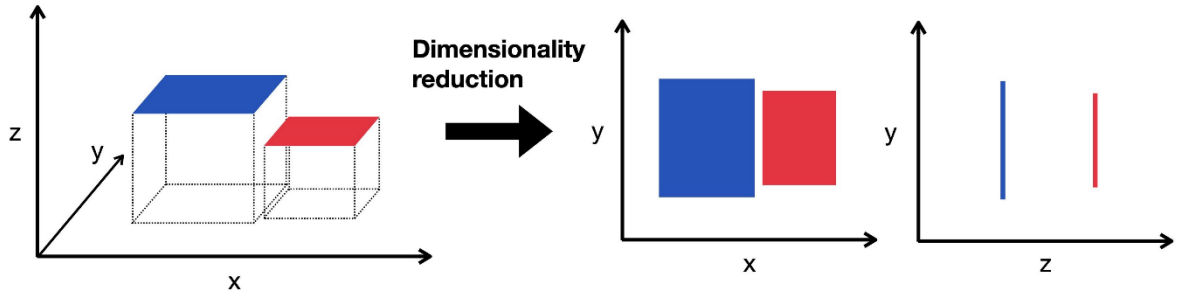

Figure 7. Extracted from Single Cell Best practices book Chapter 9*. Dimensionality reduction embeds the high-dimensional data into a lower dimensional space. The low-dimensional representation still captures the underlying structure of the data while having as few as possible dimensions. Here we visualize a three-dimensional object projected into two dimensions*

The interpretative clarity and computational efficiency of the PCA procedure represent, for sure, significant advantages as well as disadvantages. The sparse nature of single-cell RNA datasets, often resulting from dropout events, complicates the applicability of this linear dimensionality reduction technique. Consequently, the visualisation outcomes produced by PCA may not reflect the complex non-linear characteristics inherent from such datasets.

In practice, PCA is generally employed to identify and select the top 10 to 50 principal components, which are subsequently utilized in various downstream analytical procedures. In our analysis, again, we used Seurat's procedure for PCA. This functionality outputs a list of genes with the most positive and negative loadings that represent modules of genes that exhibit either correlation or negative corelation across the cells contained in our dataset. We decided to go for 30 dimensions according and then checked the elbow plot generated for later analysis generated (Fig 8).

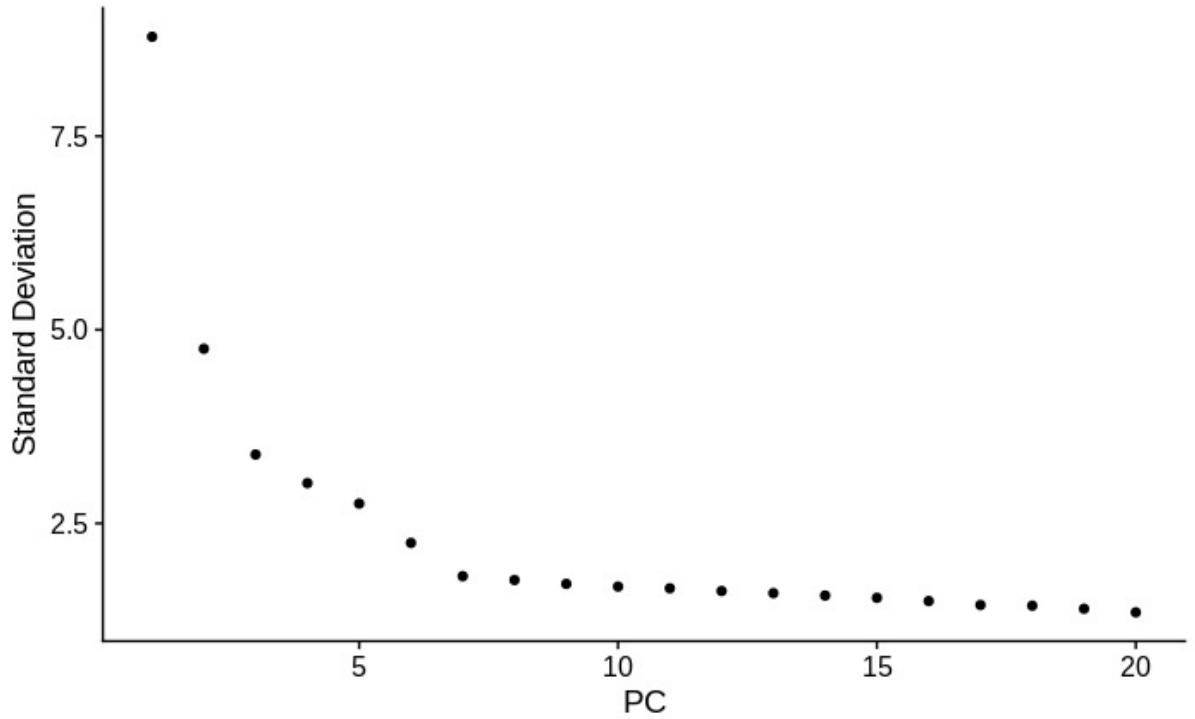

Figure 8: *Example of Elbow plot giving information on the variability explained of each of the principal components in batch 1.*

UMAP represents a non-linear dimensionality reduction method that operates on a graph-based framework, sharing fundamental similarities with t-SNE. This technique initiates by creating a high-dimensional graph representation of the dataset, followed by the optimization process aimed at ensuring that the low-dimensional graph representation retains structural fidelity to the original graph configuration.

From my side, it was preferred to use UMAP since it gives meaningful information in terms of biological differences to the distance between cells in separated clusters, while t-SNE is not capable of doing this. Therefore, Seurat functionality for UMAP was applied 20 dimensions (Fig 9).

After finishing the process, we could realise that after filtering, even using low filters, we lost many of our cells in certain specific samples making the differential expression analysis difficult. Even it is hard to visualise enough cells in the UMAP to extract visual insights. However, the main purpose of this data was supporting the findings from the immunoprofiling step to link changes in TCR / BCR repertoire upon possible conditions to gene expression and, as a mater student, learning the knowledge about the technical part of the procedure.

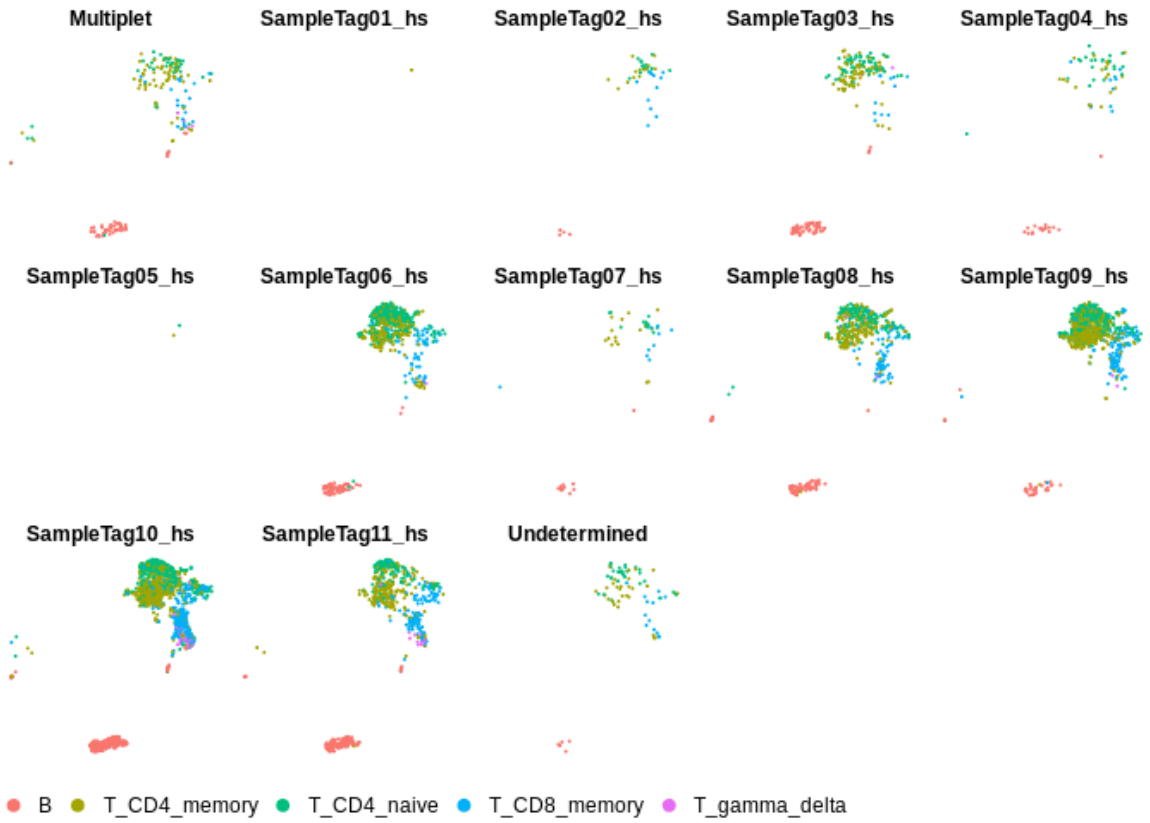

Figure 9: *UMAP on calculated from the first 20 PC. Representation is split by sample type and coloured by cell type annotation given by ABseq on batch 1.*

Even so, to deepen the knowledge on single-cell analysis, it was decided to integrate the batches to learn how to mitigate batch effect. For that, we must remind that the project started and still uses Seurat v4, sot the layers were still not present as data structures as in Seurat 5. Therefore, we used STACAS software for integration. It is a semi-supervised method that is especially suited for data with previous knowledge on cell type annotation, as it is our case. Commonly, other kind of integration are totally unsupervised, meaning that they do not take into consideration prior knowledge. While normal integration in Seurat typically relies on identifying shared features across datasets through techniques like mutual nearest neighbours (MNN) or canonical correlation analysis (CCA) without using prior knowledge of cell types, STACAS leverages known cell type labels to guide the integration process. This semi supervised approach can lead to more accurate alignment and better preservation of cell type-specific features, particularly in situations where cell type annotations are reliable and available. In cases of complex datasets or when integrating heterogeneous conditions, STACAS may improve the resolution of clustering and reduce the risk of mislabelling or misalignment between datasets. In figure 10, we can see how the performance of our integration worked perfectly and now all the batches can be visualised under the same dimensional space, allowing posterior comparison if needed or possible.

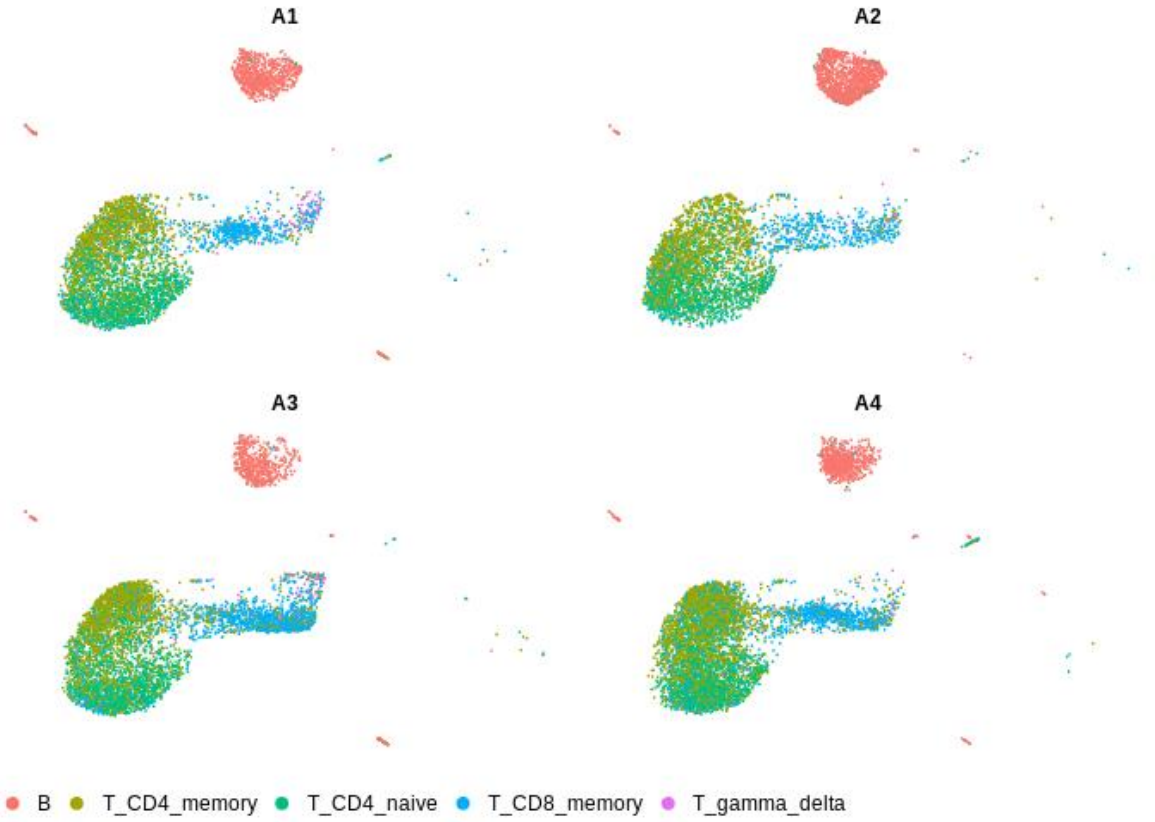

Figure 10: *Integrated UMAP obtained with the semi-supervised tool STACAS. This figure is represented by splitting by sample type and coloured by cell type annotation given by ABseq on the different batches.*

3. Single-Cell Immunoprofiling

3.1 Upstream Analysis

Our immunoprofiling, being specific, comes from the BD-Rhapsody Immune Response experiment. As it has been repeated over the thesis, the human immune system constitutes a sophisticated network comprising various organs and cell types that collectively serve to defend against pathogens, environmental antigens, and neoplastic cells. Central to the immune response are the intricate interactions among multiple cell types, each fulfilling specific roles in the immune defence mechanism. BD Rhapsody Immune Response Targeted Panel (Human) employs a multiplex polymerase chain reaction (PCR) methodology to facilitate the detection of 399 genes selected for the profiling of human immune cells. Initially, cellular lysis is performed within the BD Rhapsody Cartridge, after which magnetic retrieval is conducted to extract beads that have captured messenger RNA (mRNA) needed for complementary DNA (cDNA) synthesis.

