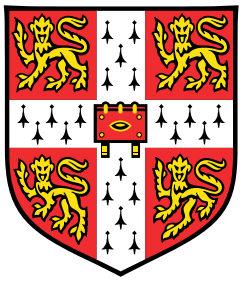
Intratumoral B and T cell receptors: reconstruction and analysis



Meltem Gürel

Cancer Research UK Cambridge Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Darwin College

November 2019

To my grandmother Türkan and my niece Deniz, who have taught me to be curious.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Meltem Gürel
November 2019

Intratumoral B and T cell receptors: reconstruction and analysis

Meltem Gürel

When cells divide, mistakes happen. However, an intricate surveillance system has evolved to detect and eliminate anomalous cells before they become detrimental to the host organism. In cancer, abnormal cells manage to escape the immune system and grow uncontrollably. In this sense, cancer can be considered as an oversight of the immune system, as immune escape is a defining feature of clinically detectable cancers. The role of the immune system in fighting cancer is becoming increasingly indisputable as our understanding of its underlying mechanisms expand owing to technological advances in genomics, cancer biology, and computational sciences. In particular, significant research effort is undertaken in the field of cancer immunotherapy, where the immune system is stimulated to recognize and attack cancerous cells. In this thesis, I investigate certain aspects of the immune system in the context of cancer by computationally reconstructing and analyzing intratumoral B and T cell receptors.

Applying a novel immune cell receptor profiling protocol to original single-cell RNA sequencing (scRNA-seq) data obtained from melanoma patients, I present a complete computational reconstruction of intratumoral immune receptors in this cancer type. The scRNA-seq results are consistent with the presence of an ongoing intratumoral immune response, likely involving tertiary lymphoid structures and the cooperation between B and T cells.

Additionally, using a dataset of paired tumor biopsies collected pre- and post-treatment, I show that B cell infiltration increases after immunotherapy in pancreatic and colorectal cancer. This thesis includes the sequences of the most clonally expanded intratumoral antibodies expressed in these biopsies which I computationally reconstructed from bulk RNA sequencing reads.

Furthermore, by combining scRNA-seq and immune cell receptor profiling of samples collected from a novel mouse model, I present a comprehensive statistical analysis of gene expression in clonal tumor-reactive T cells. I also show the distribution of tumor-reactive clones across the tumor and spleen. This study forms a first proof-of-principle effort for the in-depth assessment of tumor-reactive T and B cell clones, in-vivo, and paves the way for further, more extensive experiments.

Acknowledgements

I would like to express my genuine gratitude to my supervisor, Professor Simon Tavaré, for his guidance, support and invaluable advice. Over the past four years, he has given me the freedom to explore what I truly find interesting while making sure I stayed the course. Being his student has been a privilege. I am equally grateful to Daniele Biasci for introducing me to the world of immunology, and for all the time he has spent mentoring me throughout my studies. His endless encouragement and support, as well as expertise and constructive criticism have made this research possible. I would also like to thank James Thaventhiran and Ty So for sharing novel clinical data, and the CRUK CI Genomics Core, especially Katarzyna Kania, for all their support and assistance.

I owe huge thanks to the Tavaré group and CRUK CI for their daily support and inspiration, especially Edward Williams, Sam Abujudeh, Juliane Perner, Ioana Olan, and Ann Kaminski. I would also like to thank all my friends at Darwin College, the Darwin Boatclub, and Darbar. I am deeply thankful to my dear friends Brenda Valeiras, Elaine Gray, Robbin de Kruijf, Giles Shaw, Can Aztekin and Branislav Turčina for their incredible company and care. I would like to especially thank Emile Alexandre Marin and Michael Scherm for being my home away from home, and my family.

I am forever grateful to my life long friends Mustafa Atalay, Pınar Akçayöz De Neve, Tuğçe Ayerdoğan, Güliz Yazan, Emre Gençay, Berrak Gümüşkaya Öcal, and their families, for their constant love and support as well as all their scientific expertise which went into this thesis. İyi ki varsınız.

Without my family none of this would have been possible. I am beyond thankful to Türkan, Esin, Ahmet, Faith, Defne, Yasemin, Deniz and Levo Gürel for everything. I'm thankful to my brother, Bora Gürel, particularly for the ever so amusing sibling rivalry and for getting me into cancer research and patiently explaining to me how cancer works.

Finally, I would like to thank my parents Ayşe and Türker Gürel for their never-ending support and trust in me, their parental wisdom and scientific input, and for keeping my spirits up and throwing food at me throughout the writing process. Ebeveynlerim, çocuğunuz olduğum için büyük mutluluk ve şans, sonsuz minnet duyuyorum. Sizi çok seviyorum.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AgR Antigen receptor

APC Antigen-presenting cell

BCR B cell receptor

CC Clonotype count

CCPM Clone counts per million

CRC Colorectal cancer

CTLA-4 Cytotoxic T-Lymphocyte Associated Protein 4

GC Germinal center

GEX Gene expression

HTS High throughput sequencing

HVG Highly variable gene

ICB Immune checkpoint blockade

ICI Immune checkpoint inhibitor

Ig Immunoglobulin

IMGT The International Immunogenetics Information System

MAD Median absolute deviation

MHCI Major Histocompatibility Complex class I

MHCII Major Histocompatibility Complex class II

MNN Mutual nearest neighbor

PCA Principal component analysis

PCR Polymerase chain reaction

PD-1 Programmed cell death-1

PDAC Pancreatic ductal adenocarcinoma

QC Quality check

RACE Rapid amplification of cDNA ends

scRNA-seq single-cell RNA sequencing

SHM Somatic hypermutation

SKCM Skin cutaneous melanoma

SLO Secondary lymphoid organ

SNN Shared nearest neighbor

t-SNE t-distributed stochastic neighbor embedding

Tc Cytotoxic T cell

TCGA The Cancer Genome Atlas

TCR T cell receptor

Tfh T follicular helper cell

Th Helper T cell

TIL Tumor-infiltrating lymphocyte

TLS Tertiary lymphoid structure

TME Tumor microenvironment

Treg T regulatory cell

UMAP Uniform Manifold Approximation and Projection

UMI Unique molecular identifier

Chapter 1

Introduction

1.1 The adaptive immune system and cancer

Mutations such as copy number alterations, point mutations, and chromosome rearrangements, occur frequently during the life of a cell, and can sometimes have serious consequences for the host. Cancer is a group of diseases where the host organism's self cells behave abnormally and divide rapidly, and possibly invade healthy tissue and organs. The immune system constantly monitors for such deleterious cells and normally eliminates them before they become life-threatening. It is widely accepted that the main reason behind the evolution of the immune system is to defend the host against pathogens, i.e., any microorganism that can cause disease in the host. In order to respond to a vast array of pathogens, the immune system utilizes a range of white blood cells, known as *leukocytes*, which can be at various stages of differentiation. These immune cells are examined in two major clusters: cells of the *innate* and of the *adaptive* immune system [117, Chapter 1, p. 5-17].

Innate immune cells are the first line of defense, and they are non-specific; they target anything that is non-self and potentially harmful, such as bacteria, viruses, or pollen. *Macrophages*, which are a particular type of *phagocytic cells* - the "eating" or "devouring cells" - that engulf bacteria and viruses in order to destroy them, are examples of innate immune cells. Another example is *Natural Killer* (NK) cells. NK cells are cytotoxic; they can release toxic granules containing perforin and granzymes to kill the target cell [238]. The innate immune cells are rapid in response, but in comparison to the adaptive immune cells, they are static; the host has an absolute arsenal - inherited from ancestors - with no memory that can accumulate on their own during the lifetime of the host organism [155]. Some innate immune cells also play a role in the adaptive immune system. The

dendritic cells, for example, present antigens on their surfaces in order to trigger the adaptive immune system [126, Chapter 6, p. 186-188].

The adaptive immune system, or *acquired immunity*, is generated in response to a specific type of antigen and has an immunological memory that learns and stores information to provide increased immune response against repeating pathogens caused by the same pathogen. Compared to the innate immune response, the adaptive immune response is much slower; however, its effects are highly specific and sustained long-term [155]. The adaptive immune system relies on two major cell types: *B cells* and *T cells*. Both cell types are derived from the same type of stem cell, namely *multipotent hematopoietic stem cells*, in the bone marrow. However, after they are generated in the bone marrow, they mature and become activated following different paths [117, Chapter 2, p. 24-38].

1.1.1 B cells

In birds, B cells mature in the bursa of Fabricius, a lymphoid organ that acts as a lymph gland and gives these cells their name [79]. In mammals, they mature in the bone marrow, becoming what is called a *naive B cell*, and move into the lymphatic system to circulate throughout the host organism. B cells produce protein molecules called *immunoglobulins* (Ig) to bind to pathogens and neutralize them. Immunoglobulins produced by an individual B cell, have the same amino acid sequence and the same binding-site, which are both unique to that specific cell. These sequences are inherent to the organism's DNA and are generated via genetic recombination with a theoretical diversity greater than 10^{18} [62]. Initial Igs produced by newly formed B cells are inserted into the cell's plasma membrane. These distinct antigen-specific surface receptors, called *B cell receptors* (BCRs), recognize and bind to antigens [162, Chapter 4, p. 141-152].

Naive B cells are activated when they encounter an antigen. The B cell which recognizes a specific antigen via its BCR proliferates and differentiates into either *memory B cells* or *antibody-secreting effector cells*. Memory B cells are record keepers for the specific antigens which activated them. Upon a possible immune response recall, they can quickly reactivate and proliferate, and differentiate in order to secrete *antibodies*. An antibody is a soluble form of Ig which is not membrane-bound [126, Chapter 6, p. 198-200].

Antibody-secreting effector cells produce large amounts of antibodies in response to a particular antigen. Antibodies can either label pathogens and unwanted cells to signal destruction to other components of the immune system or can directly neutralize them themselves. Secreted in large amounts, antibodies are normally one of the most

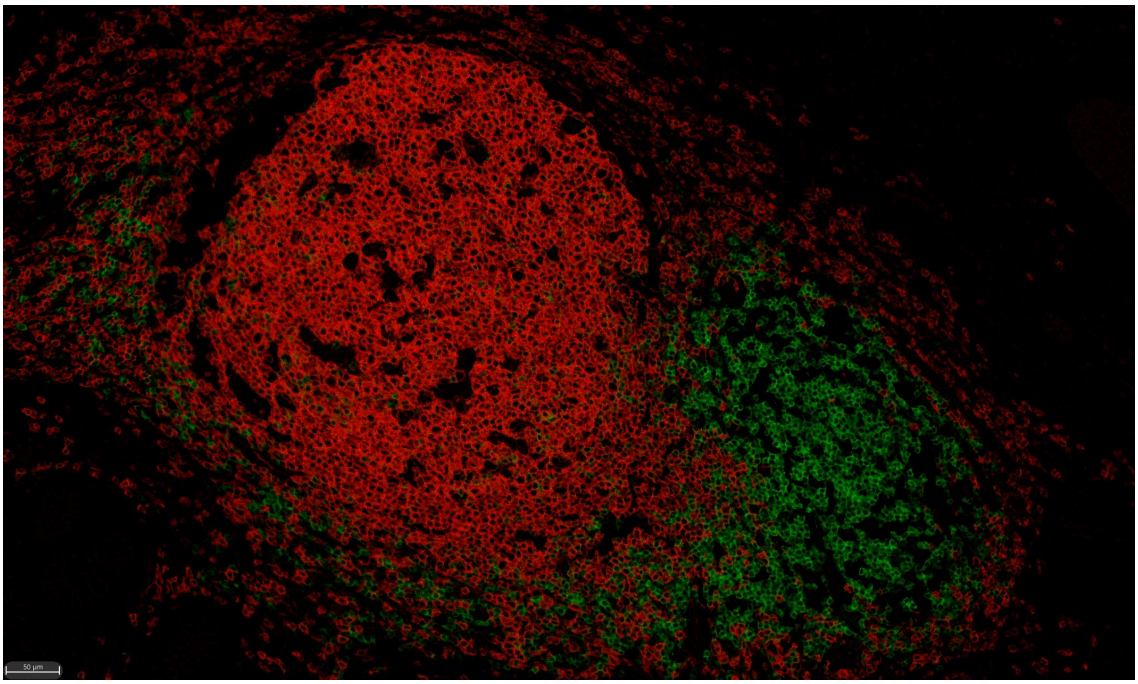


Fig. 1.1 Reactive lymphoid follicle in a lymph node, showing the classical distribution of T and B lymphocytes. B cells (CD20, texas red) are abundant in the germinal center, where activated B cells proliferate, mutate, and undergo clonal selection. B cells either differentiate into plasma cells that produce high-affinity antibodies against target antigens; or memory B cells, which then migrate out of the germinal center into the marginal zone [126, Chapter 6, p. 189-195]. T cells (CD3, FITC) are located in the peripheral zone, surrounding the germinal center. Naive B cells are also usually present in the peripheral zone (immunofluorescent stain, CD20:texas red, CD3:FITC, 200X). Image courtesy of Dr. Bora Gürel.

abundant protein components in plasma [34, Chapter 24, p. 1315]. The end-stage of an antibody-secreting effector B cell's differentiation is called a *plasma cell*. Plasma cells are much larger and dedicate most of their protein-synthesizing capacity to continuously secreting antibodies. Secreted antibodies circulate throughout the host organism to recognize free pathogens and other threats.