After the cDNA synthesis, multiple rounds of gene-specific nested PCR are performed utilizing the included primers to construct the library necessary for sequencing. The resulting amplification products from the final PCR stage are composed of sequencing adapters, a designated cell label, a unique molecular index (UMI), and up to 400 base pairs of the 3' terminus of the target gene. The assay products generated are compatible with sequencing using Illumina MiSeq, NextSeq, HiSeq2500, and HiSeq4000 sequencers, thereby providing a robust framework for the analysis of the immune response.

This experiment generates the sequence files that undergo a similar process as explained for WTA. In this case, again, we make use of BD Rhapsody provided Seven Bridges software. The alignment reads are aligned to the reference sequence serving a dual purpose, facilitating the identification of their biotype and establishing their genetic identity. As we are regarding WTA in conjunction with TCR and BCR assessments, the established pipelines identify TCR and BCR reads that exhibit alignment with recognized VDJ gene segments within the transcriptome. It is essential that such alignment occur in the correct orientation.

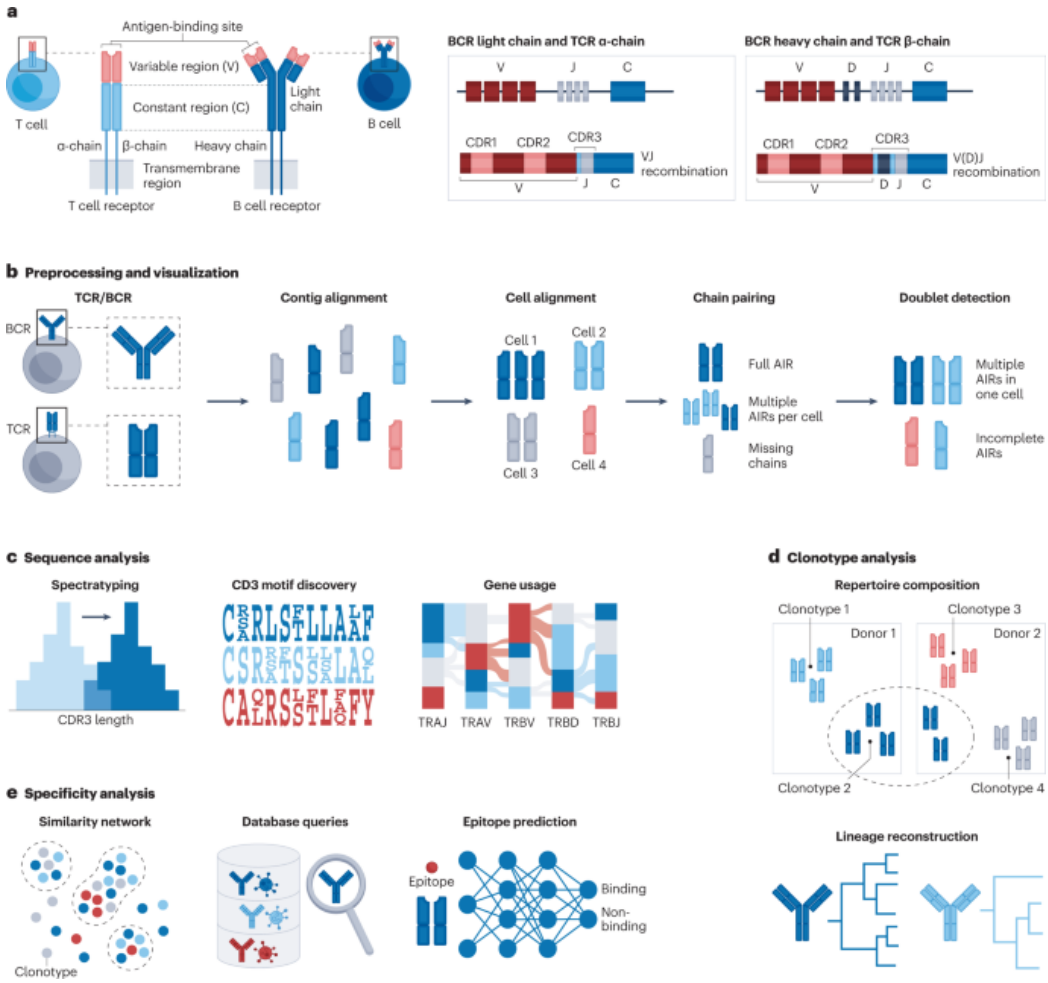

Figure 11: *Figure extracted from best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). Overview of the adaptive immune receptor analysis.*

However, in our case this is a targeted immunoprofiling assay, meaning that it will not regard WTA sequences while it indeed will have the same barcode. In this context our targeted assay that incorporates the TCR and BCRE analysis will follow a pipeline programmed to automatically append species-specific TCR and BCR gene segments to the FASTA reference file. These gene segments are sourced from the GTF utilized in the assembly of the pro-configured WTA assay reference archive. Once again, we used hg19 also called GRCh38 as mentioned in the previous section. Subsequently, the reads aligning with the TCR and BCR gene segments undergo a process of grouping, thereby distinguishing and isolating them from reads that correspond to other biotypes.

To construct the complete variable region of the immune cell receptor or to derive a consensus complementarity determining region 3 (CDR3), it is imperative to assemble the short reads. The process of read assembly functions by identifying similarities and overlaps among the reads, which indicate a common origin from a single sequence. By aligning and concatenating these reads, it becomes feasible to produce longer contiguous sequences –that we name contigs – from the short reads while simultaneously rectifying randomly distributed sequencing errors.

Before the assembly can commence, it is essential to prepare the reads categorized as TCR and BCR derived through trimming and UMI correction. Initially the 3' of the reads are trimmed, utilizing a quality score threshold of 20 to ensure that high-quality are used in assembly. Furthermore, reads are trimmed based on bead capture sequences, which serve to eliminate artifacts associated with the BD Rhapsody™ cell label sequence. These artifacts may obscure the assembly process and are often present at the 3' end of TCR or BCR reads sourced from short amplicons. Subsequently, UMI error correction is performed, with reads grouped according to their cell ID and the identified TCR or BCR chain type established during the initial alignment (e.g., TCR-Alpha, IG-Kappa).

The assembly process commences with the grouping of reads based on their respective cell IDs and chain types. These categorized read groups are subsequently processed using a transcript assembly software package known as Trinity, which yields a compilation of contig sequences. Following this, the reads are realigned to the newly formulated contigs to facilitate the generation of read and molecule counts for each individual contig. The multiple contigs originating from each cell encapsulate the rearranged VDJ mRNA sequences, exemplified by TCR Alpha and TCR Beta chain variants.

In the process of analysing contigs generated during the assembly phase, various gene segments, including V, D, J, and C segments, as well as complementarity determining regions (CDR1-3) and framework regions (FR1-4), are identified. Furthermore, the assessment includes determining productivity, defined by the absence of stop codons, the full-length status of the contig, and the corresponding protein sequence. This comprehensive analysis is conducted utilizing a software package known as IGBlast, supplemented by alignments executed through Bowtie2.

Contigs exhibiting low-quality V or J gene selection are excluded from subsequent analyses. This low quality is determined by an e-value score exceeding 10^-3, with lower scores indicating better quality. A contig is classified as "full length" when a complete amino acid sequence is established for each of the framework (FR1-4) and CDR (CDR1-3) regions. It is important to note that while the FR1 and FR4 regions may be partially defined, the overall contig can still be regarded as full length.

Ultimately, all annotated contigs are documented in an unfiltered output file that conforms to AIRR compliance standards.

This AIRR file is the one that will be used in the downstream analysis. Resulting AIRR file as set B were comprised of the same cells as set A (WTA) in a multiomic manner. Due to the multiplexing explained in the previous section, it was possible to match the information from cells in set A to cells in set B. Therefore, while we know that some information on the transcriptional side is incomplete – e.g. low-quality cells, multiplets, duplets— the immunoprofiling assay was analysed in an independent manner to be able to assess its performance without bias towards previous knowledge obtained from WTA assay.

3.2 Downstream Analysis

Upon obtaining the AIRR files the downstream analysis was ready to start. To perform this analysis, it was decided to stick with R programming language and use the well-known tool scRepertoire[77]. It is highly regarded for its ability to integrate receptor information with transcriptomic data, offering insights into clonal expansion, diversity, and dynamics in the immune system. By linking clonotype information with gene expression profiles, scRepertoire facilitates an in-depth exploration of immune responses, making it especially useful in contexts such as tumour immunology, infection studies, and autoimmune diseases. Its intuitive workflow and visualization capabilities make it a reliable choice for researchers aiming to uncover how clonotypes correlate with functional states or specific cell phenotypes, that's why it was chosen for this project.

Alternatively, despite its strengths, scRepertoire is not the only tool available for immune profiling, and alternatives may complement or expand its functionality depending on the research question. For instance, tools like Immcantation[78] provide advanced statistical methods for analysing somatic hypermutation and clonal evolution in B-cell repertoires, making it a powerful choice for studies focusing on adaptive immunity. Similarly, VDJtools [79] and scirpy [80] are other robust platforms tailored for repertoire analysis. VDJtools emphasizes cross-sample repertoire comparisons and diversity analysis, while scirpy integrates smoothly with single-cell data and offers flexibility in exploring clonotype distributions, being the later also part of the Scanpy ecosystem which is rapidly growing lately as a big open-source python community project (Scverse).

Choosing between these tools often depends on the specific focus of the analysis. For example, scRepertoire excels in integrating immune receptor analysis with broader

transcriptomic insights, making it particularly appealing for studies aiming to link clonal expansion with cell function. For this reason, and being easier to integrate with Seurat, scRepertoire is the chosen tool.

Clonotype Quantification

To start the analysis the samples were analysed in their respective batch independently and sorted out by BCR and TCR to analyse it receptor independently. For that we started checking the percentage of unique clones found in each of the samples in each batch. This value is important to take into consideration in immune analysis since it reflects the diversity of clonotypes within a sample and provides insights into the adaptive immune response.

As mentioned in the introduction section in Chapter 1, Clonotypes refer to groups of immune cells—such as T-cells or B-cells—that share the same receptor sequence, typically defined by their unique combination of variable, diversity, and joining VDJ along with the CDR3. These receptor sequences are the result of somatic recombination and are responsible for recognizing specific antigens. In our samples, all batches behave in a similar manner even either in TCR and BCR results. With some examples, our samples seem to have a high percentage of unique clones (Fig 12, 13), around 40% to 50%, which suggest a high diversity in the immune response which is not to be expected under COVID infection condition.

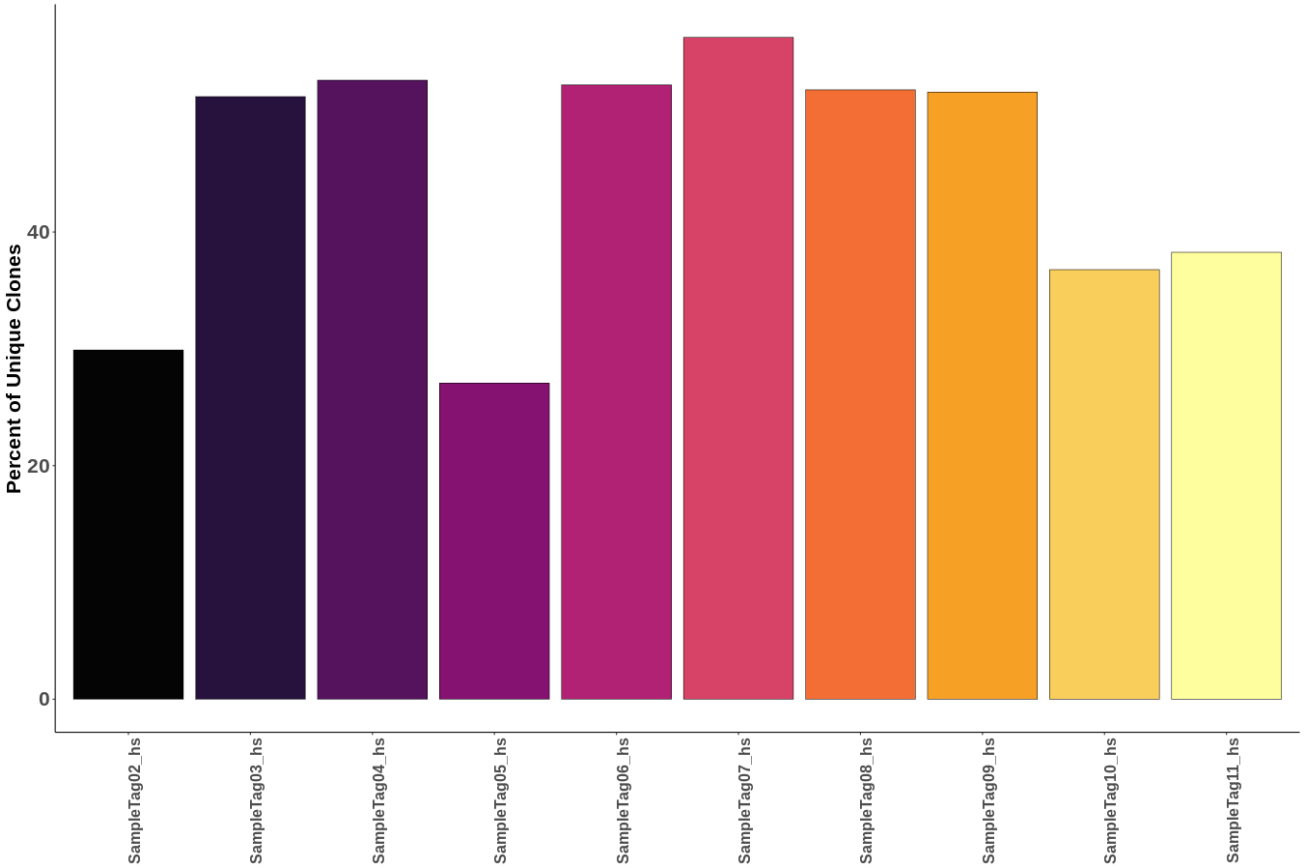

Figure 12: *Percentage of unique clonotypes in batch 1 divided by sample type under TCR context. The colour represents the sample for which the percentage of unique clonotypes has been calculated*

This diversity is often associated with a healthy immune system capable of recognizing a wide array of antigens. Conversely, lower diversity might suggest clonal expansion, where specific clones have proliferated in response to a particular antigen, as seen during infections – as our case with COVID-19 --, cancer, or autoimmune diseases. During an infection, the adaptive immune system undergoes clonal expansion, where specific T-cell and B-cell clones proliferate to target viral antigens. This expansion often results in a reduction in the percentage of unique clones as dominant clonotypes emerge in response to the infection. For instance, expanded T-cell clonotypes are typically associated with cytotoxic or helper T-cell responses targeting viral peptides, while expanded B-cell clonotypes reflect the production of neutralizing antibodies by plasma cells.

We can extract different hypothesis from this first results as first instance. If we assume that the experiment ran smoothly, the observation of ~50% unique clones in our samples suggests that while clonal expansion is occurring, there remains a substantial level of immune diversity, which could indicate a robust and multifaceted immune response. This is consistent with findings from studies on COVID-19, where individuals with severe disease often show a narrower repertoire dominated by hyperexpanded clones, while those with milder symptoms maintain greater diversity. The presence of a balanced repertoire could also reflect ongoing antigenic exposure, immune regulation, or the recruitment of new naive cells to the response. However, we know that our patients are some under severe circumstances, so it does not make much sense.

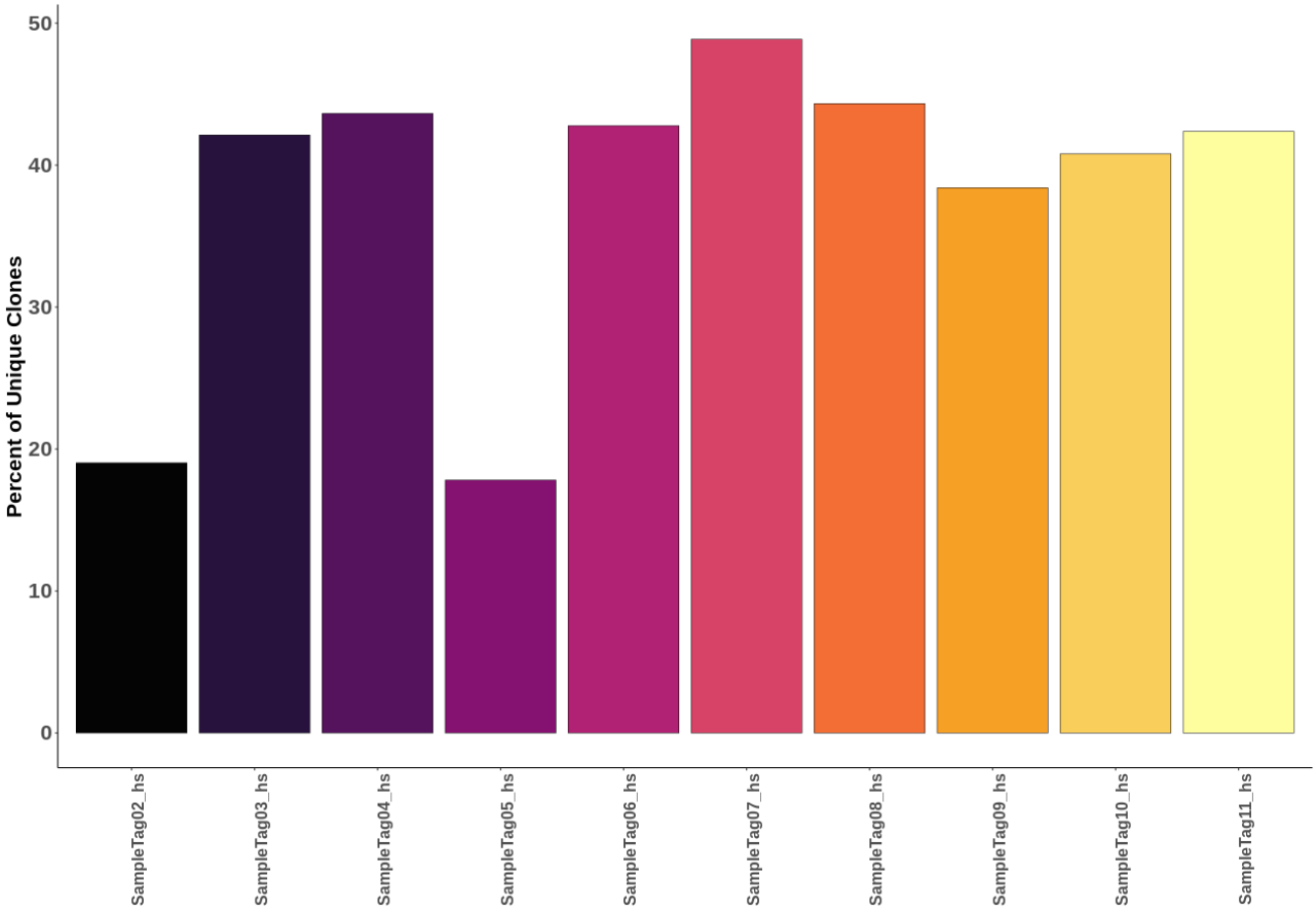

Figure 13: *Percentage of unique clonotypes in batch 1 divided by sample type under BCR context. The colour represents the sample for which the percentage has been calculated.*

Experimentally, this observation could raise questions if there are technical artifacts affecting the data. For example, insufficient sequencing depth can artificially inflate the percentage of unique clones, as rare or low-frequency clones may not be adequately captured. Conversely, overamplification during library preparation can introduce PCR artifacts, leading to an overestimation of clonotype diversity. It is also important to confirm that the definition of "unique clones" is consistent and appropriate for your analysis pipeline, as small variations in sequence alignment or clustering thresholds can influence results.

In our case, we tried from the most stringent to the laxest settings possible to try to retrieve explorable results. For that reason, t was decided to assign clonotypes by gene as shown in figures 12 and 13. This definition captures clonotypes at a level that reflects shared antigen recognition potential without requiring identical CDR3 sequences. The "Gene" setting is particularly useful when studying broader clonal relationships or when dealing with datasets where CDR3 information may be incomplete or prone to errors.

This approach is advantageous in several scenarios. By focusing on V and J gene usage, you can explore clonal expansion patterns and immune diversity at a level that balances biological relevance and analytical simplicity. This is particularly relevant in COVID-19 research, as

certain V and J gene combinations have been associated with SARS-CoV-2-specific responses. However, it is worth noting that this method may group together clonotypes that are functionally distinct, as it does not account for the CDR3 sequence, which plays a critical role in determining antigen specificity. This might lead to an underestimation of true clonal diversity. Even under these assumptions, it seems at first sight that there is no expansion involved.

Clonotype Diversity

Following the previous results, even if there were a large amount of unique clonotypes, we proceeded with the analysis of the clonal diversity in case it was possible to capture a specifically abundant clonotype. For that, we decided to check the clonal splits by separating the clonal proportions in bins with clonalProportion() function. This function provides a ranked analysis of clonotypes, grouping them into bins based on their frequency of occurrence. Unlike clonal space homeostasis, which focuses on the relative proportion of each clone to the total repertoire, clonalProportion() ranks clonotypes by their abundance and categorizes them into defined bins, such as the top 10, top 100, and so on. This ranking approach provides a clearer picture of the contribution of dominant clonotypes to the overall immune response, making it particularly useful for understanding clonal expansion and diversity since it can reveal how much of the immune response is driven by a small number of highly expanded clones versus a broader range of less abundant clonotypes.

More specifically the distribution of the splits is:

- From top 1 to 10
- From top 10 to 100
- From top 100 to 1000
- From top 1000 to 10000
- From top 10000 to 30000
- And from top 30000 onwards

Again, in these results, it was clear that there was not a clonal expansion of a specific clonal group in our data in any of the batches of any of the samples (Fig 14). We show in this report an example of the figure results, to check all the results please refer to the annex. In the example below showing the clonal expansion of TCR in under the batch 1 context, all samples have a clonal expansion of less than 30% for clonal groups ranked from the top ranked bin (ranging from top 1 to 10). Again, biologically, this suggests that the immune repertoire in your samples is relatively diverse, with no single group of highly expanded clonotypes dominating the response. Experimentally, it confirms that it is imperative to check on the number of cells and receptors available in this dataset. If the number of cells is too small, it will be needed to compensate the low statistical power with the required tools.

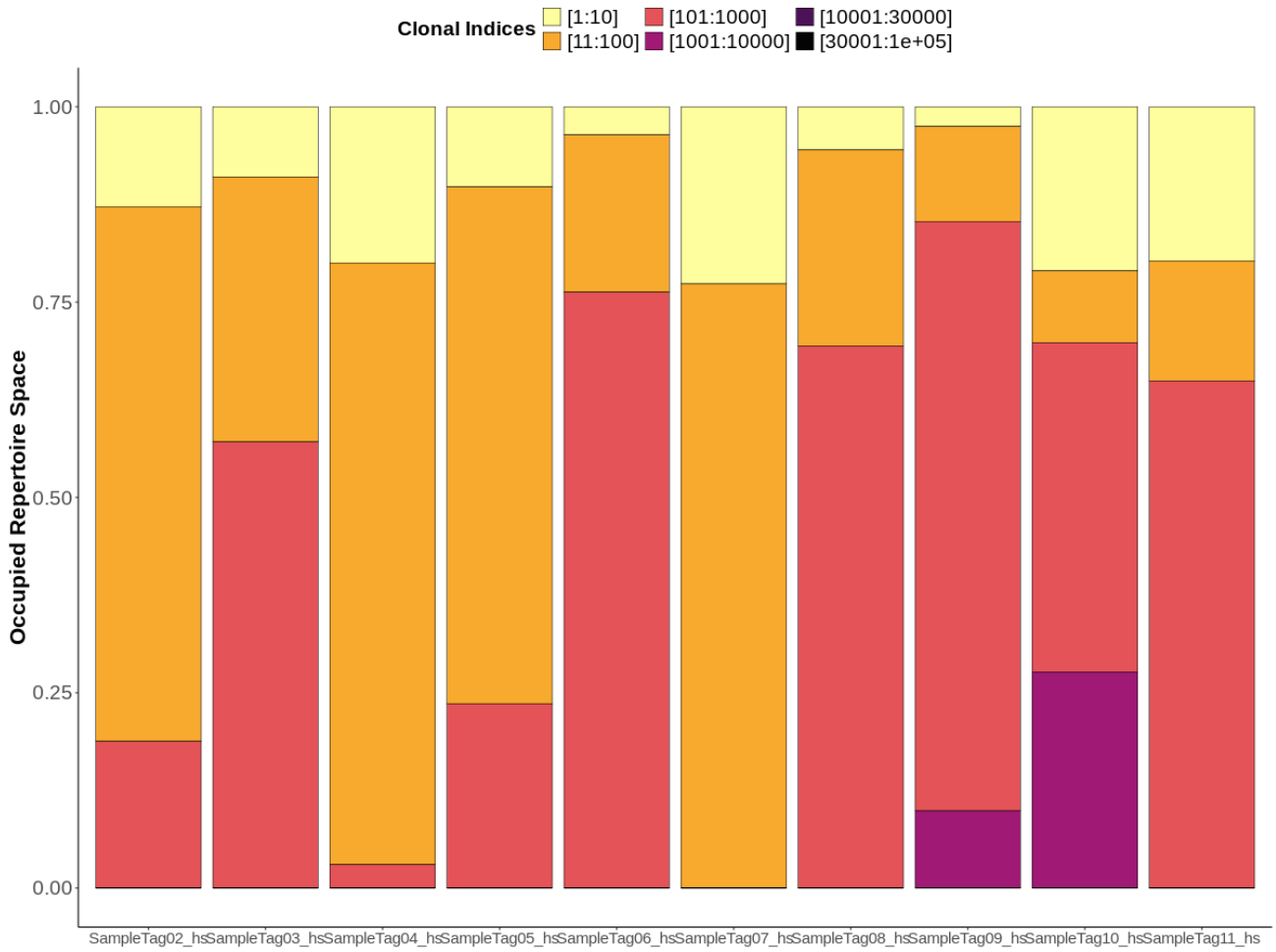

Figure 14: *Clonal proportions assessed in T-cells for batch 1. Colours represent the bins for clonal indices ranging from 1:10, 10:100, 100:1000, 1000:10000, 10000:300000, and 30K onwards.*

Integration with WTA assay in Seurat

Even with the low number of receptors available, it was decided to proceed with the planned pipeline to learn how to perform further sections of the analysis.

The combineExpression() function in scRepertoire it the utility that allowed us to integrate with gene expression profiles obtained from WTA assay. The combineExpression() function assigns clonotype metadata, such as clonal frequency or group membership, to individual cells within the expression dataset. This integration allows you to stratify cells based on clonotype characteristics and assess whether specific clonal expansions are associated with distinct gene expression signatures. For example, you can identify whether expanded T-cell clones exhibit markers of cytotoxicity, exhaustion, or proliferation, or whether expanded B-cell clones are enriched for antibody production pathways. These insights are invaluable for characterizing the functional heterogeneity of immune responses, particularly in the context of infection, where clonal expansion often drives specific effector functions.