Proliferation and differentiation take place in the *germinal centers* (GCs), which are microenvironments formed within the B cell follicles of secondary lymphoid tissues upon an immune response. Some of the activated B cells migrate to lymphoid follicles and proliferate to form germinal centers [127]. We can think of germinal centers as specialized factories that are built upon an attack to produce fighters. Although activated B cells can immediately start secreting antibodies for an initial response, in germinal centers, B cells go through various changes to provide a more effective response. These modifications are crucial to finely tune the antibody for the antigen that it needs to fight. The selected

memory or plasma B cells exiting the germinal centers will have higher-affinity antibodies [117, Chapter 11, p. 267-272].

1.1.2 T cells

T progenitor cells that leave the bone marrow migrate to the thymus in order to mature and become T cells (T from thymus). Similar to B cells, T cells also express distinct antigen-specific surface receptors, namely *T cell receptors* (TCRs). In addition, T cells express various surface markers, including CD3, CD4, and CD8, which distinguish their functionality. Thymic selection ensures that mature T cells express either one of the surface markers CD4 or CD8[1]. While BCRs can bind to antigens directly, TCRs can only recognize antigens that are bound to specific receptor molecules, namely *Major Histocompatibility Complex class I* (MHCI) and *class II* (MHCII). While MHCI molecules are expressed on the surface of all nucleated cells, MHCII expression is generally restricted to a few cell types, including *antigen-presenting cells* (APCs), such as dendritic cells and macrophages. During development, T cells undergo two selection processes. *Positive* and *negative selection*, in combination, ensure that only the self-tolerant T lymphocytes that recognize self-MHC molecules survive, while the majority are destined for programmed cell death - *apoptosis* [117, Chapter 10, p. 221-244]. It is important to note that T cell self-recognition is not restricted to thymic self-peptides, as surviving T cells of thymic selection would then assume cells of other tissues to be non-self and create an immune response towards them. The protein AIRE is responsible for the negative selection of organ-specific T cells which are not typically expressed in thymic cells [145].

Surface markers help distinguish T cells in three main types with different functions: *Helper T cells* (*Th*) express CD4, and help in the recruitment and activation of other immune cells. *Cytotoxic T cells* (*Tc*) express CD8, and are, in essence, the killer cells responsible for eliminating unwanted cells. *T regulatory cells* (*Treg*) co-express CD4, CD25, and the transcription factor FOXP3, and help distinguish between self and non-self molecules, hence monitor and control immunity. Once a threat is eliminated during a primary immune response, most effector T cells die, leaving behind some long-lived *memory T cells* at standby for a possible secondary immune response. These cells can circulate between blood, secondary lymphoid organs, non-lymphoid tissues, or reside in tissues without circulation [166].

The adaptive immune system provides *humoral* and *cell-mediated* immunity, which are dependent essentially on the functions of B and T cells, respectively. Humoral immunity

[1] Sections 4.2.1 and 4.5.2 describe the thymic selection and provide cases of possible mature $CD4^+CD8^+$ T cells.

refers to the bodily fluids where the secreted antibodies bind to antigens. B cells are the key players of the humoral immune system and, as mentioned, are responsible for producing antigen-specific antibodies. With cell-mediated immunity, the response is carried out by different T cell populations [117, Chapter 1, p. 9-17].

1.1.3 Immune system interactions with cancer

It is estimated that more than 20,000 DNA repair events take place per cell every day by specific DNA repair pathways [143]. Cells that are not repaired might acquire malignant changes. Given that it is the role of the immune system to monitor for and protect against such damaged cells, it is surprising that cancer is the second leading cause of death globally, responsible for over nine million deaths in 2018 [250].

The immune system recognizes and attacks cancerous cells continuously via *immunosurveillance* which is thought to be based on a cell's expression of tumor-specific neoantigens or tumor-associated antigens [229]. Mutated proteins from oncogenes or other genes expressed and presented on the cell surface, and abnormal over-expression of self-proteins found on tumors are targets for immunosurveillance. Both the innate and adaptive immune effector cells play a role in controlling cancer [153]. NK cells of the innate immune system can detect and kill the initial, altered self-cells via direct tumor cell lysis. Later, macrophages and dendritic cells can engulf the fragments spread from this destruction and present them to T cells, thus triggering an adaptive immune response. The adaptive immune system eliminates the remaining tumor cells and generates a long-lasting immune memory specific to the recognized tumor components. According to the immunoediting theory [205], [159], this initial stage is called the eliminating phase. However, some tumor cells can escape the host immune response and survive into the next stage, called the equilibrium phase. During this phase, the tumor is present but does not grow or further metastasize, and the immune system is keeping it under control via immunosurveillance. This stage can be long-lasting, imitating elimination. Again, however, this equilibrium can be compromised with tumor cells that can resist, avoid, or suppress the anti-tumor immune response, resulting in tumor escape. These three phases, elimination, equilibrium, and escape, represent a high level, but a broadly accepted interpretation of the interplay between cancer and immune cells [169].

1.1.4 Cancer immunotherapy

Cancer cells can bypass immune cells by avoiding immune recognition, developing resistance to attack by immune cells, or instigating an immunosuppressive tumor mi-

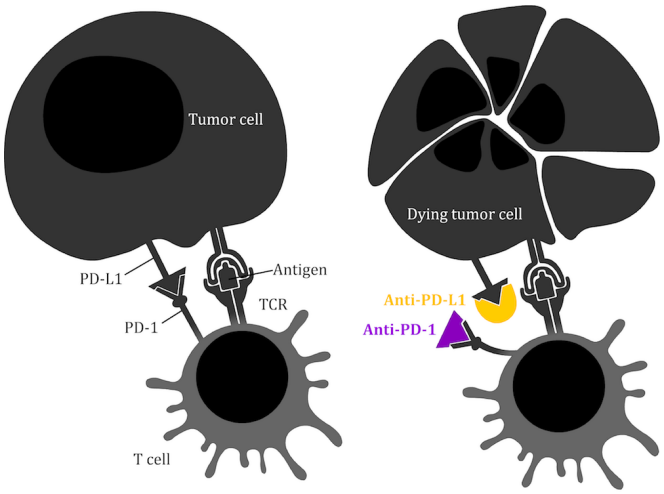


Fig. 1.2 PD-1/PD-L1 immune checkpoint inhibitors. PD-L1 binds to PD-1 and inhibits T cell killing of tumor cell. Blocking PD-L1 or PD-1 allows T cell killing of tumor cell. This illustration was drawn based on Figure 1 in [170]. Image courtesy of Ms. Güliz Yazan.

croenvironment (TME) [159], [169]. According to the cancer immunotherapy theory, it should be possible to slow or arrest tumor growth and prevent it from metastasizing by strengthening and restoring the host organism's immune system. There is a growing range of cancer immunotherapies which aim to provide long-term tumor control or complete elimination: therapeutic cancer vaccines which for the moment only prolong survival for cancer patients [157], Chimeric Antigen Receptor (CAR) T-cell therapy where a patient's T cells are modified to bind to a specific antigen on the patient's tumor cells and kill them [112], stimulating the immune system by exploiting cytokine signaling networks [133], and checkpoint blockade therapies, which have sparked the most interest [189], are all rapidly evolving treatment strategies in oncology.

Immune checkpoint inhibitor therapy

Immune checkpoint molecules are receptors found on the surface of several immune cell types, including B, T, and NK cells [232], [236], [178]. CTLA-4 (Cytotoxic T-Lymphocyte Associated Protein 4) and PD-1 (Programmed cell death-1) are the most studied immune checkpoint molecules [189]. Through such immune checkpoints, the immune cell functions are regulated, ensuring that the immune response is directed to the right cells while limiting the extent and duration of response. However, tumor cells can also exploit these immune checkpoints in order to suppress anti-tumor immune responses [52].

Immune checkpoint inhibitors (ICI) such as PD-1 inhibitors, PD-L1 inhibitors, and CTLA-4 inhibitors are drugs that block the targeted checkpoint proteins in order to stop

them from binding and therefore promote immune-mediated elimination of tumor cells. Figure 1.2 illustrates the blocking of the PD-1 or PD-L1 checkpoint molecules. Immune checkpoint therapies using such blocking monoclonal antibodies have demonstrated positive clinical outcome in several cancers including non–small-cell lung cancer (NSCLC), renal cell cancer, and has been particularly successful in melanoma [189]. Ipilimumab, an anti-CTLA-4 monoclonal antibody therapy, was the first drug to show long-term survival in patients with advanced melanoma [98]. It has been shown that among 1,861 melanoma patients treated with ipilimumab, the median overall survival was 11.4 months, with possible continued response up to 10 years after starting therapy [203]. Nivolumab, an anti-PD-1 monoclonal antibody, improved the overall survival of patients with melanoma and NSCLC [237], and pembrolizumab, another anti-PD-1 monoclonal antibody, increased progression-free survival and overall survival in patients with advanced melanoma and had less toxicity than ipilimumab [191]. Dual immune checkpoint blockade of anti-PD-1 and anti-CTLA-4 has demonstrated an overall survival rate of 58% at three years in the nivolumab-plus-ipilimumab group and 52% in the nivolumab only group, and 34% in the ipilimumab only group [252]. However, 59% of patients administered the combination immunotherapy experienced toxicity, compared to 21% of those in the nivolumab only group, and 28% in the ipilimumab only group. With the recent publication of the five-year follow-up [130], it has been reported that the overall survival at five years is 52% in the nivolumab-plus-ipilimumab group, 44% in the nivolumab group, and 26% in the ipilimumab group.

Owing to such drugs, immunotherapy has become a powerful clinical strategy for cancer treatment. Numerous other treatments are in clinical and preclinical developmental stages [81]. However, the outcome is not the same for every patient; not every patient responds the same to therapy. Checkpoint blockade therapy so far only helps a minority of patients, and its effects are not long-lived [140], [169]. In a study, it was observed that the type, density, and location of immune cells within the tumor site could predict clinical outcome in colorectal cancer [73]. A scoring system, namely *Immunoscore*, based on the quantification of $CD3^+$ and $CD8^+$ cell populations both at the tumor site and the invasive margin was devised following this finding. With this scoring system, tumors are classified based on their immune infiltration. This led to the concept of hot (highly infiltrated) and cold (non-infiltrated) tumors. Furthermore, studies analyzing the TME have observed that response to immune checkpoint blockade favors pre-existing T cell infiltration in tumors [26]. Effective immune response post-anti-PD-1 therapy requires the reactivation and clonal-proliferation of antigen-experienced immune cells to be present in the TME [214].