While combineExpression is a robust and widely used tool for this purpose, alternative approaches can also be considered depending on the specific needs of your analysis. One

alternative is using the Scannpy framework, which offer tools for integrating clonotype metadata as part of their single-cell analysis workflows. This platform allows for flexible downstream analysis, such as differential gene expression, trajectory inference, or clustering, while incorporating clonotype information. For instance, you can add clonotype metadata to Scanpy's *.obs* data structure to annotate cells with clonotype-related attributes.

From the representation of the combined UMAP of VDJ with expression (Fig 15), after integrating using scRepertoire, it was needed to show information on both ends of the assay. The ClonalOverlay() function in scRepertoire is a powerful visualization tool that allows to overlay clonotype information directly onto dimensionality reduction plots, such as those generated through UMAP or t-SNE. This function enables to visually map how clonotypes are distributed across different clusters or cell populations within your dataset, providing critical insights into the relationship between clonal dynamics and cellular phenotypes. In the context of the COVID-19 study, this functionality is particularly important for understanding how clonally expanded cells are associated with specific immune subsets or functional states. The visual output of ClonalOverlay also facilitates comparisons between different timepoints or conditions. For example, it is possible to track how the spatial distribution of clonotypes evolves during infection or even recovery. This is particularly relevant in longitudinal studies, where the immune repertoire may undergo significant changes over time. In our case, it is interesting to show that the multiplet section of the figure represents a big number of cells. After having commented throughout the chapter that we are missing statistical power in our analysis, it is interesting to notice the big number of cells that we are missing because of demultiplexing process. This fact will be developed in detail in the following limitations and discussion of the chapter.

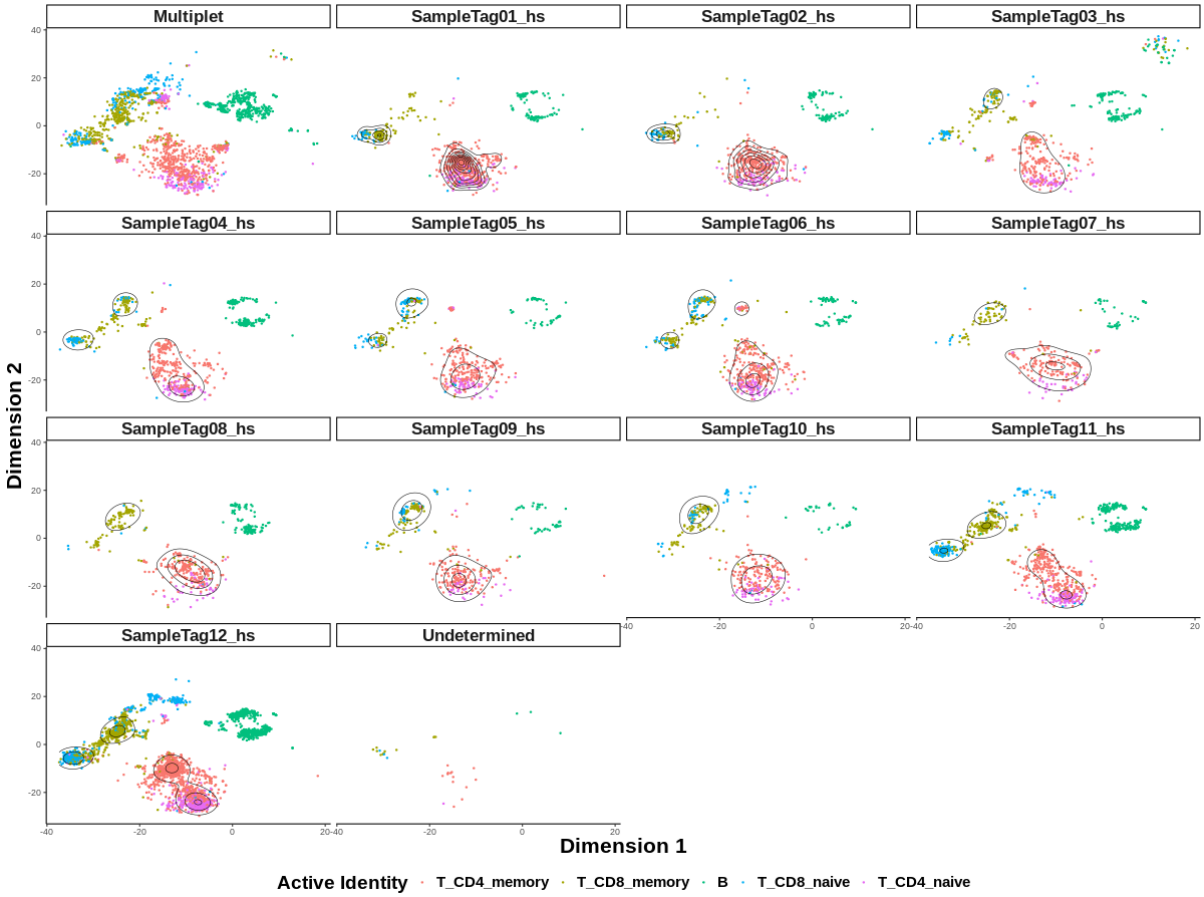

Figure 15: *TSNE combining the celltype information extracted from the WTA assay in Seurat with the distribution of the most dominant clonotypes in the VDJ analysis. This allows to identify the position in the transcriptomic space of the most dominant immune receptors. This figure shows only batch 4.*

4. Limitations and Discussion

Several limitations should be considered in the context of this analysis, which impact the interpretation of the results and guide future improvements. One of the primary challenges was the retrieval of a relatively low number of cells across the samples. This limitation reduces the overall statistical power and potentially biases the analysis, as key immune cell subsets or clonotypes may be underrepresented. A limited number of cells can also hinder the ability to accurately assess immune diversity, clonal expansion, and repertoire dynamics, especially when investigating complex immune responses such as those observed during COVID-19 infection.

Another significant issue encountered was the presence of a substantial proportion of cells classified as multiplets or undefined. Multiplets, where two or more cells are mistakenly captured and sequenced as a single entity, artificially inflate the number of detected clonotypes and compromise the biological accuracy of the data. Similarly, undefined cells, which could not be confidently assigned to a specific clonotype or cell type, introduce ambiguity into the dataset and reduce the reliability of downstream analyses. These artifacts

are likely due to the limitations of the initial demultiplexing method, underscoring the need for more robust strategies in future experiments.

To address these challenges, a different approach to demultiplexing will be necessary in subsequent steps. Improved computational methods or experimental protocols that are better equipped to resolve multiplets and accurately assign clonotypes could significantly enhance data quality. For instance, employing methods such as computational doublet detection tools or optimizing sample barcoding strategies may help reduce the prevalence of multiplets. Similarly, adopting algorithms designed for more precise demultiplexing could improve the classification of undefined cells.

Finally, an additional limitation of the current analysis is the absence of direct V(D)J sequencing data in these samples. While scRepertoire was used to analyse immune repertoires, its reliance on pre-annotated TCR and BCR sequences limits the scope of the analysis. To address this, future work will focus on inferring V(D)J information directly from whole transcriptome sequencing (WTS) data. While this approach has its challenges, such as lower sensitivity compared to dedicated V(D)J sequencing, it offers an opportunity to extract immune receptor information from existing datasets, potentially uncovering clonotypic insights that were not accessible in this analysis.

Chapter 3: Demultiplexing in Droplet-based Sequencing

This project presents a framework with the aim of improving the statistical power of the VDJ dataset. To do so, it is proposed the following Diagram (Figure 16). The methodologies, results and limitations of the first step involving the genotype-based demultiplexing will be presented and discussed.

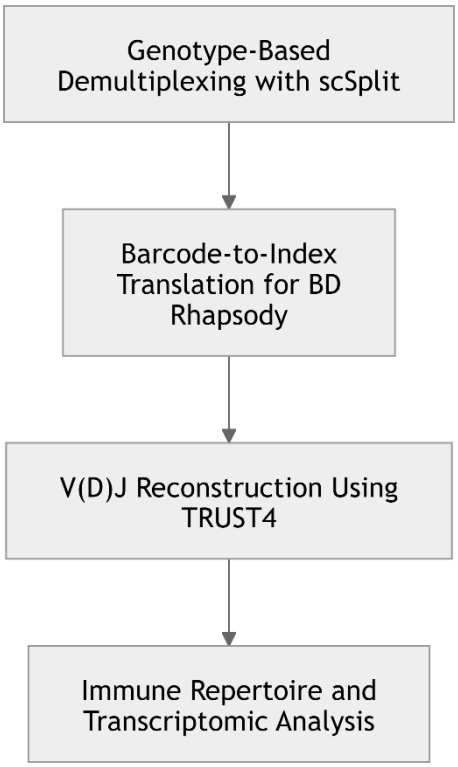

Figure 16. *Flowchart diagram of the simplified framework proposed in this project combining demultiplex and reconstruction of VDJ from scRNAseq WTA assay.*

1. Introduction

Demultiplexing is a cornerstone of droplet-based single-cell sequencing workflows, playing a critical role in ensuring the accuracy and interpretability of the data. In this sequencing approach, individual cells are encapsulated within tiny droplets, where molecular barcodes and unique molecular identifiers (UMIs) label their RNA or DNA content. These barcodes are essential for linking sequencing reads back to their respective cells, allowing to disentangle the pooled data and achieve true single-cell resolution. Without effective demultiplexing, the data would lose its biological fidelity, resulting in errors such as incorrect cell assignments, artificially inflated gene expression values, or the misrepresentation of clonal dynamics. This would ultimately hinder the ability to draw meaningful conclusions from any study.

The importance of demultiplexing becomes particularly evident in multiplexed experimental designs, where samples from different individuals, conditions, or timepoints are combined within a single sequencing run. Multiplexing offers numerous advantages, including increased efficiency and reduced batch effects, but it also demands robust computational tools to ensure that cells are correctly assigned to their original samples. For studies that rely on immune repertoire analysis, such as this one, the accuracy of demultiplexing is even more critical. Clonotypes identified through V(D)J profiling must be linked to their respective transcriptional profiles with precision, as any errors in cell assignment can obscure key biological insights into clonal expansion, immune diversity, and functional heterogeneity.

Robust demultiplexing is also vital for addressing technical challenges inherent to droplet-based sequencing, such as the occurrence of multiplets. Multiplets arise when two or more cells are encapsulated within the same droplet, leading to hybrid profiles that distort downstream analyses. Effective demultiplexing strategies can identify and remove these multiplets, preserving the integrity of the data and ensuring that the observed gene expression patterns accurately reflect single-cell behaviour.

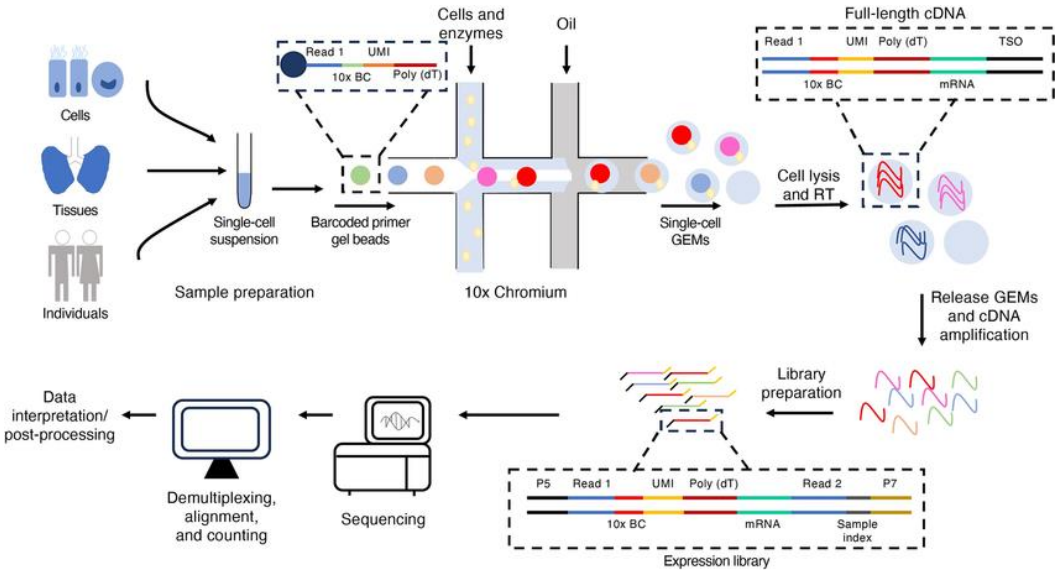

Figure 17: Extracted from Swaminath S, Russell AB (2024) The use of single-cell RNA-seq to study heterogeneity at varying levels of virus–host interactions. *Single-cell sequencing workflow for droplet-based approaches exemplified by the 10x Genomics Chromium platform Infected cells from cell culture, tissues/organoids, or infected individuals are dissociated into a single-cell suspension. The cell suspension is loaded onto a microfluidic chip, and cells are partitioned into nanoliter-scale Gel Beads-in-emulsion (GEMs) droplets containing barcoded gel beads and reagents for reverse transcription (RT). Following cell lysis, the beads capture the mRNA molecules.*

Similarly, cells classified as undefined due to ambiguous barcodes or low-quality data can complicate analyses and reduce the reliability of the results. Addressing these limitations requires the use of advanced demultiplexing methods capable of resolving these ambiguities and enhancing the quality of the dataset.

In the context of this project, demultiplexing takes on an additional layer of importance due to the integration of single-cell transcriptomics with immune repertoire data. By linking V(D)J information to gene expression profiles, it becomes possible to study the functional states and phenotypes of clonally expanded cells, offering valuable insights into the immune response during infection. However, this level of integration relies on accurate cell assignments at the demultiplexing stage. Any errors in this process could misrepresent the relationship between clonal dynamics and cell functionality, potentially undermining the conclusions of the analysis.

This chapter will explore the role of demultiplexing in droplet-based sequencing and its significance for single-cell studies, with a particular focus on its implications for immune repertoire analysis. It will discuss the challenges posed by technical artifacts such as multiplets and undefined cells, while also addressing the strategies used to overcome these limitations. Additionally, this chapter will highlight the methodologies employed in this study to ensure the robustness of the demultiplexing process, setting the stage for the subsequent analyses of immune clonal dynamics and transcriptomic profiling. Through this discussion, the chapter aims to underscore the central role of demultiplexing in preserving the biological resolution and statistical power required for meaningful discoveries in single-cell research.

2. Original BD-rhapsody demultiplexing

The BD Rhapsody system enables sample multiplexing using the BD Single-Cell Multiplexing Kit, which assigns specific Sample Tags to different cell samples. These tags facilitate the identification of sample origins during sequencing. The system supports up to 12 species-specific Sample Tags for human or mouse samples and up to 24 species-agnostic Sample Tags in the flexible kit, ensuring compatibility across a variety of experimental designs.

When running the BD Rhapsody Sequence Analysis Pipeline, users can specify sample names associated with Sample Tags, and the pipeline integrates Sample Tag barcodes into the reference files. Reads aligning to a Sample Tag are used to identify the sample of origin for each putative cell. The sample determination algorithm then evaluates and categorizes cells based on their Sample Tag data.

The algorithm initially identifies high-quality singlets, which are cells where more than 75% of Sample Tag reads come from a single tag. Noise from other Sample Tags is attributed to contamination from oligo manufacturing, incomplete washing, or residual labelling during cell preparation. For each Sample Tag, a minimum read count threshold is established based on the lowest number of reads observed in high-quality singlets, ensuring robust identification. Cells exceeding this threshold are labelled with the corresponding Sample Tag and may include singlets or multiplets (cells associated with multiple tags).

Noise levels for each Sample Tag are calculated to determine a per-cell noise trend. A trend line is fitted to the data, correlating total Sample Tag counts with noise levels. Cells that

deviate significantly from this trend are flagged as potential multiplets, indicating the presence of multiple cells in the same droplet from different samples.

To improve sample calling accuracy, the algorithm subtracts expected noise counts from each Sample Tag, based on the trend line and overall noise contribution. Cells with adjusted Sample Tag counts above the threshold are identified as called cells. Multiplets are flagged when counts for multiple Sample Tags exceed their respective thresholds, while cells without sufficient Sample Tag counts are categorized as Undetermined.

3. BD Rhapsody Algorithm Downsides

As described through our COVID-based case study dataset, it seems that the number of cells classified as multiplets is rather high. After presenting the demultiplexing algorithm offered by BD it is possible to identify several potential challenges demultiplexing process, as outlined in the documentation, which may impact data quality and the accuracy of sample identification.

Firstly, the demultiplexing algorithm is sensitive to noise caused by residual Sample Tag labelling, incomplete washing during cell preparation, or barcode contamination from oligo manufacturing. These issues can lead to false positive associations between cells and Sample Tags, especially in cases with low Sample Tag counts. Noise management is partially addressed through thresholding and subtraction, but this step may not eliminate the impact of contamination. Furthermore, the determination of the minimum Sample Tag read count threshold for singlets is based on empirical observations of high-quality singlet cells. If the threshold is not appropriately set or is too conservative, it may misclassify valid cells as undetermined or fail to detect low-abundance cell populations. On the other hand, overly lenient thresholds could increase the inclusion of noisy or ambiguous cells, compromising the accuracy of downstream analysis.

Also, multiplets, where multiple cells from different samples are encapsulated in the same droplet, pose a challenge to accurate sample identification. The process does not resolve multiplets but rather categorizes them, potentially leading to a loss of single-cell resolution in affected droplets as happening in our study.

4. SCsplit demultiplexing algorithm

To tackle the complexities of sample demultiplexing, we have integrated scSplit [81], a sophisticated computational tool that employs genotype information to accurately assign single-cell sequencing reads to their respective donors in multiplexed experiments. scSplit provides a reliable methodology for determining sample origins, particularly beneficial in studies where cells from multiple individuals are combined in a single sequencing run. By utilizing genetic variation as a natural barcode, scSplit ensures high precision in demultiplexing, negating the need for supplementary experimental alterations, such as Sample Tag labelling. This feature is especially advantageous in contexts where precision is

paramount, including investigations of immune repertoire diversity or disease-specific cellular responses.

scSplit operates by analysing single-nucleotide polymorphism (SNP) information from the sequencing data. During its workflow, the software matches the SNP profile of each cell to a reference panel of donor genotypes, which can be derived from whole genome sequencing (WGS), exome sequencing, or other genetic datasets. By aligning sequencing reads to the SNPs and calculating the likelihood of a match to each donor, scSplit probabilistically assigns cells to their donors. Cells with ambiguous or insufficient SNP coverage remain unclassified to avoid introducing noise into the analysis. This genotype-based approach allows for demultiplexing that is independent of experimental barcodes, eliminating potential biases introduced during cell labelling or preparation.

One of the key strengths of scSplit is its ability to differentiate between singlets (cells originating from a single donor) and multiplets (cells containing mixed genetic material from two or more donors). Multiplets are a common artifact in droplet-based sequencing and can significantly affect downstream analyses if not properly identified. scSplit explicitly models multiplets by evaluating combinations of SNP profiles, assigning probabilities for each possible donor pair. This enables the tool to flag and categorize multiplets with high confidence, ensuring the integrity of the single-cell dataset. It is interesting, because we will be able to understand if our multiplets are produced by encapsulating two cells or by containing two mixed labels.

In addition to its precision, scSplit provides robust scalability and adaptability, making it well-suited for extensive multiplexing experiments. It is compatible with various droplet-based sequencing platforms, such as 10x Genomics and BD Rhapsody, and can manage complex experimental designs involving multiple donors. Furthermore, the software's dependence on SNP data allows for seamless integration into existing workflows, minimizing disruption when donor genotypes are accessible. This genotype-driven approach is particularly beneficial for human studies, as the genetic diversity among individuals yields a rich reservoir of information for effective demultiplexing.

Application on our dataset

To apply the SCsplit into our dataset we made use of Demuxafy [82], framework to enhance donor assignment and doublet removal through the consensus intersection of multiple demultiplexing and doublet detecting methods. Demuxafy significantly improves droplet assignment by separating singlets from doublets and classifying the correct individual. In this framework, they suggest different combinations of tools to enable each user to choose and run the demultiplexing and doublet detecting analyses of their choice smoothly and efficiently.

We decided to use Scsplit since we didn't have available information on patients' genotypes, so we had to extract from 3' data. The scSplit tool was utilized in this study to perform demultiplexing of single-cell data based on genotype information. Following the

documentation provided in the scSplit manual, we executed a series of computational steps to assign cells to their corresponding donors and detect potential multiplets. The process began with the preparation of required input files. The key inputs for scSplit are:

1. **Genotype Information for Donors**: This was obtained from external references. Specifically, we used VCF File available at de demuxafy page. The genotype data provided the single-nucleotide polymorphisms (SNPs) used for donor identification.
2. **Single-Cell Gene Expression Data**: The single-cell RNA sequencing (scRNA-seq) data, in the form of a count matrix, was processed to extract SNP information from the sequencing reads. It was the matrices from the WTA assay.
3. **Reference Genome Alignment**: The scRNA-seq reads were aligned to a reference genome, allowing for the identification of SNP loci within the single-cell data. We used the same hg19 reference as in previous steps.

Once these inputs were prepared, scSplit was executed using the following steps:

1. **Preprocessing**: We first processed the scRNA-seq data to filter cells based on quality metrics, ensuring that only high-quality cells with sufficient SNP coverage were included in the analysis. This step also involved aligning sequencing reads to SNP loci to generate a cell-by-SNP matrix.
2. **Donor Matching**: Using the genotype reference file, scSplit performed probabilistic matching of SNP profiles for each cell. For each cell, the algorithm calculated the likelihood of its SNP pattern being derived from each donor. This enabled the assignment of cells to individual donors based on their highest likelihood match.
3. **Multiplet Detection**: scSplit explicitly modelled multiplets by evaluating combinations of SNP profiles that could arise from two or more donors. Cells exhibiting mixed genetic material were flagged as multiplets, allowing for their exclusion from downstream analyses.
4. **Output Generation**: The results of the analysis were output in the form of annotated cell metadata. Each cell was assigned a donor label or flagged as a multiplet or unclassified, depending on the confidence of the SNP matching.

All code included in this chapter will be found in the demultiplexing.sh script in the repository. After performing the whole process and looking at the results for our batches, we were afraid to realise that the performance of an alternative algorithm was not able to improve the original demultiplex even with SNP data (Table 3) when comparing the previous data (Table 2 in Chapter 2).

Sample	Batch 1	Batch 2	Batch 3	Batch 4
Multiplet	365	472	598	1032
SampleTag01_hs	4	637	543	729
SampleTag02_hs	52	692	482	462
SampleTag03_hs	257	509	468	503
SampleTag04_hs	93	3	302	618
SampleTag05_hs	7	175	61	507
SampleTag06_hs	841	1081	482	554
SampleTag07_hs	69	648	1192	287
SampleTag08_hs	682	864	675	421
SampleTag09_hs	1131	72	168	318
SampleTag10_hs	2170	6	28	440
SampleTag11_hs	814	628	1671	710
SampleTag12_hs		258	1613	1782
Undetermined	99	61	181	49

Table 3: *Cell frequencies of each sample contained in each of the batches of the WTA experiment (setA) after QC and with the SCsplit demultiplexing strategy.*

5. Discussion and Limitations

In this study, we employed both scSplit and BD Rhapsody's native demultiplexing pipeline to assign single-cell data to their respective samples in a multiplexed droplet-based sequencing experiment. While scSplit provided a genotype-based computational approach to demultiplexing, its performance was notably less accurate than the BD Rhapsody pipeline. This discrepancy highlights the critical impact of data quality, input requirements, and methodological differences on demultiplexing outcomes.

The weaker performance of scSplit is likely attributable to the lack of direct SNP genotype information for our patient cohort. scSplit relies heavily on high-quality SNP profiles to differentiate between donors with high confidence. In our case, SNP data had to be inferred indirectly from the 3' whole transcriptome assay (WTA) reads. While WTA can capture some SNP information from expressed genes, this approach is inherently limited. The 3' WTA focuses on the ends of transcripts, which restricts SNP detection to regions close to the transcript termini. Additionally, SNP representation in expressed genes may not comprehensively reflect genome-wide variation, leading to incomplete and potentially biased SNP profiles. This limitation likely reduced the accuracy of scSplit's donor assignments, as the algorithm had less genetic information to work with, resulting in lower confidence in singlet detection and a higher rate of ambiguous or incorrect classifications.
In contrast, the BD Rhapsody pipeline takes use of Sample Tag barcodes specifically designed for demultiplexing. These barcodes are experimentally assigned to each sample during cell preparation, providing a direct and robust means of identifying sample origins. This method does not depend on SNP availability or inference and is therefore unaffected by the limitations of transcriptomic SNP detection. The pipeline's use of predefined thresholds for high-quality singlet identification, noise estimation, and multiplet detection further

strengthens its reliability, which would special shine particularly in experiments with well-prepared samples.

Another factor influencing the performance gap is the handling of noise and multiplets. The BD Rhapsody algorithm is optimized to account for noise stemming from Sample Tag contamination or incomplete washing steps, while scSplit's ability to model and subtract noise is limited by the quality and coverage of the inferred SNP data. Furthermore, the BD Rhapsody pipeline's reliance on explicit experimental barcoding makes it less prone to errors introduced by low-coverage or incomplete data, providing a more accurate resolution of singlets and multiplets.

This finding underscores an important consideration for future experiments: the choice of demultiplexing strategy must align with the quality and type of available data. In cases where genotype information is unavailable or limited, approaches like scSplit may struggle to perform effectively, especially in the context of complex or highly multiplexed datasets. On the other hand, experimental barcoding methods, such as BD Rhapsody's Sample Tags, offer a more robust alternative, albeit with the requirement of careful preparation and additional experimental steps.

From a methodological perspective, this comparison highlights the trade-offs between computational and experimental approaches to demultiplexing. Computational methods like scSplit provide flexibility and cost-effectiveness by taking int account existing data but are inherently constrained by the quality and completeness of the input. Experimental approaches, while requiring additional resources and preparation, offer superior accuracy and robustness in situations where computational inference may be compromised.

We also tried other algorithms like freemuxlet with no better results, for which we considered that improving the results on demultiplex would be not possible at least with our current means.

Chapter 4: Immune Receptor Reconstruction from single-cell RNA-seq

1. Introduction

As explained throughout this project, the immune system is a complex network of cells and chemicals responsible for protecting the body from infections and maintaining homeostasis. T cells and B cells mediate the adaptive immune response, which is important to its function. These lymphocytes identify antigens via highly specialized receptors: the T cell receptor (TCR) for T cells and the B cell receptor (BCR) for B cells. The diversity of these receptors, collectively known as the immunological repertoire, allows the immune system to recognize an almost unlimited number of antigens.

Moreover, single-cell RNA sequencing has transformed immunology by allowing for high-resolution analysis of cellular heterogeneity in complex tissues. Unlike bulk RNA sequencing, which provides an averaged signal across a population of cells, scRNA-seq captures each cell transcriptome profile. This skill is especially useful in immunology, where cellular variety is an indicator of function and malfunction.