It is thus essential to analyze the TME prior to therapy as well as post-therapy, in order to identify patients for whom immune checkpoint therapies could result in positive outcomes. Methods that can expose clonally expanded immune cells and fully characterize the TME could help transform immunotherapy into a more broadly applicable treatment for cancer [188], [266].

1.1.5 Tertiary lymphoid structures

The cell-mediated and humoral immune responses against cancer are thought to be primarily initiated in the secondary lymphoid organs (SLOs) such as the lymph nodes and the spleen [48]. APCs that have processed antigens from the tumor site travel to the SLOs where they present these peptides to B and T cells, initiating their activation. Following proliferation and differentiation, immune cells migrate into the tumor in order to eliminate the tumor cells [80]. Hence, SLOs are considered crucial in initiating and regulating an adaptive immune response to cancer as they provide the grounds for interaction between immune cells and tumor antigens.

However, recent studies have shown that SLOs are not the only sites where the anti-tumor immune response is generated [201]. Tertiary lymphoid structures (TLSs) are ectopic lymphoid-like structures that develop in non-lymphoid tissues at sites of infection or chronic inflammation, including tumors. TLSs have been observed in the stroma, invasive margin, and within the center of many solid tumors, including breast, pancreatic, NSCLC, and melanoma [48]. TLSs organizationally resemble lymph nodes. In primary cancers, TLSs have been reported to comprise a T cell zone, a B cell zone with characteristics of a germinal center mainly containing B cells, and antigen-presenting cells such as DCs, and macrophages [80]. TLSs provide a local setting for T and B cell activation, proliferation, and differentiation through the presentation of antigens from the neighboring tumor [201]. A dense population of plasma cells has been observed surrounding TLSs in various tumors [123], [132], [231] and for example, in ovarian cancer, the presence of tumor-infiltrating antibody-producing plasma cells has been reported to strengthen the prognostically favorable cytolytic T cell responses [123]. Studies suggest that tumor-infiltrating B cells could be presenting tumor antigens to T cells [35], [261]. Moreover, it has been hypothesized that TLSs enable the local production of antibodies [103]. Although it is accepted that $CD8^+$ T cells drive anti-tumor immunity, observations suggest that TLSs may be facilitating an effective anti-tumor immune response achieved via the coordination of T and B cells.

Furthermore, an association between the presence of TLS and a favorable prognosis has been reported for many solid tumors, although the outcome depends on parameters

including cancer type and disease stage. An extensive review is presented in [201]. Besides established biomarkers such as the quantification of pre-existing anti-tumor T cells, PDL1 expression at the tumor site, or tumor mutational burden, the presence of TLSs is being regarded as a prognostic and predictive factor in patient response to immunotherapies, as well as a criterion for patient selection for immunotherapy [48].

1.2 B and T cell antigen receptors guide the adaptive immune response

B and T adaptive immune cells utilize their antigen-specific surface receptors to recognize a vast array of threats against the host organism. This highly specific recognition is achieved via the considerable diversity of BCRs and TCRs and decides the fate of a cell; the cell's development, survival, and activation are heavily influenced by signals received via its BCR or TCR.

BCRs are composed of two identical heavy chains encoded by the IGH gene locus and two identical light chains encoded by the IGK or IGL genes, respectively. TCRs are composed of either α and β or γ and δ chains where each chain is encoded by a different germline TRA, TRB, TRG, or TRD gene locus, respectively (see Figure 1.3). Each of these chains is made up of one of the many possible germline variable (V), diversity (D), joining (J), and constant (C) gene segments. One of each V, D (only in BCR heavy and TCRβ chains), J, and C segment are combined through somatic DNA rearrangements during B and T cell development. This process is known as *V(D)J recombination* and is the major driver of the receptor diversity. Additional diversity between receptors is generated by random insertion and deletion of nucleotides at the V-(D)-J junction sites. The pairing of heavy and light, and α and β or γ and δ chains is also a factor of diversity [117, Chapters 4, 5, and 9].

BCR and TCR chains contain three complementarity determining regions: *CDR1* and *CDR2* are encoded solely by V genes in the germline DNA segments, whereas *CDR3* is encoded by the terminal of V genes, the D genes in the case of BCR heavy and TCRβ chains, the beginning of J genes, and the junctional sites between those genes, making it the most variable region in a chain (see Figure 1.4) [117, Chapter 5, p. 115], [117, Chapter 9, p. 207]. CDR3 regions are also the primary antigen-binding site of the receptors [74]. Due to these characteristics, antigen receptors (AgRs) are, in general, identified by their CDR3 regions [137].

When an adaptive immune cell recognizes a specific antigen, it might proliferate, leading to clonal expansion. Cells expressing identical receptors are said to be of the

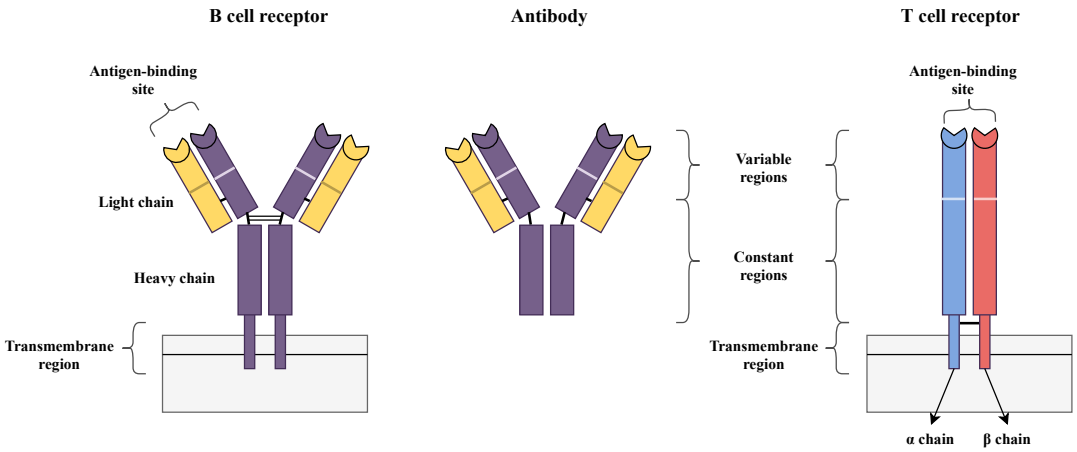


Fig. 1.3 B and T cell receptor structure. Figure adapted from [174].

same *clone* and their AgRs of the same *clonotype* [137]. A clone can be defined by the set of unique CDR3 regions its receptors encode, meaning that cells with receptors that have identical sequences in their CDR3 region are considered to be of the same clone. With some exceptions [239], it is highly unlikely that two cells with no common predecessor express the same receptor, hence cells of the same clone are considered to have proliferated from the same ancestor cell. When a B or T cell is antigen challenged, it divides and undergoes clonal expansion. Thus the AgR sequences serve as a unique "clonal index" which decodes antigen specificity and cell lineage.

In theory, the number of unique BCR and TCR sequences can surpass 10^{18} [62]. However, there are only 3.72 $\times 10^{13}$ non-bacterial cells in the human body [20], and some BCR and TCR tend to be more frequent than others due to antigen specificity and other factors, hence the full possible collection cannot be realized in a single host. The adaptive immune repertoire refers to the collection of B and T cell receptors within an individual host, and is formed throughout a lifetime, shaped adaptively in response to antigen challenge. The clonal diversity and distribution, and abundance of BCRs and TCRs are used to characterize this dynamic repertoire.

1.2.1 BCRs

In addition to the combinatorial, junctional, and coupling diversity, BCRs achieve increased functional diversity through *isotype/class switching* their constant regions (C-region), *somatic hypermutation*, and *affinity maturation*.

Mammals have five classes of antibodies: IgA, IgD, IgE, IgG, and IgM, where each has a distinct heavy chain C-region α, δ, ϵ, γ, and μ, respectively. In humans, IgG and IgA

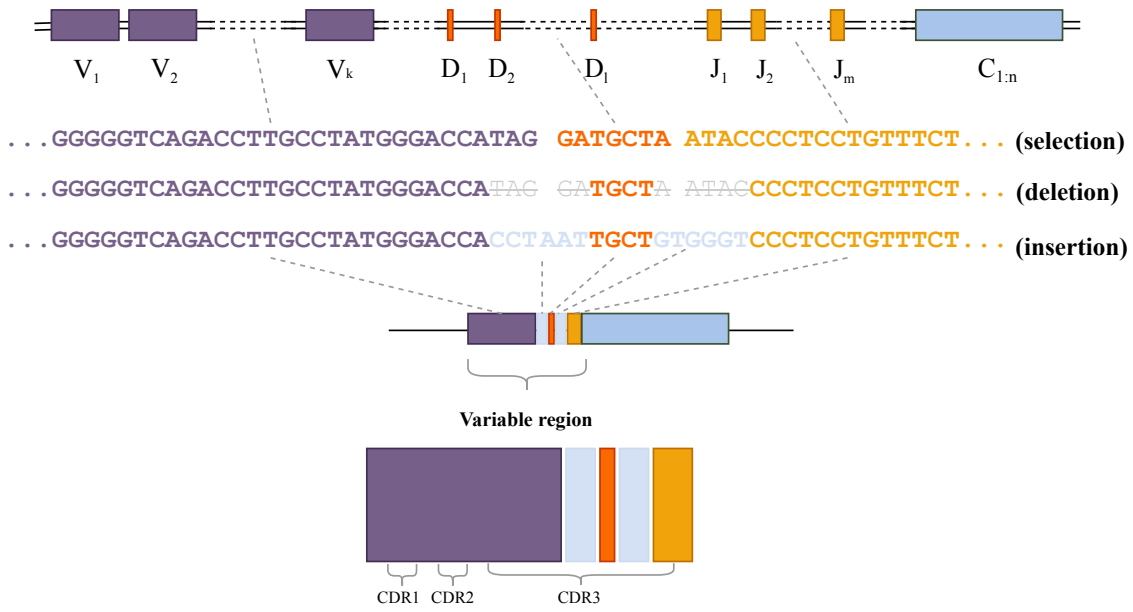


Fig. 1.4 The V(D)J recombination process, in which individual V, D, and J genes are randomly selected (selection). These genes are then joined together after a process that deletes random nucleotides on the boundaries (deletion) and inserts random nucleotides in the light blue N-region (insertion). The diversity and specificity of an antigen receptor are, to a large extent, determined by the recombination site CDR3. I have illustrated this process based on details provided in [152, Chapter 4, p. 85-110].

have further subclasses IgG1, IgG2, IgG3, and IgG4, having γ_1, γ_2, γ_3, and γ_4 heavy chain C-regions, respectively, and IgA1 and IgA2 with α_1 and α_2 heavy chain C-regions. Each class has different functional properties and determines what follows antigen binding. Naive B cells express IgD and IgM, either as BCRs or secreted antibodies. IgM is secreted into the blood during a primary antibody response, upon initial exposure to an antigen. IgG is secreted in large quantities during secondary immune responses and makes up the majority of immunoglobulin in the blood. As an immune response evolves, a BCR can "switch" its "class" to alter its functional role. Following antigen binding, by excising unwanted isotypes, active B cells may express IgG, IgA, and IgE, or continue expressing IgM [34, Chapter 24, p. 1315-1318]. While the exact process behind class-switching is unknown, it has been observed that cytokines play a major role in determining which isotype a B cell will express [117, Chapter 5, p. 121-122].