While it is possible to capture the cell receptors of T and B cells using 5' sequencing and immune profile sequencing techniques, the possibility of recycling other techniques like default 3' scRNA-seq to reconstruct the repertoire of immune cells is bound to enhance the statistical power of the analysis. It also opens a door to wider meta-analysis on the matter, giving extra value to data that was not regarded with this end. Nonetheless, reconstructing the immunological repertoire from scRNA-seq data is difficult due to the technology's inherent complexity and limitations. This kind of data is generally sparse given that only a part of the transcriptome is recorded from each individual cell. This sparsity, along with biological variability and technical noise, is a factor that may impede the recovery of full receptor sequences. Also, it must be considered that the VDJ region actively contribute to receptor diversity and, thus, have a high sequence similarity, making them challenging to precisely assemble and annotate.

Also, another layer of difficulty to the reconstruction is added by the complexity of determining the appropriate pairing of receptor chains. In T cells context, this process entails the reconstruction of the α and β chains of the TCR whereas for B cells, the heavy and light chains of the BCR. Conventional single-cell RNA sequencing protocols frequently fail to maintain this pairing information, Most widely used scRNA-seq platforms, like 10x Genomics, rely on droplet-based microfluidics to encapsulate single cells with barcoded beads. These platforms capture the mRNA from individual cells and attach unique molecular identifiers (UMIs) and barcodes for each cell. While this works well for transcriptomic profiling, the challenge arises because TCR and BCR chains are transcribed from separate loci in the genome. This separation means the mRNAs for the alpha and beta chains of a TCR or the

heavy and light chains of a BCR are not physically linked and are sequenced independently, making it impossible to determine their pairing directly. Additionally, conventional scRNA-seq generates short-read sequences to cover only fragments of mRNA. As a result, assembling full-length receptor sequences requires inference, and it becomes particularly hard to associate paired chains without physical linkage.

If that was not enough of a challenge, scRNA-seq focuses on polyadenylated mRNA, but the receptor chains may be transcribed at different rates or with different levels of completeness, thus, leading to uneven coverage, compounding the difficulty of reconstructing paired receptors.

2. Immune Repertoire Reconstruction State-of-the-Art

Having mentioned the importance of immune repertoire reconstruction from common RNA-seq data and the challenges that arise for immune repertoire reconstruction, in this section it will be presented current tools that try to overcome this challenge of reconstructing the immune repertoire from transcriptomic data.

2.1 TRUST4

The reconstruction of immune repertoires from transcriptomic data is a critical step in understanding adaptive immune responses, particularly in studies where specialized V(D)J sequencing is not available. TRUST4 has emerged as one of the most advanced tools in this domain, offering a powerful, flexible, and accurate method for reconstructing immune receptors from RNA-seq data. Its innovative design and performance make it an asset for analysing immune repertoires in both bulk and scRNA-sequencing contexts, but it is important to clarify and expand upon its capabilities while situating it within the broader landscape of tools in this field. In this project it is the tool that we decided to apply.

TRUST4 represents a significant advancement due to its hybrid approach, combining reference-guided assembly with de novo assembly. While TRUST4 benefits from existing reference databases, such as those from IMGT (International ImMunoGeneTics), it is not strictly dependent on them. This flexibility allows TRUST4 to reconstruct novel or highly divergent T-cell receptor (TCR) and B-cell receptor (BCR) sequences that might otherwise be missed by alignment-based methods. This capability is particularly critical for studying non-model organisms or identifying rare, previously uncharacterized immune receptor variants. TRUST4 ability to assemble full-length receptor sequences, even including the highly variable complementarity-determining region 3 (CDR3), ensures a comprehensive view of the immune repertoire, as CDR3 is the key determinant of antigen specificity.

Unlike many tools limited to specific sequencing platforms, TRUST4 demonstrates broad compatibility with various technologies, including 10x Genomics, Smart-seq, and bulk RNA-seq. This flexibility makes it adaptable for diverse datasets and experimental designs. TRUST4

is particularly useful in this project due to its ability to process data generated from BD Rhapsody, a platform not directly supported by most existing tools. Its robustness in handling sparse and fragmented scRNA-seq data, where coverage of immune receptor transcripts can be highly variable, makes TRUST4 a reliable choice for reconstructing immune repertoires even in challenging datasets with low input quality or cell numbers.

However, some adaptations are needed to be made to use less used methods like BD Rhapsody. In this project, TRUST4 plays a pivotal role in overcoming challenges related to the low number of immune cells retrieved. By reconstructing TCR and BCR sequences from droplet-based sequencing data, it will enable to increase the statistical power of our already existing immunoprofiling experiment by adding more cells into the analysis. TRUST4 ability to work with scRNA-seq data ensures that even in cases of sparse receptor coverage, the immune repertoire can still be meaningfully characterized.

2.2 Other Tools

While TRUST4 stands out for its hybrid assembly approach and flexibility, other tools have also made significant contributions to the field of immune repertoire reconstruction from RNA-seq data. For example, BraCeR [83] and TraCeR [84] are widely used tools designed for reconstructing BCR and TCR sequences, respectively, from single-cell RNA-seq data. Both rely on reference-guided approaches and excel at linking reconstructed receptor sequences to the transcriptomic profiles of individual cells. However, their reliance on references makes them less capable of detecting novel or divergent sequences compared to TRUST4.

MiXCR [85] is another popular tool for bulk and scRNA-seq data that employs a highly efficient alignment-based method for reconstructing TCR and BCR sequences. While MiXCR is faster than most tools and highly accurate when reference sequences are well-curated, it struggles in cases where receptor sequences deviate significantly from the known database.

IgBlast [86] and IMSEQ [87] are additional alignment-based tools commonly used for immune repertoire analysis. While they are excellent for targeted V(D)J sequencing data, their use in transcriptomic datasets is limited due to their strict reliance on predefined references.

3. Applying Trust4

TRUST4 requires as input the alignment of RNA-seq read in BAM files, the file that contains the genomic sequence, the coordinates of the VDJ genes, and the reference database sequence containing the annotation information. In our case we extracted that information from the widely known IMGT database. IMGT [88] is a comprehensive database and bioinformatics resource specializing in immunoglobulins and T cell receptors which provides standardized nomenclature, classification, and a suite of tools for analysing these sort molecules of the immune system. As an alternative input, TRUST4 also accepts files in sequence files (either fasta or fastq) in paired-end format or single end format.

At first instance, the approach chosen was to run the software with the alignment file. An alignment file contains information comparing a large number of sequences, which can be composed by DNA, RNA, or amino acids. These files are used to allow to analyse how distinct sequences compare or match to one another, which is frequently done as part of sequence comparison or similarity searches. In the context of single cell sequencing, it is mandatory to generate alignment files, since it helps to map these raw sequencing reads to the appropriate locations in the genome, making it possible to identify which genes or regions are being expressed or sequenced in each individual cell.

To obtain the alignment file. We picked up the intermediate stage .bam files generated with the Seven Bridges software provided by BD Rhapsody as explained in Chapter 1 of this document.

Having the alignment file, TRUST4 is ready to run. Upon identifying the barcode within the entries stored in the .bam files, TRUST4 only assembles the reads with the same barcode together. For 10X Genomics data, usually the input is the BAM file from cell-ranger, and you can use "--barcode" to specify the field in the BAM file to specify the barcode: e.g. "--barcode CB". The sequences used in this thesis were not from 10x, but from BD. BD alignment files, while similar format to 10x, store header information differently. BD generated an index calculated from the combination of nucleotides of the barcode captured instead of the real barcode present in the sequence. Additionally, the position of the information and format of such is different to 10x. These natural differences between technology providers returned unsatisfactory results in the first applications of the software under default configurations. Less than 10% of the cells were able to be reconstructed due to inability to properly identify the cells.

To try to tackle this issue TRUST4 was ran again specifying the format of the barcode in the header with the parameter –barcode. Specifically, it was indicated that the barcode was captured in the format CB:Z:$BC where $BC stands for the index that bd generates from the barcodes. However, possibly due to the difference in position of the header of the bam entries the software retrieved a similar interspecific result. Since it was possible that the problem could be related to TRUST4 not finding the barcode specified in the header within the sequence, a new approach was regarded to tackle the problem. As previously mentioned, BD used indices that are calculated from the barcode sequence. In other words, it gives a kind of code name to the barcode to make the naming more efficient. In this approach the idea was to substitute the value in $BC by the actual barcode, to do so, it was needed to backtrack manually the process done by BD to obtain the expected barcode. The code for this process is stored in function index_to_sequence() in the repository. Firstly, is important to remark the structure of

Given the problems with alignment files, it was decided to try to improve the performance of the software by using the de novo approach and use fastq files as input. As previously mentioned, the fastq files used in this project are paired-end meaning that the r1 file contains

the information regarding the barcodes. These barcodes are original V1 beads from BD. The bead is composed by 3 CLS (composed label sequences) and the UMI. In total the barcode goes from position 0 to 52, and the capture sequence corresponds from nucleotide 60 onwards. Therefore, by taking advantage of this knowledge it was possible to retrieve a positive output from TRUST4 which in this case was capable of reconstruct a decent number of cells.

However, another problem arises with this approach. BD Rhapsody, as previously explained, translates its barcodes into indices based on the sequence of the barcode to optimize the process. If we recall, each bead is composed by 3 CLs along each R1 read. Two common sequences (L1, L2) separate the three CLSs, and the presence of L1 and L2 relates to the way the capture oligo nucleotide probes on the beads are constructed. By design, each CLS has one of either 96 or 384 predefined sequences (depending on bead version), which has a Hamming distance of at least four bases and an edit distance of at least two bases apart. A cell label is defined by the unique combination of predefined sequences in the three CLSs. Thus, the maximum possible number of cell labels is either 963 or 3843. In the final data tables, the three-part cell label is converted to a single integer index between 1-3843.

With this information, we need to replicate their process to translate the real barcodes stored in the results of BD rhapsody. For that, a self-made code was included in our framework e. The code is included in the GitHub repository as barcode_translate.py.

3.1 Script Methodology

This section describes the development and implementation of a Python-based solution designed to translate barcodes obtained from TRUST4 V(D)J reconstruction into well-defined indices for downstream analysis. The task was particularly challenging due to the biological nature of RNA sequencing, where nucleotide mismatches, such as insertions, deletions, and substitutions, can result from sequencing errors. These issues directly affect the integrity of the cell labels embedded in barcodes. To address this, a custom logic was implemented to enable error-tolerant matching, ensuring robust and accurate reconstruction of cell indices.

The goal of this process was to match cell-specific barcode sequences to predefined reference keys, even in the presence of sequencing artifacts. This translation is essential for mapping reconstructed immune receptor data back to the specific cells from which they originated, enabling a seamless integration of V(D)J and Trust4 inferred data. The solution accounted for common sequencing issues by implementing similarity-based matching and logical handling of imperfect barcode sequences.

1. Barcode Quality assessment and matching Logic: The script begins by parsing the output of TRUST4, specifically the file containing reconstructed receptor sequences and their associated barcodes. Each barcode consists of several sections, and the quality of the sequencing is assessed by comparing key portions of the barcode (linker regions) against predefined reference sequences. A helper function similarity_percentage() is used to

calculate the similarity between barcode segments and their respective reference sequences using a sequence-matching algorithm. This approach allows for tolerance of small sequencing errors by computing a similarity ratio. Barcodes with high similarity scores (≥90%) are classified as good quality, while others are flagged as low quality or irreconstructible. Perfect matches (100% similarity) are separately tracked.

2. Barcode Decomposition and Logical Matching: Each barcode was decomposed into three sections (cls1, cls2, cls3) corresponding to different parts of the cell label. These sections were matched against predefined reference keys (A96_cell_key1, A96_cell_key2, A96_cell_key3) that define the expected nucleotide sequences for each position. For cases where an exact match could not be found due to sequencing errors, the script employed an error-tolerant matching logic using the assing_label function. This function iteratively compared the input sequence with all reference keys, selecting the most similar reference based on the highest similarity percentage.

3. Reconstruction of Cell Indices: If all three sections of a barcode (`cls1`, `cls2`, `cls3`) were successfully matched to their respective references, the indices of the matched references were combined into a unique identifier using the label_sections_to_index() function. This reconstructed index represented the cell final identifier, allowing for accurate tracking and integration of its V(D)J immune receptor data.

4. Quality assessment and Metrics: The script tracked barcode quality at multiple levels. Barcodes were categorized as:

- **Perfect Matches**: Barcodes with 100% similarity to the reference sequences in all regions.
- **Good Quality Matches**: Barcodes with ≥90% similarity but not perfect matches.
- **Irreconcilable Barcodes**: Barcodes with <90% similarity in one or more sections.

Metrics such as the fraction of barcodes successfully reconstructed, the fraction of irreconcilable barcodes, and the number of cells with perfectly translatable barcodes were computed and reported. These metrics provided a comprehensive overview of the data quality and the efficiency of the reconstruction process. Only the reconstructed cells were perfect and good quality matches were stored as for reconstruction.

4. Results

After performing TRUST4 and the barcode translation in all the batches, these were the final outputs retrieved (Table 5)

Metric	Batch 1	Batch 2	Batch 3	Batch 4
Fraction of good quality with matching CL1	0.879	0.881	0.880	0.875
Fraction of good quality with matching CL2	0.578	0.576	0.577	0.572
Fraction of good quality with matching CL3	0.400	0.402	0.401	0.396
Number of Cells with easy translation	495	740	811	1108
Fraction over Good Quality Cells with easy translation	0.331	0.328	0.330	0.325
Fraction over All Cells with easy translation	0.240	0.241	0.241	0.238
Fraction of Low-quality Linker	0.270	0.269	0.271	0.274
Fraction of Good quality Linker	0.453	0.455	0.454	0.450
Fraction of Perfect quality Linker	0.277	0.276	0.275	0.276
Number of reconstructed cells	382	652	708	926
Total number of output cells by Trust4	2,055	3,113	3,371	4,667
Total number of cells WTA	13,484	20,193	21,935	31,025
Total number of cells WTA After filtering	6511	5974	6968	7345

Table 5. *Comparison of barcode reconstruction and translation metrics across four batches of single-cell data during V(D)J immune repertoire analysis from whole transcriptome sequencing (WTA).* *Metrics include the quality of barcode matching, the number of cells successfully translated, and the fraction of reconstructed cells. Proportions are consistent across batches, demonstrating the reproducibility of the approach despite variations in total cell input.*

5. Discussion and Limitations

The reconstruction of V(D)J immune receptor sequences from WTA data presents unique challenges, particularly when dealing with barcodes affected by sequencing noise and RNA-specific errors like nucleotide insertions and deletions. Across all batches, the results indicate consistent performance in barcode matching and reconstruction metrics, suggesting the robustness of the implemented workflow.

The fraction of good quality matching for CL1 remained high (0.875–0.881), underscoring the reliability of this label segment during reconstruction. In contrast, CL2 and CL3 exhibited lower matching rates (0.572–0.578 and 0.396–0.402, respectively), reflecting the increased complexity of these regions and their susceptibility to sequencing errors. Despite these

differences, the proportions remained stable across batches, indicating minimal variability attributable to sequencing platform performance or batch-specific conditions.

Figure 18. *Decision Tree describing the path of filtering in the data used for reconstruction in the TRUST4 process. Data goes from the total number of WTA cells inferable to the final number of cells reconstructed.*

The number of cells with easy translation scales proportionally with the total cell count per batch, as does the number of reconstructed cells. Batch 4, having the largest cell input (31,025 cells), yielded the highest number of successfully translated and reconstructed cells (1,108 and 926, respectively). However, the proportion of cells with easy translation relative to all cells (0.238–0.241) and good quality cells (0.325–0.331) remained consistent, demonstrating that the workflow maintains performance across datasets of varying sizes. It is interesting to note that, while Batch4 had almost 3 times more cells than batch3 the number of

reconstructed cells was not especially higher. It is possibly due to the large amount of multiplets found in that dataset, therefore further pointing out the importance of technical problems in the dataset.

Linker quality metrics revealed that roughly 27% of barcodes exhibited low-quality linkers, while the remaining 73% were classified as good or perfect quality. These values are in line with the expected performance of whole transcriptome sequencing data when applied to immune repertoire reconstruction, where noise in barcode capture is a known limitation.

The total number of reconstructed cells per batch was a fraction of the total input cells, highlighting the challenges of recovering full V(D)J sequences from transcriptomic data. Nevertheless, the consistent proportions of reconstructed cells across batches validate the reliability of the TRUST4-based workflow. The similarity-based matching logic and error tolerance implemented in this study were critical for mitigating the effects of barcode mismatches and recovering meaningful immune repertoire data.

However, it is true that the total number of cells that were able to reconstruct was low compared to the total amount that we could have reconstructed. In the end, of the fraction of cells reconstructed in function of the total number of cells stored in the sequence files ranged from 3% to 7%. It is true, however, that TRUST4 was capable of reconstructing ranging from 12% to about 15% of the total available cells and that many are lost in the barcode to index translation process. The statistical power gained in the filtered cells ranged from 5% to 10% as well.

Therefore, while the method performance shows the robustness and reliability of the implemented workflow, the inherent challenges associated with sequencing noise and RNA-specific errors were not able to be overcome as far as to gain enough statistical power to perform a proper analysis with this data.

Chapter 5: General Discussion and Limitations

For this project, we propose an innovative framework that enhances the statistical power of single cell VDJ immunoprofiling focused on reconstructing immune receptor repertoires from whole transcriptome RNA-sequencing data and optimizing demultiplexing workflows. All in the context of BD Rhapsody. The results reflect the potential and limitations of these methodologies when applied to complex single-cell datasets and highlight key areas for improvement in experimental and computational workflows.

One significant limitation of the project was the platform-specific constraints of BD Rhapsody. While many tools, such as TRUST4 and other computational pipelines, are optimized for 10x Genomics data, BD Rhapsody often lacks direct compatibility with these widely used resources. Even if many tools sell their product or software as being able to adapt to many situations, the reality is quite tricky at the time of applying it on real datasets. For example, TRUST4, which demonstrated robustness in reconstructing V(D)J immune receptor sequences, faced compatibility issues with BD Rhapsody's barcode and BAM file formats. These incompatibilities required additional preprocessing steps, as we needed to translate BD barcodes into indices, which introduced an additional layer of complexity and potential data loss. Even to specifically allocate the barcodes we had to understand the barcode to allow TRUST4 to identify the sequences properly. Future work could explore alternative approaches to this step, such as alignment-based methods like those used in sequence matching, which may offer improved efficiency and accuracy in translating barcodes to indices after TRUST4 processing.

Furthermore, BD Rhapsody's reliance on short 3' sequencing reads presented challenges for V(D)J reconstruction, as immune receptor regions often require full-length or 5' reads to capture the diversity of complementarity-determining regions (CDR3). This technical limitation likely reduced the accuracy and statistical power of the reconstructed repertoire, suggesting that adopting a 5' sequencing approach could significantly improve results. TRUST4 paper [89] also highlights that the performance of the software lacks when presented against 3' sequencing datasets, as we could see through the results of the Reconstruction chapter. Implicitly they state that: "Besides the 5' scRNA-seq data, we also evaluated TRUST4 on the 3' 10x Genomics PBMC data, with only 335 cells having reconstructed CDR3s".

Adopting 5' sequencing in future studies could significantly enhance the recovery of full V(D)J regions, improve diversity metrics, and allow for more detailed immune repertoire analysis. While 3' sequencing has its advantages in scalability and cost-efficiency, the trade-offs in resolution and completeness make 5' sequencing a more suitable approach for such analyses. However, it would remain a challenge to infer VDJ data from 3' scRNAseq.

Another critical issue encountered was the high prevalence of multiplets in the BD Rhapsody datasets, which likely stemmed from technical errors during cell preparation or droplet encapsulation. Multiplets inflate cell counts and complicate both demultiplexing and

downstream analyses by creating hybrid profiles that do not represent single-cell data. This issue was particularly pronounced in Batch 4, which had the largest input dataset and the highest number of flagged multiplets. Addressing these experimental challenges is essential for improving data quality, and future studies should prioritize optimizing droplet-based workflows to reduce multiplet rates. Enhanced protocols for sample handling, droplet generation, and cell washing could mitigate these artifacts and yield cleaner datasets for downstream analyses.

The demultiplexing framework also revealed critical insights into the trade-offs between computational and experimental approaches. While scSplit demonstrated the potential of genotype-based demultiplexing, its performance was constrained by the lack of direct genotyping and the sparse SNP coverage inherent to 3' WTA data. The BD Rhapsody Sample Tag-based pipeline proved more reliable, offering robust sample assignments that were unaffected by SNP data quality. However, scSplit showcased the flexibility and cost-effectiveness of computational methods, particularly when high-quality and personalized SNP data are available. Future work could explore hybrid demultiplexing strategies, combining experimental barcoding with computational inference to maximize accuracy and adaptability. For instance, integrating genotype-informed methods with enhanced SNP inference algorithms could improve sample origin determination, even in the absence of direct genotyping data.

The V(D)J reconstruction framework developed in this project proved to be robust and consistent across batches, despite the inherent limitations of 3' sequencing. TRUST4 successfully reconstructed immune repertoires from 12–15% of cells in the sequence files, with further data loss occurring during barcode-to-index translation. Ultimately, only 3–7% of total input cells were reconstructed, underscoring the challenges of working with short-read WTA data. Nevertheless, the pipeline's ability to consistently perform across datasets of varying sizes demonstrates its reliability as a tool for immune repertoire analysis. To improve these outcomes, future work could focus on developing algorithms specifically tailored to handle the sparse and fragmented nature of 3' sequencing data, potentially incorporating machine learning to predict and recover missing information in the V(D)J regions. Furthermore, the development of tools that are platform-agnostic or that include direct support for BD Rhapsody data would greatly enhance the inclusivity and applicability of bioinformatics workflows for immune profiling.

Therefore, concluding the project, the results underscore the challenges and opportunities of taking into consideration WTA data for immune profiling. While BD Rhapsody offers unique capabilities, such as Sample Tag-based demultiplexing, its limitations—particularly regarding barcode complexity, multiplets, and short-read sequencing—require careful consideration. Future studies could benefit from adopting 5' sequencing technologies for more comprehensive V(D)J reconstruction and expanding the compatibility of tools like TRUST4 to work seamlessly with BD Rhapsody formats. Additionally, addressing technical errors in sample handling and exploring hybrid approaches that combine experimental barcoding with

computational inference could further enhance the quality and reliability of the data. Despite these challenges, the workflows developed in this project demonstrate the feasibility of immune repertoire reconstruction from WTA data and provide a foundation for future advancements in single-cell immune profiling

Chapter 6: Key notes on ethical considerations, sustainability and diversity

The ethical framework of this project emphasizes the protection of participant privacy, informed consent, and compliance with international ethical and legal standards. By adhering to protocols approved by institutional review boards (IRBs) and following the Declaration of Helsinki, the study ensures that participants' rights and dignity are upheld at every stage. The use of anonymized and de-identified data further safeguards participant privacy, addressing the ethical responsibility of minimizing risks associated with sensitive personal information. The stringent adherence to GDPR and U.S. privacy regulations demonstrates a robust commitment to cross-jurisdictional ethical compliance.

Socially, the study contributes to the broader understanding of immune responses, particularly in the context of diseases such as COVID-19, which disproportionately affect marginalized populations. By including participants from diverse racial and demographic backgrounds, the research promotes equity and inclusivity in scientific discovery. This diversity is essential for ensuring that findings are generalizable and not biased toward specific populations, thereby addressing disparities in healthcare and fostering more equitable therapeutic strategies.

The implications of this project extend beyond academic research to societal benefits, as it provides foundational knowledge for improving diagnostics, vaccines, and therapies targeting immune-related diseases. The focus on V(D)J reconstruction from whole transcriptome sequencing opens avenues for cost-effective immune profiling, potentially making these technologies more accessible in resource-limited settings. However, the study also highlights the challenges of working with non-standard sequencing platforms like BD Rhapsody, underscoring the need for greater inclusivity in the development of bioinformatics tools to accommodate diverse datasets and technologies.

Sustainability

The project aligns with sustainable research practices by maximizing the use of existing datasets and computational resources. Instead of generating new sequencing data, the study takes advantages of a previously generated dataset, reducing the environmental impact associated with sample collection, sequencing, and reagent use. This approach minimizes waste and energy consumption, reflecting a commitment to resource efficiency. Additionally, the computational framework developed in this study is designed for scalability and adaptability, ensuring that it can be reused and expanded upon in future projects. By addressing limitations in current workflows, such as compatibility with BD Rhapsody data, the project contributes to the long-term sustainability of bioinformatics by creating tools that are versatile and robust across different experimental setups. Furthermore, the framework itself aims to allow the reutilization of already existing tools, being key for sustainable praxis.

However, the heavy reliance on computational resources, including secure remote servers, highlights the importance of energy-efficient data storage and analysis solutions. Future considerations could include incorporating cloud platforms with low carbon footprints or using energy-efficient algorithms to further align with sustainability goals.

Diversity and Inclusion

This project demonstrates a clear commitment to diversity in its participant pool, incorporating individuals from various racial and demographic backgrounds. This inclusivity ensures that the findings are applicable to a wide range of populations, addressing the historical underrepresentation of minority groups in biomedical research. Such representation is crucial for understanding variations in immune responses and personalising medical interventions to diverse patient populations.

The study also underscores the importance of diversity in scientific collaboration, involving researchers from different geographic and institutional backgrounds. By fostering international collaboration between European and American institutions, the project highlights the value of cross-cultural exchange in advancing scientific innovation.

Finally, the development of computational tools that adapt to non-standard sequencing platforms like BD Rhapsody indirectly promotes inclusivity by expanding the accessibility of advanced bioinformatics workflows. Supporting diverse experimental platforms ensures that smaller laboratories or those using fewer common technologies are not excluded from advancements in single-cell analysis.

References

1. Lai, C. C., Shih, T. P., Ko, W. C., Tang, H. J. & Hsueh, P. R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *Int J Antimicrob Agents* **55**, (2020).
2. Cucinotta, D. & Vanelli, M. WHO Declares COVID-19 a Pandemic. *Acta Biomed* **91**, 157–160 (2020).
3. Krammer, F. SARS-CoV-2 vaccines in development. *Nature 2020 586:7830* **586**, 516–527 (2020).
4. Symptoms of COVID-19 | CDC. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.
5. Clinical Spectrum | COVID-19 Treatment Guidelines. https://www.covid19treatmentguidelines.nih.gov/overview/clinical-spectrum/.
6. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology 2020 19:3* **19**, 155–170 (2020).
7. Pustake, M., Tambolkar, I., Giri, P. & Gandhi, C. SARS, MERS and CoVID-19: An overview and comparison of clinical, laboratory and radiological features. *J Family Med Prim Care* **11**, 10 (2022).
8. Li, T. *et al.* Phylogenetic supertree reveals detailed evolution of SARS-CoV-2. *Scientific Reports 2020 10:1* **10**, 1–9 (2020).
9. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).
10. Wang, M. Y. *et al.* SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. *Front Cell Infect Microbiol* **10**, 587269 (2020).
11. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nature Reviews Molecular Cell Biology 2021 23:1* **23**, 3–20 (2021).
12. Huang, Y., Yang, C., Xu, X. feng, Xu, W. & Liu, S. wen. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica 2020 41:9* **41**, 1141–1149 (2020).
13. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology 2021 19:7* **19**, 409–424 (2021).
14. Carabelli, A. M. *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology 2023 21:3* **21**, 162–177 (2023).
15. Radvak, P. *et al.* SARS-CoV-2 B.1.1.7 (alpha) and B.1.351 (beta) variants induce pathogenic patterns in K18-hACE2 transgenic mice distinct from early strains. *Nature Communications 2021 12:1* **12**, 1–15 (2021).
16. Nonaka, C. K. V. *et al.* SARS-CoV-2 variant of concern P.1 (Gamma) infection in young and middle-aged patients admitted to the intensive care units of a single hospital in Salvador, Northeast Brazil, February 2021. *International Journal of Infectious Diseases* **111**, 47 (2021).
17. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature 2021 599:7883* **599**, 114–119 (2021).
18. Liu, Y. *et al.* The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. *Nature 2021 602:7896* **602**, 294–299 (2021).