Antibodies produced during the early stages of an immune response usually have much lower binding strength compared to those produced later on. This is achieved overtime via fine-tuning their ability to bind an antigen through a process called somatic hypermutation (SHM). Point mutations accumulate in the V-regions of both the heavy and light chains, with the overall goal of producing diversity, but not affinity. Antibodies

undergoing SHMs can end up with higher, lower, or fixed strength antigen binding. SHMs can also produce non-productive antibodies. In order to increase a B cell's affinity for a specific antigen, it undergoes *affinity maturation*, which is the repeated process of SHMs and subsequent selective survival of high-affinity antibodies. Mutations tend to accumulate in the CDRs. As affinity maturation is guided via antigen binding, it is the CDRs that end up with the most mutations after affinity maturation [117, Chapter 5, p. 117-122].

Clonal expansion, class switching, SHMs, and affinity maturation all take place in intratumoral TLSs and GCs [231]. Naive B cells entering a GC clonally expand and accumulate SHMs. Cells that generate antigen affinity increasing mutations in their antibodies are selected for class switching. They then differentiate into plasma or memory B cells and exit the GCs (see Figure 1.5). Cells which, after SHM, produce antibodies with a decreased antigen affinity are signaled for apoptosis. Similarly, in GC found in TLSs, B cells undergo terminal differentiation into effector cells and migrate from the TLS B-cell follicle to the tumor stroma [76].

In this thesis, when describing BCR reconstruction, the terms BCR and antibody are used interchangeably if not specified otherwise since their sequences are identical. I focus on BCRs in Chapters 2 and 3 in which I present complete computational reconstructions of intratumoral BCR repertoires.

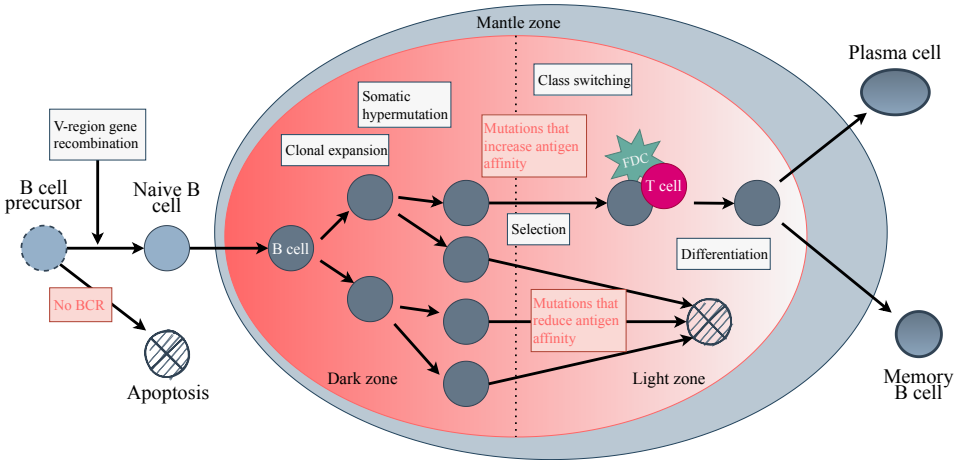


Fig. 1.5 B cell affinity maturation in GC. Naive B cells entering the GC are fine tuned for a given antigen. B cells with decreased affinity are signalled for apoptosis while ones with increased affinity differentiate into memory and plasma B cells. Figure adapted from [127].

1.2.2 TCRs

TCRs are always membrane-bound, and they guide cell-to-cell adaptive immunity [162, Chapter 4, p. 153-155]. In humans and mice, the TCR$\alpha\beta^+$ lineage makes up more than 90% of the T cells in the peripheral immune system [126, Chapter 6, p. 190]. In general, TCRs do not undergo SHMs when a T cell is exposed to antigen. Thus, the antigen affinity of a given T cell clone and their TCR clonotype remain static after V(D)J recombination. However, SHM was observed in the TCRα of mice and TCRβ of HIV-infected patients, as well as the variable domain of TCR in camel and shark [21].

In Chapter 4 of this thesis, I give a detailed description of T cells and TCRs and present a comprehensive statistical analysis of gene expression of clonal tumor-reactive T cells coupled with their TCR repertoire profiles.

1.3 Strategies for antigen receptor reconstruction

In the past, immune cell receptor studies would only recover a tiny fraction of the complete repertoire [11]. Early studies characterizing the TCR repertoire were conducted at the protein level using flow cytometry and monoclonal antibodies [64]. Experiments at the genomic level were initially based on CDR3 sequence length rather than the actual nucleotide sequences [83]. Characterization of the TCR repertoire at the nucleotide sequence level was achieved via molecular cloning techniques coupled with Sanger sequencing [200]. Although Sanger sequencing provides specificity, it is not reasonable to assume capture of the immense diversity using a method with such low capacity - studies were only able to sequence receptors in the 10^2–10^3 range [11]. [222] provides an extensive review of methods developed to recover AgRs and assess repertoire diversity prior to high throughput sequencing (HTS).

With the advent of HTS, research focusing on AgR repertoires gained significant momentum. Cost-effective massively parallel sequencing of millions of DNA molecules offered a comprehensive view of receptor sequences, including V(D)J regions and the complete CDR3 sequence. In 2009, using the HTS platform Roche 454 - which can provide, on average 500 bp length reads - for immune repertoire analysis for the first time, [248] described the diversity of the antigen-binding domain of the Ig heavy chain recovered from 14 zebrafish to analyze V(D)J usage and antibody sequences. Their work was followed by others using the 454 sequencing platform to cover the entire V(D)J region of the human BCR and TCR repertoires [29], [247]. Studies next utilized the Illumina HiSeq platform, which offers shorter read length but much higher read throughput at

a lower cost, allowing for deep sequencing of repertoires [70], [192]. Since then, HTS profiling of immune receptors has been used extensively to study adaptive immunity [39].

However, the alignment of AgRs has to go beyond those developed for genome sequencing as their intrinsic nature, especially of BCRs, present particular challenges in recovering V(D)J sequences. First, as receptors are a product of V(D)J recombination, junctional insertions and deletions (indels), and possible SHMs, there are no reference sequences for CDR3s in the genome. Second, the edit distances between two different V(D)J sequences can be less than the acceptable sequencing error rate, i.e., sequencing reads may result in erroneous sequence variants with mismatches, and it may be challenging to decide whether the sequence represents the same CDR3 with technical errors or a completely different V(D)J sequence.

1.3.1 Targeted repertoire sequencing

Currently, there are two main approaches taken to reconstruct AgRs: targeted repertoire sequencing and computational reconstruction of unselected RNA-seq reads. Targeted repertoire sequencing methods selectively target TCR/BCR molecules; using polymerase chain reaction (PCR) and HTS, they generate a large number of short to mid-length DNA/RNA sequences covering regions of interest of AgRs. They employ a target enrichment step to increase the sensitivity, using either genomic DNA or cDNA as starting material. Multiplex PCR, targeted in-solution enrichment, and 5′RACE-switch-oligo nested PCR are the most commonly used enrichment methods [17]. Multiplex PCR uses a multiplex pool of forward PCR primers complementary to V segments, and depending on whether the starting material is gDNA or cDNA, either a pool of reverse primers designed for the J segments, or for the constant regions, respectively [59], [242]. Using quantitative multiplex immunofluorescence, and HTS of TCRs, [242] analyzed samples from 46 patients with metastatic melanoma obtained before and during anti-PD-1 therapy (pembrolizumab). Pre-treatment samples obtained from responding patients showed a more clonally expanded TCR repertoire, suggesting these patients had already mounted an adaptive immune response prior to therapy. They also concluded that successful anti-PD-1 outcome requires pre-existing $CD8^+$ T cells that are negatively regulated by PD-1/PD-L1-mediated adaptive immune resistance. Targeted/bait-based enrichment uses RNA baits which are complementary to the sequences of interest and tolerate a few mismatches. 5′ rapid amplification of cDNA ends (RACE) technologies decrease the amount of PCR bias that may result from using multiple primers for different V and J genes. This method adds only a single universal oligo at the 5′ end of the transcript, and a reverse primer to the C region of the transcript [70]. 5′RACE-switch-

oligo method uses transcript RNA as input material while the two others can use genomic DNA or transcript RNA.

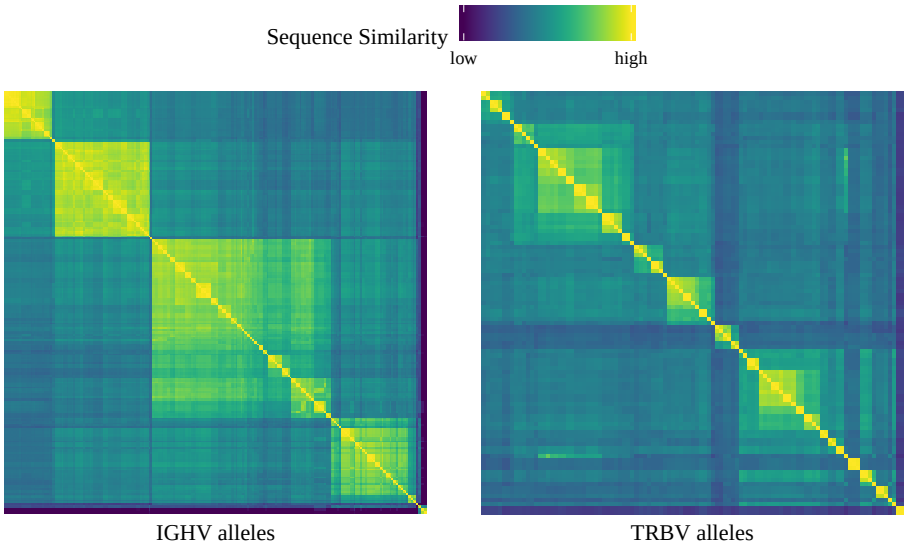


Fig. 1.6 Levenshtein distance matrix of human IGHV and TRBV alleles. Each cell is colored by the sequence similarity between any two alleles of a BCR heavy (left) and TCRβ (right) variable group, where high corresponds to a Levenshtein distance of 0, and low to 250 and above. The median length of human IGHV allele sequences is 353 bp, and human TRBV allele sequences is 344 bp.

Apart from the amplification technique, the choice of input material is also important and depends on the research question. Genomic DNA offers better stability and allows for better quantification of AgR clones as each cell contains a single DNA copy of the gene encoding each receptor chain, however, it does not provide any information on the gene expression level and may lead to sequencing errors because of introns or possible residuals of V(D)J recombinations [222]. With RNA, we can obtain the final BCR and TCR products along with dynamic expression level information, while excluding the non-productive sequences that are functionally irrelevant. However, when using RNA, quantification can be challenging as an immune cell will contain more than one immune receptor transcript, and BCR and TCR representation in plasma B cells and activated T cells will be quite high. RNA-based methods can utilize *unique molecular identifiers* (UMIs), which are a sequence of 8–12 random nucleotides attached to each initial cDNA, in order to detect and quantify unique mRNA transcripts after PCRs and sequencing [120]. These unique molecular barcodes allow for BCR/TCR transcript quantification and error correction during data analysis. As mentioned before, AgRs may differ by single nucleotides. Therefore it is imperative to distinguish between PCR or sequencing

errors, and low-frequency clonotypes. UMIs provide a reliable technique to distinguish technical errors from distinct AgRs or daughter AgRs with SHMs.

Once AgRs have been sequenced, the raw HTS platform sequence data files are converted to FASTQ format [47] and are then processed to match the obtained sequence to the known germline gene sections. However, unlike standard alignment where short reads are mapped to the most similar regions of the reference, this gene matching process is a bit tricky. The BCR and TCR loci are made up of many similar V, D, and J genes, which need to be distinguished accurately. As an example, Figure 1.6 shows this similarity with a heatmap of all the human BCR heavy and TCRβ V allele edit distances. Also, these BCR/TCR variable regions accumulate insertions and deletions and possible SHMs during recombination and affinity maturation, diverging them from the actual gene sequences. Table 1 in [96] lists the current receptor gene assignment and profiling tools that were developed specifically for targeted AgR sequencing datasets.

1.3.2 Computational reconstruction and repertoire profiling of antigen receptors from bulk RNA-seq reads

The alternative to targeted AgR sequencing is computationally reconstructing AgRs from RNA-seq reads obtained from any sample containing immune cells. The downside of this method is that a small fraction of such sequencing reads will come from BCRs and TCRs, and a high sequencing depth will be required to pick up the AgR signal in the RNA-seq read pool [33]. However, as new computational methods develop, this is becoming less of a challenge. [33] extracted TCR sequence information from the RNA-seq data of 6,738 tumor samples taken from The Cancer Genome Atlas (TCGA) [164], with a typical yield of one unique TCR per 10 million reads. However, their poor detection rate is dependent on using computational methods that are not explicitly designed for unselected RNA-seq data. In comparison, [137], with a specialized method, detected an order of magnitude more distinct CDR3 sequences in the same dataset.

The fraction of BCR and TCR reads covering CDR3s in RNA-seq data varies from sample to sample depending on the degree of immune cell infiltration, and can range from 10^{-4} to 10^{-7} for BCRs and 10^{-5} to 10^{-7} for TCRs [24]. Moreover, as the length of RNA-seq paired-end reads is typically in the 50-100 bp range, successful detection of target V(D)J junctions requires alignment with very short fragments of germline V and J genes (12-15 bp). Another challenge in the evaluation of such short sequences is the high probability of false-positive alignments. Efficient reconstruction of antigen receptors

from raw RNA-seq reads thus requires the extraction of as many true CDR3 sequences as possible while ensuring nearly zero CDR3-like false-positives.

Targeted amplicon sequencing based methods cannot evaluate AgR variation in the context of genetic diversity. RNA-seq based transcriptome analysis is now quite affordable and routine in biological research and clinical studies. Extracting AgRs directly from the readily available RNA-seq data provides the ability to assess receptor clones within the gene expression space. Furthermore, the quantity of starting material is a factor that can make targeted amplicon sequencing based methods unfavorable. Especially in cancer studies, tumor tissue samples taken during ongoing treatment can be quite limited as they are often from needle biopsies. If only low quantities of starting material are available, it is beneficial to extract the BCR/TCR information directly from RNA-seq data along with other -omics data, rather than having to separate the sample for transcriptome, BCR, and TCR profiling.

Moreover, with the ongoing "fight against cancer", cancer researchers have been amassing data in centralized repositories such as TCGA [164] and the International Cancer Genome Consortium (ICGC) [104]. AgR reconstruction methods can employ such readily available databases for immune receptor profiling, providing an additional layer of information to pre-existing research without extra wet-lab time and cost. In this regard, [137] demonstrated their computational algorithm named TRUST [138] on unselected tumor RNA-seq data taken from the TCGA cohort to characterize the TCR repertoire of tumor-infiltrating T cells. One of their findings is that the diversity of T cell clonotypes positively associates with cancer somatic mutation load. Similarly, [24] employed MiXCR [25] to reconstruct AgR repertoires of 458 cutaneous melanoma

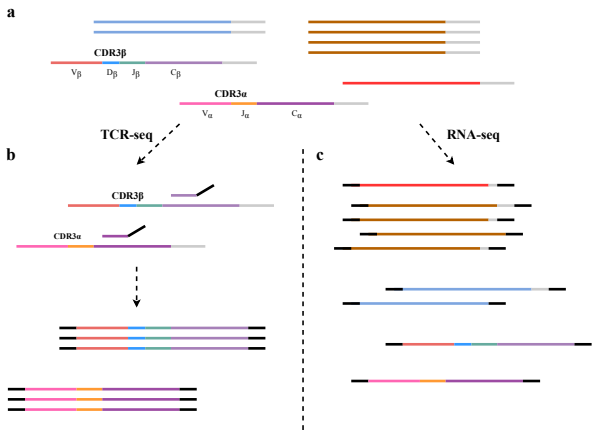


Fig. 1.7 Simplified representation of targeted AgR sequencing vs RNA-seq. Each horizontal line represents an mRNA transcript where each color denotes a unique gene sequence. (a) The mixed pool of all possible mRNA in a given sample including the recombined AgR transcripts (multicolored) and the non-AgR sequences (blue, brown, red). (b) Targeted sequencing amplifies the selected CDR3 region of receptor transcripts (displayed as a color gradient). (c) RNA-seq generates fragments from all transcripts present in the mixed pool, including fragments of the CDR3 encoding sequence. Figure adapted from [33].

(SKCM) patients from 48+48- bp paired-end RNA-seq data taken from TCGA. They associated a combination of high intratumoral IGH expression levels and high levels of IGH clonality with more prolonged survival. Patients with abundance and high clonality of IGH showed better prognosis compared to patients with a similar abundance of IGH but diverse repertoires.

1.3.3 Reconstructing antigen receptors from single cell RNA-seq data

The immune system is a complex network of cells that vary in type, state, and differentiation stage. In order to grasp the full heterogeneity of the immune system and understand the intricate underlying mechanism, studies need to be conducted at single-cell resolution. In the past 50 years, flow cytometry has provided significant developments in high-throughput quantitative analysis of cells and other particles [177]. In immunology, microscopy and flow cytometry are used to classify and quantify populations of immune cells based on the specific surface marker and intracellular protein expression. Recent computational flow cytometry tools and software have also begun providing reproducible and unbiased results that manual analysis could not [198]. However, such methods are still limited by the number of parameters for cell-type classification and the reliance on prior knowledge of known markers [44].

Single-cell RNA-seq (scRNA-seq) is an ideal method to reveal cellular heterogeneity. scRNA-seq measures the gene expression levels within a single cell, allowing for the classification of individual cells by transcriptome analysis rather than surface markers. In immunology, with scRNA-seq, cellular complexity can be investigated in far greater detail than conventional methods, exposing new cell types and functionally diverse subpopulations of known cell types [246], [22]. scRNA-seq has also allowed for the in-depth study of the adaptive immunity against cancer. [235] analyzed a total of 4,645 malignant and tumor cells from 19 human melanoma samples. They examined $CD8^+$ T cells to identify dysfunction programs. Similarly, [140] analyzed the tumor-infiltrating immune cells in human melanoma to reveal the trajectories of $CD8^+$ T cell dysfunction. Profiling immune cells from melanoma patients using scRNA-seq, [197] determined that TCF7 (TCF1 in mice) expression can predict positive clinical response. Many other studies have dissected tumor tissues using scRNA-seq to characterize the TME [175], [221], [184], [13], [202], [66].

The usual scRNA-seq workflow consists of single-cell capture, mRNA reverse transcription, cDNA amplification, library preparation, high-throughput sequencing, and

data analysis [44]. In some protocols, to eliminate the amplification bias, UMIs are attached to each transcript within a cell during reverse transcription [107]. The absolute copy number of a transcript in a single cell can be found by counting the number of distinct UMIs. The count of sequenced UMI reads that align to a specific gene represents the expression level of that gene. Each cell is presented as a vector of gene expression levels, and the cell vectors merged on genes make up the digital (gene x cell) expression matrix of the captured cells.

scRNA-seq is an ideal method to study the vast diversity of AgRs, along with their relation to their cells' type and state. Similar to bulk data, the sequences of BCRs and TCRs can be assembled from scRNA-seq reads or reconstructed using targeted AgR sequencing. With "bulk" approaches, information about the pairing of heavy and light chains in the case of Ig, and the α and β chain for TCR, cannot be obtained with certainty. In clonal repertoires, one may think to pair the most clonal heavy and light clonotypes together. However, the most frequent heavy or light chains may be expressed by several clones. They would be the most common chain, but not necessarily the most clonally expanded clonotype. Furthermore, in the case of allelic inclusion [227], [30], chains cannot be assumed to pair up. In this regard, scRNA-seq is superior to bulk RNA-seq in reconstructing AgR repertoires.

scRNA-seq also allows for the incorporation of single-cell gene expression profiles with BCR/TCR information at the cellular level in order to uncover the links and transitions between transcriptional states. [197], via single-cell gene expression and TCR analysis, showed that T cells can transition from one state to the other. This was later confirmed by [140], where again combining scRNA-seq and TCR-seq, they showed that the dysfunctional $CD8^+$ T cell state rather than being a discrete pool constituted a gradient spectrum going from transitional state to early dysfunction, and highly dysfunctional.

1.3.4 Ontology for immunogenetics

The International Immunogenetics Information System (IMGT), created in 1989 by Professor Emeritus Marie-Paule Lefranc, is the most widely used reference in immunogenetics and immunoinformatics [135]. It is a centralized repository for immunogenetics-related data, specializing in AgRs and MHCs of all vertebrate species. The IMGT/GENE-DB database [134] contains a well-annotated set of genomic V, D, J, and C genes. This database provides the reference for most AgR gene assignment tools, as well as for tools that can directly extract immune repertoires from whole-transcriptome RNA-seq data sets. These tools are reviewed in Section 1.3.5. IMGT/GENE-DB statistics [106] of

IGH, IGK, IGL, TRA, and TRB genes for homo sapiens and mus musculus species are provided below in Tables 1.1 and 1.2. The tables below also present the median allele lengths of each IMGT group. Note how short D and J genes are; often, the D segment remains unidentified due to its short length. Also, as shown previously with Figure 1.6, V gene segments are very homologous. Considering that it is the V(D)J region which has indels and SHMs the most, it becomes apparent that AgR reconstruction is more tedious than standard alignment. Differentiating between allelic differences and V(D)J recombination sequence alterations require extra care.

	Locus	**IGH**				**IGK**			**IGL**		
Species	IMGT group	IGHV	IGHD	IGHJ	IGHC	IGKV	IGKJ	IGKC	IGLV	IGLJ	IGLC
Homo sapiens	Number of genes	185	37	9	12	109	5	1	79	11	14
	Number of alleles	524	44	19	93	148	9	5	133	14	29
	Median length (bp)	353	23	52	1131	353	38	321	353	38	318
Mus musculus	Number of genes	352	31	4	9	177	5	1	8	5	4
	Number of alleles	508	38	9	26	208	10	3	14	5	5
	Median length (bp)	351	17	48	1032	351	38	321	351	38	315

Table 1.1 Number and median length in basepairs of IMGT IG genes and alleles of homo sapiens and mus musculus species per IMGT group.

	Locus	**TRA**			**TRB**			
Species	IMGT group	TRAV	TRAJ	TRAC	TRBV	TRBD	TRBJ	TRBC
Homo sapiens	Number of genes	57	61	1	77	2	14	2
	Number of alleles	115	71	1	168	3	16	6
	Median length (bp)	338	61	422	344	16	50	537
Mus musculus	Number of genes	132	60	1	35	2	14	2
	Number of alleles	279	71	2	61	2	19	4
	Median length (bp)	340	61	408	344	13	49	518

Table 1.2 Number and median length in basepairs of IMGT TRA and TRB genes and alleles of homo sapiens and mus musculus species per IMGT group.

1.3.5 Software tools and pipelines reconstructing antigen receptors from unselected RNA-seq data

Before discussing state of the art, it is necessary to mention that to date, there is no gold standard method, and that it is difficult to define one given our limited knowledge of immunogenomics. Certain dogmas such as each BCR has one heavy and one light chain, or that a lymphocyte can have either a BCR or a TCR but not both, are being proven wrong as our knowledge expands [7], [227], [30]. The tools developed need to be more flexible

in their assumptions. Benchmarking on real data sets is not straightforward, as it is not possible to know the actual repertoire of a sample before measuring it. Benchmarking is often carried out in silico; however, the simulated data may not be reflective of the actual sample space as it is far cleaner and incorporates possible false assumptions. That being said, every available method has its advantages and disadvantages, and it is up to the researcher to decide which tool, and which configuration will most accurately answer their research question.