19. Jangra, S. *et al.* SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. *Lancet Microbe* **2**, e283–e284 (2021).
20. Cherian, S. *et al.* SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. *Microorganisms* **9**, (2021).
21. Montazersaheb, S. *et al.* COVID-19 infection: an overview on cytokine storm and related interventions. *Virology Journal 2022 19:1* **19**, 1–15 (2022).
22. Eberhardt, N. *et al.* SARS-CoV-2 infection triggers pro-atherogenic inflammatory responses in human coronary vessels. *Nature Cardiovascular Research 2023 2:10* **2**, 899–916 (2023).
23. Fairweather, D. L. *et al.* COVID-19, Myocarditis and Pericarditis. *Circ Res* **132**, 1302–1319 (2023).
24. Zheng, Y. Y., Ma, Y. T., Zhang, J. Y. & Xie, X. COVID-19 and the cardiovascular system. *Nature Reviews Cardiology 2020 17:5* **17**, 259–260 (2020).
25. Inciardi, R. M., Solomon, S. D., Ridker, P. M. & Metra, M. Coronavirus 2019 disease (Covid-19), systemic inflammation, and cardiovascular disease. *J Am Heart Assoc* **9**, (2020).
26. Diamond, M. S. & Kanneganti, T. D. Innate immunity: the first line of defense against SARS-CoV-2. *Nature Immunology 2022 23:2* **23**, 165–176 (2022).
27. Meidaninikjeh, S. *et al.* Monocytes and macrophages in COVID-19: Friends and foes. *Life Sci* **269**, 119010 (2021).
28. Hirayama, D., Iida, T. & Nakase, H. The Phagocytic Function of Macrophage-Enforcing Innate Immunity and Tissue Homeostasis. *Int J Mol Sci* **19**, (2018).
29. Duque, G. A. & Descoteaux, A. Macrophage cytokines: Involvement in immunity and infectious diseases. *Front Immunol* **5**, 117833 (2014).
30. Marshall, J. S., Warrington, R., Watson, W. & Kim, H. L. An introduction to immunology and immunopathology. *Allergy, Asthma and Clinical Immunology* **14**, 1–10 (2018).
31. Wang, Z., Li, S. & Huang, B. Alveolar macrophages: Achilles' heel of SARS-CoV-2 infection. *Signal Transduction and Targeted Therapy 2022 7:1* **7**, 1–9 (2022).
32. Knoll, R., Schultze, J. L. & Schulte-Schrepping, J. Monocytes and Macrophages in COVID-19. *Front Immunol* **12**, 720109 (2021).
33. Wendisch, D. *et al.* SARS-CoV-2 infection triggers profibrotic macrophage responses and lung fibrosis. (2021) doi:10.1016/j.cell.2021.11.033.
34. Mandal, A. & Viswanathan, C. Natural killer cells: In health and disease. *Hematol Oncol Stem Cell Ther* **8**, 47–55 (2015).
35. Abel, A. M., Yang, C., Thakar, M. S. & Malarkannan, S. Natural Killer Cells: Development, Maturation, and Clinical Utilization. *Front Immunol* **9**, 1 (2018).
36. Patente, T. A. *et al.* Human dendritic cells: Their heterogeneity and clinical application potential in cancer immunotherapy. *Front Immunol* **10**, 422571 (2019).
37. Mellman, I. & Steinman, R. M. Dendritic cells: Specialized and regulated antigen processing machines. *Cell* **106**, 255–258 (2001).
38. Wculek, S. K. *et al.* Dendritic cells in cancer immunology and immunotherapy. *Nature Reviews Immunology 2019 20:1* **20**, 7–24 (2019).
39. Zhou, R. *et al.* Acute SARS-CoV-2 Infection Impairs Dendritic Cell and T Cell Responses. *Immunity* **53**, 864-877.e5 (2020).
40. Ho Lim, C. & Ito, M. Monocyte and dendritic cell defects in COVID-19. *Nature Cell Biology 2021 23:5* **23**, 445–447 (2021).
41. Chaplin, D. D. Overview of the Immune Response. *J Allergy Clin Immunol* **125**, S3 (2010).

42. Bonilla, F. A. & Oettgen, H. C. Adaptive immunity. *Journal of Allergy and Clinical Immunology* **125**, S33–S40.
43. Chen, S. *et al.* The role of B cells in COVID-19 infection and vaccination. *Front Immunol* **13**, 988536 (2022).
44. Akkaya, M., Kwak, K. & Pierce, S. K. B cell memory: building two walls of protection against pathogens. *Nature Reviews Immunology 2019 20:4* **20**, 229–238 (2019).
45. Goel, R. R. *et al.* Distinct antibody and memory B cell responses in SARSCoV-2 naïve and recovered individuals following mRNA vaccination. *Sci Immunol* **6**, 1–19 (2021).
46. Hoffman, W., Lakkis, F. G. & Chalasani, G. B Cells, Antibodies, and More. *Clin J Am Soc Nephrol* **11**, 137 (2016).
47. Luckheeram, R. V., Zhou, R., Verma, A. D. & Xia, B. CD4+T Cells: Differentiation and Functions. *Clin Dev Immunol* **2012**, 12 (2012).
48. Crotty, S. T follicular helper cell differentiation, function, and roles in disease. *Immunity* **41**, 529 (2014).
49. Berger, A. Science commentary: Th1 and Th2 responses: what are they? *BMJ : British Medical Journal* **321**, 424 (2000).
50. Gil-Etayo, F. J. *et al.* T-Helper Cell Subset Response Is a Determining Factor in COVID-19 Progression. *Front Cell Infect Microbiol* **11**, 624483 (2021).
51. Corthay, A. How do Regulatory T Cells Work? *Scand J Immunol* **70**, 326 (2009).
52. Zhang, N. & Bevan, M. J. CD8+ T Cells: Foot Soldiers of the Immune System. *Immunity* **35**, 161 (2011).
53. Raskov, H., Orhan, A., Christensen, J. P. & Gögenur, I. Cytotoxic CD8+ T cells in cancer and cancer immunotherapy. *British Journal of Cancer 2020 124:2* **124**, 359–367 (2020).
54. Minervina, A. A. *et al.* SARS-CoV-2 antigen exposure history shapes phenotypes and specificity of memory CD8+ T cells. *Nature Immunology 2022 23:5* **23**, 781–790 (2022).
55. Lorente, L. *et al.* HLA genetic polymorphisms and prognosis of patients with COVID-19. *Med Intensiva* **45**, 96 (2021).
56. Ouedraogo, A. R. *et al.* Association of HLA-DRB1*11 and HLA-DRB1*12 gene polymorphism with COVID-19 in Burkina Faso. *BMC Med Genomics* **16**, 1–7 (2023).
57. Jin, P. & Wang, E. Polymorphism in clinical immunology - From HLA typing to immunogenetic profiling. *J Transl Med* **1**, 1–11 (2003).
58. Williams, T. M. Human Leukocyte Antigen Gene Polymorphism and the Histocompatibility Laboratory. *J Mol Diagn* **3**, 98 (2001).
59. Hennecke, J. & Wiley, D. C. T cell receptor-MHC interactions up close. *Cell* **104**, 1–4 (2001).
60. Tourigny, M. R., Mazel, S., Burtrum, D. B. & Petrie, H. T. T Cell Receptor (TCR)-β Gene Recombination: Dissociation from Cell Cycle Regulation and Developmental Progression During T Cell Ontogeny. *J Exp Med* **185**, 1549 (1997).
61. Snyder, T. M. *et al.* Magnitude and Dynamics of the T-Cell Response to SARS-CoV-2 Infection at Both Individual and Population Levels. *medRxiv* (2020) doi:10.1101/2020.07.31.20165647.
62. Gutierrez, L., Beckford, J. & Alachkar, H. Deciphering the TCR Repertoire to Solve the COVID-19 Mystery. *Trends Pharmacol Sci* **41**, 518–530 (2020).
63. Wang, Y. *et al.* Analysis of TCR Repertoire by High-Throughput Sequencing Indicates the Feature of T Cell Immune Response after SARS-CoV-2 Infection. *Cells* **11**, (2021).
64. Hozumi, N. & Tonegawa, S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc Natl Acad Sci U S A* **73**, 3628–3632 (1976).
65. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302**, 575–581 (1983).

66. Wen, W. *et al.* Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. *Cell Discovery 2020 6:1* **6**, 1–18 (2020).
67. Blood ATlas Consortium, M. *et al.* Authors COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium Correspondence In brief ll Resource A blood atlas of COVID-19 defines hallmarks of disease severity and specificity COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. *Cell* **185**, 916–938 (2022).
68. GitHub - s-andrews/FastQC: A quality control analysis tool for high throughput sequencing data. https://github.com/s-andrews/FastQC.
69. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
70. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
71. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491–499 (2017).
72. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).
73. Wickham, H. *et al.* Welcome to the Tidyverse. *J Open Source Softw* **4**, 1686 (2019).
74. 7. Normalization — Single-cell best practices. https://www.sc-best-practices.org/preprocessing_visualization/normalization.html.
75. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nat Methods* **20**, 665–672 (2023).
76. Greenacre, M. *et al.* Principal component analysis. *Nature Reviews Methods Primers 2022 2:1* **2**, 1–21 (2022).
77. Borcherding, N., Bormann, N. L., & Kraus, G. (2020). scRepertoire: An R-based toolkit for single-cell immune receptor analysis. F1000Research, 9. https://doi.org/10.12688/F1000RESEARCH.22139.2/DOI
78. Arunkumar, M., & Zielinski, C. E. (2021). T-Cell Receptor Repertoire Analysis with Computational Tools—An Immunologist's Perspective. Cells, 10(12), 3582. https://doi.org/10.3390/CELLS10123582
79. Shugay, M., Bagaev, D. v., Turchaninova, M. A., Bolotin, D. A., Britanova, O. v., Putintseva, E. v., Pogorelyy, M. v., Nazarov, V. I., Zvyagin, I. v., Kirgizova, V. I., Kirgizov, K. I., Skorobogatova, E. v., & Chudakov, D. M. (2015). VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. PLoS Computational Biology, 11(11). https://doi.org/10.1371/JOURNAL.PCBI.1004503
80. Sturm1, G., Szabo, T., Fotakis, G., Haider1, M., Rieder, D., Trajanoski, Z., & Finotello, F. (2020). Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. Bioinformatics, 36(18), 4817–4818. https://doi.org/10.1093/BIOINFORMATICS/BTAA611
81. Xu, J., Falconer, C., Nguyen, Q., Crawford, J., McKinnon, B. D., Mortlock, S., Senabouth, A., Andersen, S., Chiu, H. S., Jiang, L., Palpant, N. J., Yang, J., Mueller, M. D., Hewitt, A. W., Pébay, A., Montgomery, G. W., Powell, J. E., & Coin, L. J. M. (2019). Genotype-free demultiplexing of pooled single-cell RNA-seq. Genome Biology, 20(1), 1–12. https://doi.org/10.1186/S13059-019-1852-7/FIGURES/4
82. Neavin, D., Senabouth, A., Arora, H., Lee, J. T. H., Ripoll-Cladellas, A., Franke, L., Prabhakar, S., Ye, C. J., McCarthy, D. J., Melé, M., Hemberg, M., & Powell, J. E. (2024). Demuxafy: i mprovement in droplet assignment by integrating multiple single-cell demultiplexing and

doublet detection methods. Genome Biology, 25(1), 1–24. https://doi.org/10.1186/S13059-024-03224-8/FIGURES/5

83. Lindeman, I., Emerton, G., Mamanova, L., Snir, O., Polanski, K., Qiao, S. W., Sollid, L. M., Teichmann, S. A., & Stubbington, M. J. T. (2018). BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. Nature Methods 2018 15:8, 15(8), 563–565. https://doi.org/10.1038/s41592-018-0082-3.
84. Stubbington, M. J. T., Lönnberg, T., Proserpio, V., Clare, S., Speak, A. O., Dougan, G., & Teichmann, S. A. (2016). T cell fate and clonality inference from single-cell transcriptomes. Nature Methods 2016 13:4, 13(4), 329–332. https://doi.org/10.1038/nmeth.3800
a. Bioinformatics, 36(18), 4817–4818. https://doi.org/10.1093/BIOINFORMATICS/BTAA611
85. Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. v., & Chudakov, D. M. (2015). MiXCR: software for comprehensive adaptive immunity profiling. Nature Methods 2015 12:5, 12(5), 380–381. https://doi.org/10.1038/nmeth.3364
86. Ye, J., Ma, N., Madden, T. L., & Ostell, J. M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Research, 41(Web Server issue). https://doi.org/10.1093/NAR/GKT382
87. Kuchenbecker, L., Nienen, M., Hecht, J., Neumann, A. U., Babel, N., Reinert, K., & Robinson, P. N. (2015). IMSEQ—a fast and error aware approach to immunogenetic sequence analysis. Bioinformatics, 31(18), 2963–2971. https://doi.org/10.1093/BIOINFORMATICS/BTV309
88. Lefranc, M. P. (2003). IMGT, the international ImMunoGeneTics database®. Nucleic Acids Research, 31(1), 307. https://doi.org/10.1093/NAR/GKG085.
89. Song, L., Cohen, D., Ouyang, Z., Cao, Y., Hu, X., & Liu, X. S. (2021). TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. Nature Methods 2021 18:6, 18(6), 627–630. https://doi.org/10.1038/s41592-021-01142-2