IgBlast is one of the earliest BCR and TCR variable domain sequence V(D)J mapping tools [259]. IgBlast carries out read-mapping via BLAST [27] with an optimized set of parameters. However, IgBlast does not extract the sequence of CDR3 region directly from HTS data, meaning it is a mapping tool and not a complete reconstruction tool. Also, IgBlast is designed to analyze the V domain sequence. If the query sequence only contains D or J gene and not the V gene, it will not be analyzed adequately by IgBlast. Many reconstruction pipelines such as TraCeR [227] and BraCeR [144], use IgBlast in their workflow to analyze the assembled contigs.

A variety of software tools have been developed to reconstruct AgRs from unselected bulk and single-cell RNA-seq data directly. As mentioned before, the V(D)J recombination and possible SHMs renders the RNA-seq reads from AgRs unmappable to the reference sequence. TRUST [138], V'DJer [161], and MiXCR [25] reconstruct antigen receptors from RNA-seq data using different underlying algorithms to map the short RNA-seq reads to germline genes. TRUST first maps all reads to the reference genome and finds read pairs where one of the pairs properly maps to a TCR gene and the other read pair is unmappable, assuming so due to V(D)J recombination. Then taking the unmapped reads, it constructs an overlap matrix, represented by an undirected graph, where nodes correspond to reads and edges indicate partial sequence overlap. Disjoint subgraphs (cliques) represent potentially distinct CDR3 sequences. All reads in a clique are assembled *de novo* to obtain contigs, which are then annotated with amino-acid sequences, and the associated V and J genes using IMGT. Contigs not annotated as CDR3 regions are discarded to reduce false-positive rates. TRUST is developed specifically for TCR reconstruction, although recent updates include BCR CDR3 calling function as well [102]. V'DJer similarly reconstructs AgR sequences by extracting fully or partially Ig gene mapped and unmappable RNA-seq reads and with them constructs a De Bruijn graph [180]. The graph is traversed, producing possible contigs that are filtered for homology with V and J segments. The final set of assembled contigs spanning most of the V(D)J region and a portion of the constant region is returned with annotation from IMGT. V'DJer returns the most abundant portions of the BCR repertoire. MiXCR

enables assembly of both BCR and TCR clonotypes from various raw sequencing data such as whole-genome sequencing, RNA-seq, and paired-end sequenced cDNA library of IGH gene prepared using 5′RACE-based protocols. The MiXCR workflow consists of three main steps: (i) building alignments of sequencing reads to reference V, D, J and C genes, (ii) assembling clonotypes using the alignments, and (iii) exporting filtered alignments. The alignment step is an implementation of the k-mer chaining algorithm proposed by [142] and can handle short alignments and alignments with mismatches, indels, and big gaps. Identical sequences are clustered into clonotypes and corrected for PCR and sequencing errors. The clustering algorithm looks for fuzzy matches between clonotypes and organizes the matched clonotypes in a tree, where each daughter node is highly similar to its parent but has a significantly smaller abundance. Only parent nodes are considered as clonotypes. Lastly, clonotype sequences are aligned to reference V, D, J, and C genes using the Smith-Waterman algorithm [224]. Benchmarking with in silico generated data demonstrated MiXCR to have an order of magnitude higher efficiency in terms of reconstructing true clonotypes and filtering for false-positives when compared to TRUST [24]. Comparing the number of TCR-seq-confirmed RNA-seq-extracted high-frequency clonotypes of lymph node metastasis, MiXCR demonstrated better accuracy than that of TRUST. However, since the MiXCR-conducted-benchmarking, TRUST has been updated, clarifying the claimed performance gaps [101]. MiXCR, when compared with V'DJer, demonstrated significantly better computing performance [24]. Average analysis time of 10^8 RNA-seq reads, took 28 hours with V'DJer, and 4 hours with MiXCR. To analyze all IGH, IGK, and IGL chains, V'DJer needs to be run three times for each chain [24].

There are also pipelines specific to the reconstruction of AgRs from scRNA-seq data. TraCeR [227] and BraCeR [144] are two pipelines that reconstruct TCR and BCR sequences from raw scRNA-seq data making use of various available bioinformatics tools. They use Bowtie2 [129] to align the scRNA-seq reads to custom in silico combinations of all known chosen reference V and J gene alleles taken from IMGT. Using Trinity [84], they assemble reads into BCR/TCR contigs. Next, IgBLAST is used for the analysis of assembled contigs. For BCR/TCR expression quantification, they use Kallisto [32], or Salmon [176]. BraCeR additionally uses BLAST to determine the BCR isotype. VDJPuzzle [190] is a similar pipeline that reconstructs productive, full-length BCRs of both heavy and light chains and can extract SHMs on the V(D)J region. It makes use of multiple rounds of alignment and assembly; first to the reference genome to filter reads for initial contig assembly, another to re-align all reads to the assembled contigs in order to retrieve any possible misses followed by a second assembly and a third alignment for error

correction. In the VDJPuzzle pipeline, paired-end reads are trimmed using Trimmomatic [23] and aligned to the reference genome using Tophat [115]. Paired reads with at least one pair aligned to any of the V, D, J, or C genes are extracted using BEDtools [186], and de novo assembled using Trinity [84]. Resulting contigs are matched to IMGT to find complete, productive, and in-frame BCRs. Reads are re-aligned against the primary repertoire using Bowtie2, and aligned reads are used for assembly of a second round of contigs with Trinity. Contigs corresponding to a BCR from the primary repertoire are merged. For error correction, paired-end reads are re-aligned a third time, to the merged repertoire using BWA [139]. A semi de novo pipeline called BASIC reconstructs a BCR sequence by anchoring reads to the V and the C genes and then extends the sequence by progressively concatenating overlapping reads to the anchored sequence [42].

BALDR [243] is another BCR reconstruction pipeline, and similar to BraCeR makes use of seven different bioinformatics tools and performs de novo assembly of RNA-seq reads. The one tool that is the workhorse of all these pipelines is the Trinity assembler [84]. Trinity constructs and traverses a set of de Bruijn graphs [2] where each graph represents the transcriptional complexity at a given gene or locus and then processes each graph independently to fully reconstruct a significant fraction of the transcripts. TRAPeS is a TCR reconstruction tool which (i) for each chain, identifies the V and J segments by searching for paired reads with one read mapping to the V segment and its mate mapping to the J segment, (ii) identifies a set of putative CDR3-originating reads as the set of unmapped reads whose mates map to the V, J, and C segments, and (iii) constructs the CDR3 region with the putative CDR3 reads [5]. BRAPeS is an extension of TRAPeS which reconstructs a BCR in two extra steps: in addition to TRAPeS (i)-(iii) steps, BRAPeS determines the BCR isotype by running RSEM [136] on the reconstructed sequence with all possible constant segments added to it and then corrects for SHMs [4].

[2] De Bruijn graphs, named after Nicolaas Govert de Bruijn, are directed graphs that represent overlaps between sequences of symbols [54]. Every possible $(k-1)-mer$ of the given symbol sequence is assigned to a node and edges connect any two nodes that have a consecutive perfect overlap of $(k-2)-mer$, which means that the two nodes are the prefix and the suffix of a *k-mer* from the input sequence. The edges correspond to this *k-mer*. De Bruijn graphs can be used for the *de novo* assembly of sequencing reads by determining which reads should be concatenated to form sequence contigs. Possible contigs are produced by traversing a graph of sub-reads (sequences of nucleotides) of length L that overlap consecutively by $L-1$ bases. For example, in transcriptomics, each path in a de Bruijn graph would represent a possible transcript.

1.4 Aims of this thesis

The overall aim of this thesis is to investigate the anti-tumor immune response to better understand adaptive immunity in cancer. I focused on B and T cell antigen receptors as they guide the adaptive immune response.

In the scRNA-seq human melanoma study presented in Chapter 2, I aimed to incorporate AgR detection with scRNA-seq in order to investigate the clonal adaptive immune response in melanoma. By extracting both gene expression and AgR clonotype data at the single-cell level, I aimed to find the possible subpopulations of immune cells driving the intratumoral immune response.

By analyzing the bulk RNA-seq colorectal cancer (CRC) and pancreatic ductal adenocarcinoma (PDAC) dataset presented in Chapter 3, I focused on the tumor infiltration of B cells, specifically in CRC and PDAC. I aimed to study the effect of a particular immunotherapy drug on the B cell responses towards these cancers. Using paired data from a clinical trial, I aimed to find immune receptors expanded after therapy.

In the scRNA-seq mouse melanoma study presented in Chapter 4, I aimed to link the TCR repertoire with whole-transcriptome scRNA-seq data. Combining the two at a single cell level allowed me to investigate the relationship between T cell phenotypes and TCR repertoires. By using a novel transgenic mouse model that can help track the clonal T cell response to tumor antigens, I investigated the clonality and diversity of tumor-reactive TCRs.

1.5 Overview of work done

In this thesis, I investigate certain aspects of the immune system in the context of cancer by computationally reconstructing and analyzing intratumoral B and T cell receptors from both bulk and single-cell RNA-seq data. In the remainder of this thesis, RNA-seq will refer to bulk RNA-seq unless otherwise noted.

1.5.1 Chapter overviews

In Chapter 2, applying a novel single-cell immune profiling protocol on original data obtained from melanoma patients, I present a complete computational reconstruction of immune receptors. Patient biopsies were collected prior to immunotherapy, and their repertoire analysis depicts the baseline immunological status of patients. With the abundance of plasma cells and the fully class-switched antibodies in these biopsies, I show the presence of an ongoing integrated immune response which involves B and T

cell cooperation. In this dataset, single-cell RNA-sequencing results are consistent with the presence of tertiary lymphoid structures.

In Chapter 3, using a clinical trial dataset of paired biopsies, I show that B cell infiltration increases after immunotherapy in pancreatic and colorectal cancers. I then present the sequences of post-therapy clonal antibodies which I computationally reconstructed from RNA-seq data.

In Chapter 4, I focus on T cells and TCRs. I describe our approach that combines single-cell gene expression profiles with TCR information at the cellular level in order to uncover the links and transitions between transcriptional states. By linking T cells via their TCRs, I show the differentiation of tumor-infiltrating T cells caused by the interplay between the tumor and immune response. I present a novel transgenic mouse model that can help track the clonal $CD8^+$ T cell response to tumor antigens. I describe how, using this mouse model, we were able to harvest and sequence tumor-reactive T cells, and how we obtained their gene expression profile and TCR repertoires. I give a detailed account of my analysis on the obtained dataset from the single-cell RNA sequencing of the mouse model, focusing on investigating the clonality and diversity of T cells.

1.5.2 Original contribution

In Chapter 2, I processed a novel, linked single-cell RNA-seq and AgR repertoire dataset of two melanoma patients and identified paired heavy and light antibody chains which could be further analyzed for antigen binding. The paired nucleotide sequences are presented in Appendix A.

In Chapter 3, I performed a complete BCR repertoire analysis of 37 PDAC and CRC patient samples taken before and after a new type of immune treatment [65]. To the best of my knowledge, BCRs have not been reconstructed and analyzed from tumor tissue of PDAC, which is one of the most aggressive and lethal forms of cancer. With this clinical trial data, I showed a snapshot of the adaptive immune response pre and post immunotherapy, and showed increased BCR clonality and expansion post-treatment. I further identified clonal BCRs which could potentially be used to identify antigens that drive PDAC and CRC intratumoral immune responses. The possible pairings of nucleotide sequences are presented in Appendix A.

In Chapter 4, I presented a combined TCR and gene expression analysis of tumor-reactive T cells at an unprecedented level of detail. This work also confirms the usability of the presented novel mouse model in AgR signaling experiments.

1.5.3 Other PhD work

I began my PhD in September 2015. Between October 2015 and December 2017, I worked on developing a method, namely *CamSeq*, to sequence concatenated amplicons on the Oxford Nanopore's MinION sequencer. I worked with Dr. James Hadfield and Dr. Sarah Field to design and implement an amplicon sequencing method to detect low-frequency alleles in cancer samples using the MinION sequencer. For this purpose, I wrote a pipeline that takes in Nanopore MinION sequencing reads and outputs possible mutations. I also developed a signal detection tool using time-point change detection algorithms to call mutations directly within the signal space rather than on the base-called sequences. This work was in collaboration with AstraZeneca and has been handed over to their labs to improve further.

Between December 2016 and September 2018, together with Dr. Daniele Biasci, I worked on standardizing the application of existing methods in liquid biopsy data analysis [91]. I developed *Varbench* [92], a suite of tools that compare somatic variant callers for targeted deep sequencing data, run the best performing one, and report the allele frequencies of the called variants using estimate intervals. Varbench comprises four pipelines, namely `varbench:simulate`, `varbench:compare`, `varbench:estimate` and `varbench:validate`. `varbench:simulate` and `varbench:compare` pipelines allow users to systematically asses, compare, and choose the best combination of tools for the data at hand. The `varbench:estimate` pipeline is designed to robustly estimate and quantify uncertainty for allele frequencies, with particular attention to low-frequency mutations. Importantly, the estimates obtained can be used to detect changes during longitudinal ctDNA monitoring in patients. Finally, `varbench:validate` assists the user in selecting mutations for downstream validation.

Chapter 2

Characterization of the melanoma TME and antibody repertoire

2.1 Introduction

Although melanoma accounts for only 1% of skin cancers, it makes up the majority of skin cancer-related deaths [8]. Immune checkpoint blockade (ICB) therapies have drastically increased life expectancy in melanoma patients [130], however still a considerable number of patients do not show a positive outcome post-treatment. Melanomas are considered as immunologically "hot" tumors, meaning they harbor high levels of tumor-infiltrating T cells, resulting in an inflamed TME, which increases the effectiveness of checkpoint blockade immunotherapies [72]. Hence it is essential to study the TME of melanoma at baseline at single-cell level in order to understand the mechanisms of anti-tumor immunity prior to immunotherapy.

Majority of theories on inflamed TMEs focus on T cell responses. In this study, I investigated the presence of tumor-infiltrating B cells (TIL-Bs) in melanomas, characterizing them based on both functional phenotypes and clonality. I achieved this by linking the whole-transcriptome scRNA-seq data with the BCR repertoire at a single-cell level. In Section 2.2, I review the role of B cells in tumor immunity and immunotherapy, and explain the mechanisms via which diverse BCR repertoires are generated. I further describe how a novel protocol, The Chromium Single Cell Immune Profiling Solution, allows the gene expression profile, and full-length, paired, BCR and TCR clonotypes to be obtained from the same input sample [269]. In Section 2.3, I present the two melanoma samples that I used for this study, and in Section 2.4 describe how they were pre-processed with the aforementioned protocol. Here, I also point out the shortcomings of the protocol's analysis pipelines and how they can be further improved. In Section 2.5,

I give a detailed account of my analysis on the obtained scRNA-seq datasets, while in Section 2.6, I present the antigen receptor repertoires obtained again from the same datasets. In Section 2.7, I present my findings which include the sequences of clonal intratumoral antibodies, and conclude this chapter with a discussion in Section 2.8.

2.2 Background

2.2.1 Anti-tumor immunity from a B cell perspective

Very recently, the five-year follow-up of the combined nivolumab-plus-ipilimumab immunotherapy trial in advanced melanoma reported that the overall survival at five years was 52% in the nivolumab-plus-ipilimumab group, compared to the 44% in the nivolumab and 26% in the ipilimumab group [130]. While only a decade ago, patients diagnosed with late-stage metastatic melanoma would survive at most six months [43], now, with the advances in immunotherapies, the five-year survival rate is one in two patients. However, a significant number of patients still show innate or acquired resistance to immunotherapies [169]. The study of the TME prior to treatment can provide insight into patient response.

The role of the T cell-mediated adaptive immunity in the anti-tumor immune response is well established [235], [13], [202], [197], [140], and is the focus of successful cancer immunotherapies [189]. It has been widely accepted that patients with melanoma tumors infiltrated with T cells, the so-called “hot tumors”, have better long-term survival. The role of B cell-mediated humoral adaptive immunity is less clear and has only recently gained attention [241], [254].

B cells have been shown to infiltrate the tumors of melanoma, breast, and colorectal cancers, among others [128], [151], [216]. In some tumors, B cells make up 25% of all the intratumoral cells, and 40% of TILs are TIL-B in some breast cancer tumors [260]. B cells are a heterogeneous population with functionally discrete sub-types. Depending on their phenotype, intra-tumoral B cells may promote or inhibit anti-tumor immunity. Studies of TIL-Bs have shown associations with positive prognosis [56], [141], [251]. B cells might positively contribute to anti-tumor immunity through the production of antibodies which target tumor antigens, and of cytokines and chemokines which can recruit other immune cells, or via presenting antigens and partaking in other immunoregulation mechanisms [241], [141], [251], [254]. On the other hand, there are cases where TIL-Bs result in the progression of the tumor. For example, [185] demonstrated that IL-35 secreting TIL-Bs were associated with facilitating tumor cell proliferation in pancreatic cancer, and [213]

showed that tumor-infiltrating plasma cells which express IgA, interleukin (IL)-10 and PD-L1 suppressed the cytotoxic function of T cells induced by chemotherapy in three different mouse prostate cancer models. Both [185] and [213] were conducted using mouse models.

In a clinical study, $CD20^+$ B cells were detected around $CD8^+$ T cells in the tumor stroma, demonstrating a positive prognosis in ovarian cancer [167]. B cells might be acting as APCs to T cells, promoting intratumoral T cell proliferation [260], thus indirectly contributing to anti-tumor immunity. Furthermore, [123] described the positive impact of tumor-infiltrating plasma cells on the anti-tumor immune response by showing that $CD8^+$ TILs had increased prognostic benefits when detected alongside plasma cells, $CD20^+$ TIL-Bs, and $CD4^+$ TILs, suggesting cooperation of B and T cells within the TME to promote anti-tumor immunity. Additionally, a recent study showed that tumor-associated B cells are crucial to melanoma-associated inflammation as they may be guiding T cells and other immune cells to the tumor [86]. The same study, using scRNA-seq, also showed that $CD19^+CD20^-CD38^+CD138^-$ plasmablast-like and naive-like B cell frequencies to be significantly higher in the pre-immunotherapy melanomas of patients responding to anti-PD-1 therapy. B cells are also relevant in ICB therapies. It is not just T cells that express PD-1, PD-L1, and CTLA-4. When either PD-1 or CTLA-4 is blocked, the proliferation of memory B cells and antibody production increase [260].

B cells can secrete lymphotoxin (LT) which has been reported to be tumor-promoting in some contexts [10] but also to play an important role in the formation and maintenance of TLSs [201]. While TIL-Bs revealed contradicting results in human and mouse models, the number of TLSs showed a positive association with favorable outcomes in both human disease and mouse models. Various human cancers, including melanoma, have reported positive outcomes associated with a high number of TLSs. Additionally, it has been shown that high amounts of B cells within TLSs predict longer patient survival in lung cancer [77].

In summary, B cells might play a beneficial role in anti-tumor immune response with the antibodies they secrete, and by cooperating with T cells within the TME, as well as possibly by aiding in the development and maintenance of intratumoral TLSs. However, the role of B cells in the anti-tumor immune response needs to be further studied in human cancers, paying attention to the differences in molecular B cell sub-types.

2.2.2 Generation of BCR diversity

The vast BCR diversity is generated by a series of complex processes [117, Chapter 5, p. 117-122]. During the maturation of B cells, the BCR genes are rearranged from many different possible gene alleles to form a complete chain. During this V(D)J recombination step, the receptor is tested for functionality and eliminated when it shows self-antigen reactivity in order to prevent autoimmunity.

As described in Section 1.1.1, matured naive B cells get activated when they encounter an antigen and travel to the germinal centers in the secondary or tertiary lymphoid structures in order to gain further affinity via additional diversification processes. This section will discuss in detail the different processes which contribute to BCR diversification.

Pre-antigen-challenge diversification

Before antigen encounter, the BCR repertoire of a human can generate more than 10^{12} unique Ig molecules, referred to as their *preimmune antibody repertoire* [34, Chapter 24, p. 1319].

Combinatorial diversity is achieved via somatic DNA recombination which takes place in the bone marrow. The different Ig loci which contain the gene fragments that concatenate to make up the different variable domains are located on different chromosomes (Chr) in the human genome: IGH is on Chr14, IGK on Chr2, and IGL on Chr22. The gene segments consist of different germline sequences. There are also several pseudogenes which result in nonfunctional variable regions when undergoing recombination. The number of gene segments are presented in Table 1.1. The recombination steps of the heavy and light variable regions follow a specific order which is shown in Figure 2.1. I have illustrated this V(D)J recombination process based on details provided in [34, Chapter 24, p. 1316-1324].

Junctional diversity happens due to imprecise DNA joining. V(D)J joining is mediated by RAG1 and RAG2 (recombination activating genes) recombinase [75]. During the joining of Ig gene segments, individual nucleotides are often deleted from their ends, and one or more random nucleotides may be inserted. This random loss-and-gain increases the diversity of the CDR3 region significantly. However, it can also cause a shift in the reading frame, producing nonfunctional genes [117, Chapter 5, p. 115].

Coupling diversity is a result of the combination of different variable regions of the heavy and light chains. Though thought to be random, it has been reported that pairing preferences do exist for a small proportion of germline gene segments [109].

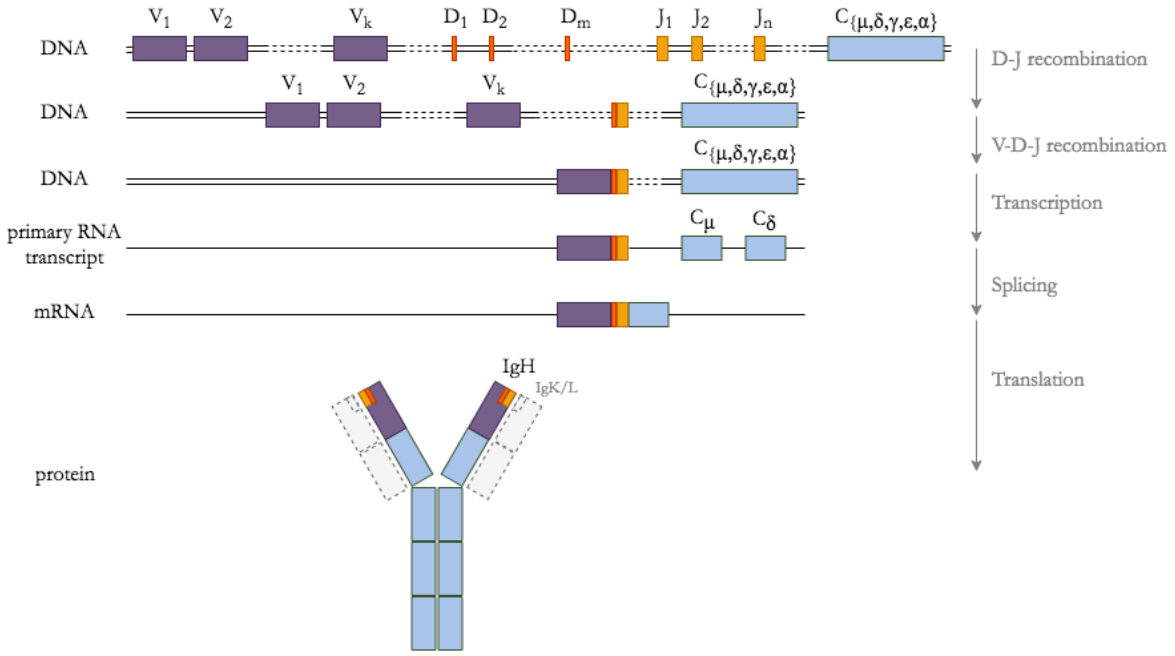


Fig. 2.1 The V(D)J recombination of the immunoglobulin heavy chain from germline gene segments. The construction of the variable domain of the heavy chain begins with the randomly chosen D and J genes' recombination, and then the V gene joins the DJ segment. The constant domain is then joined through RNA splicing of the primary RNA. In immature B cells, the variable domain is transcribed with the constant μ or δ chain, producing either of the mRNAs through alternative splicing, which are finally translated into either IgM or IgD. Heavy Igs are then paired with a light Ig which are shown in dashed lines. I have illustrated this V(D)J recombination process based on details provided in [34, Chapter 24, p. 1316-1324]

Post-antigen-challenge diversification

After encountering the appropriate antigen, B cells may start a process aimed at making high-affinity antibodies in order to produce a more efficient immune response.

Somatic hypermutations (SHMs) occur in the variable domains of the heavy and light chains mediated by the activation-induced cytidine deaminase (AID) enzyme [179], which is expressed and active in GC and TLS-activated mature B cells [231]. SHM takes place in the dark zone of the GCs (see Figure 1.5). The SHM rate is about one nucleotide per 10^3 base pairs per cell division [162, Chapter 10, p. 410]. If mutations lead to a nonfunctional BCR, the B cell might die by apoptosis. Otherwise, it can migrate into the light zone. In the light zone, low-affinity BCR-bearing cells are stimulated for survival, proliferation, and possible re-entry into the dark zone for consequent rounds of affinity maturation, whereas very low-affinity BCR bearing cells die by apoptosis. Final

B cells leave the GC differentiating into antibody-producing plasma cells or memory B cells.

Class switching allows B cells to switch from making one class of antibody to making another in their heavy chains. On stimulation by antigens, some B cells are activated to secrete IgM antibodies which dominate the primary antibody response. Only after antigen stimulation, with the cytokines secreted by helper T cells, do many B cells switch to making IgG, IgE, or IgA antibodies. The differentiated memory B cells and plasma cells express these class-switched antibodies. In an individual's Ig repertoire, a particular antigen-binding site can be distributed among various classes. In a B cell, once a switch has happened from IgM and IgD to one of IgG, IgE, or IgA, the change is irreversible at the DNA level as every heavy constant coding gene sequences between the assembled VDJ sequence and the particular constant heavy coding sequence is deleted [34, Chapter 24, p. 1322-1324].

2.2.3 Integrated immune cell profiling

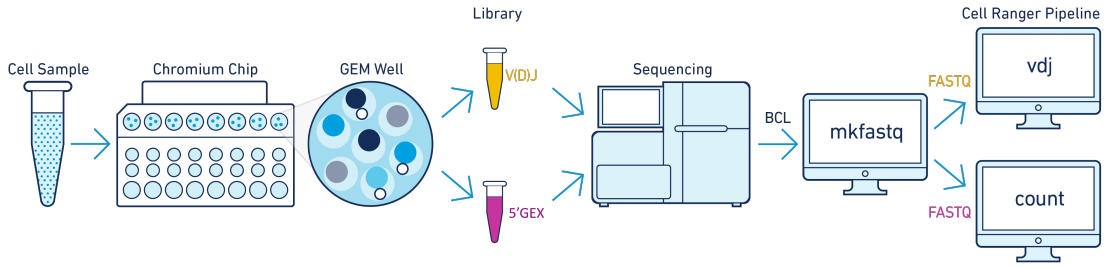


Fig. 2.2 Workflow describing the Chromium Single Cell Immune Profiling Solution. Image adapted from [2].

10x Genomics provides a commercial method to link BCR/TCR targeted sequencing with whole-transcriptome analysis of human and mouse single-cells in parallel employing the droplet-based microfluidic platform Chromium 10x [269]. There are roughly 750,000 unique 10x barcoded gel beads that can index a cell in an input sample. In the Chromium controller, 10x barcoded gel beads are mixed with cells, enzyme, and partitioning oil and captured in droplets making up a GEM; gel bead in emulsion partition (see Figure 2.2). The GEM provides the micro-environment in which cells are barcoded and reverse transcribed into cDNA. Hence, all cDNA within the same cell are labeled with the same barcode. The GEMs are then broken so that all barcoded cDNA are mixed together. The mixed cDNA undergoes amplification and is divided for target enrichment and library preparation of B and T cell V(D)Js, and for 5′ gene expression library preparation.

V(D)J enrichment of B and T cell is performed on amplified cDNA by PCR, which uses reverse-primers designed for the constant regions and a universal forward-primer designed for a sequence added at the 5′ end [1]. The prepared libraries are then RNA-sequenced yielding reads with cell barcodes. 10x provides two pipelines which produce a (gene x cell) expression matrix from the 5′ gene expression library reads, and V(D)J repertoires from the B and T cell enriched libraries. These pipelines are described in Section 2.4.1.

2.3 Human melanoma samples

In this study, I had access to two patient samples. Both patients are enrolled in the MelResist clinical trial, which is an ongoing study investigating the resistance to gene-targeted melanoma therapies [51]. Both patients have metastatic melanoma. The samples were collected prior to immunotherapy.

The patient referred to as *patient1* received radiotherapy prior to sample collection. The sample was taken surgically from this patient's lung metastasis. This patient was operated on as the tumor was compressing the lung [James Thaventhiran, personal communication, August 8, 2018].

The second patient, *patient2*, did not receive any prior treatment. The sample was collected with a needle biopsy [James Thaventhiran, personal communication, August 8, 2018].

2.4 Pre-processing of RNA-seq data

patient1 and *patient2* samples were processed using the 10x Genomics Chromium Single Cell Immune Profiling Solution to produce 10x-barcoded libraries. For sequencing, the libraries were run on the HiSeq 4000 platform outputting 150 x 150 bp paired-end reads.

Prior to Cell Ranger v3.1, which introduced changes to the assembly algorithm[2], the required read lengths for GEX libraries were 26 x 98 bp, and 150 x 150 bp for the V(D)J enriched libraries. When sequencing *patient1* and *patient2* samples, in order to cut the sequencing cost, all libraries were placed on the same lane and sequenced for 150 x 150 bp

[1] It should be noted that the enrichment kits target neither the constant regions of TCR γ and δ chains, nor the IgE antibody. However, it is possible to design custom primers to capture these sequences.

[2] Changes to the assembly algorithm allow V(D)J libraries to be sequenced at 26 x 91 bp instead of 150 x 150 bp. This enables V(D)J and GEX libraries to be sequenced in a single run. So while our sequencing step was cutting costs and simplifying the workflow by increasing the sequencing length of the GEX libraries, 10x Genomics achieved this via improving their assembly algorithm which allowed for the shortening of the V(D)J read configuration.

paired-end reads. Therefore, following sequencing, V(D)J enriched library reads were trimmed to 26 x 98 bp. Read 1 (R1) of GEX reads includes 16 bp of barcode and 10 bp of UMI. Read 2 (R2) of GEX reads includes 91 bp of RNA-seq and 7 bp of the i7 index. R1 of V(D)J reads includes 16 bp of barcode, 10 bp of UMI, 13 bp of switch oligonucleotides, and 111 bp of RNA-seq. R2 of V(D)J reads is all 150 bp of RNA-seq.

The raw data files generated by the Illumina platform are in Binary Base Call (BCL) format. Using Illumina's `bcl2fastq`, BCL files were converted into FASTQ format which is a text-based sequence file format that records both raw sequence data and quality scores. At this stage, for each library, there were multiple FASTQ files, ready to be processed for UMI counting, and V(D)J reconstruction. The two main Cell Ranger pipelines; `cellranger count` and `cellranger vdj` perform these steps, respectively.

2.4.1 The Cell Ranger pipelines

`cellranger count` performs GEX analysis on the RNA-seq data. The pipeline aligns short reads to the specified reference transcriptome using the STAR aligner [57], performs de-duplication of cDNA PCR duplicates, removes barcodes not associated with cellular GEM partitions, and generates (gene x barcode) matrices by counting unique molecules. `cellranger vdj` assembles all reads within a single cell, then annotates the assembled contigs by aligning them to the provided BCR and TCR reference sequences. Alignment is performed by seed-and-extend, where seeds are 12-mer perfect matches. Contigs that are not productive are filtered out. Cells are grouped together into clones if they share the same set of productive CDR3 nucleotide sequences by exact match. Hence, a clonotype is defined as the set of exactly matching CDR3 nucleotides. This definition does not account for B cell lineages arising via SHMs.

With `cellranger count`, I processed the FASTQ files in order to align, filter, and count UMIs and generate a (gene x barcode) count matrix for each sample. I mapped the sequences to the GRCh38 transcriptome.

2.4.2 Improving the Cell Ranger pipelines for accuracy

Cell detection

I processed the *patient1* sample in June 2018, and the *patient2* sample in August 2018, with the then-current Cell Ranger version 2.2.0. The 10x Chromium platform has a cell recovery rate of 65%. With approximately 14,000 loaded cells [3], the expected cell count

[3]For the *patient1* sample, I do not have evidence of the actual cell count being 14,000. This load was requested by the sequencing facility, but the actual numbers were not counted, but for the *patient2*

for both samples was roughly 8K. However, overestimation in cell counting before loading the sample and cell loss during sample preparation is common. Also, achieving a stable cell count from human samples taken from the clinic, and maintaining it, is usually more difficult when compared to cell harvest from mice.

`cellranger count` estimated 3,582 cells for the *patient1* sample, and 993 cells for the *patient2* sample. These numbers were well below the expected count even after considering the possible overestimation and deterioration of cells. Also, for both samples, many of the reads [4] were not assigned to cell-associated barcodes. With droplet-based sequencing, it is possible for some GEMs to have extracellular RNA rather than an intact cell. Therefore we cannot just assume GEMs with positive UMI counts to be valid cells. GEMs containing cells need to be differentiated from GEMs containing extracellular RNA.

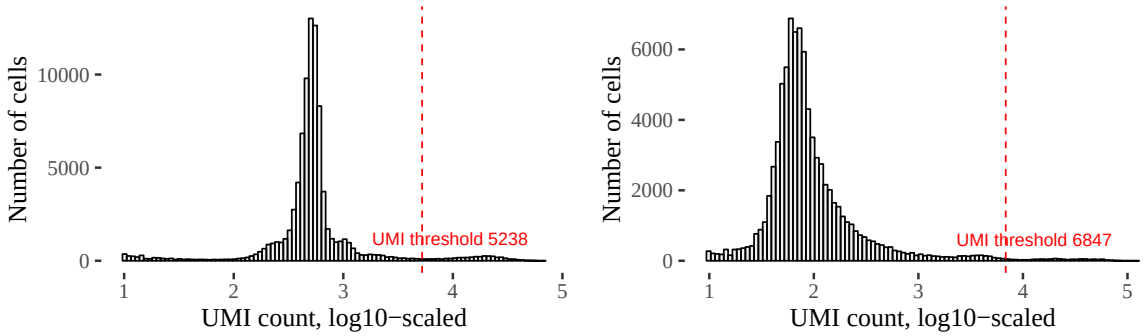
Let N be the expected number of recovered cells (which is 3000 by default and can be specified by the user), and m be the top 1% of N barcodes. Cell Ranger version 2.2.0 cell calling algorithm sorts the UMI count of all barcodes in decreasing order and finds the m^{th} UMI count. The algorithm then determines a barcode as a valid cell if its UMI count is greater than 10% of this count. However, in the context of tumor samples, this assumption that RNA content varies by an order of magnitude among cells does not hold. The TME comprises various type of cells, such as tumor cells, small TILs, large plasma cells, and so on [111]. As the TME contains such cells that can be extremely diverse in RNA content, this assumption would only recover the larger cells and consider smaller cells to be background noise (see Figure 2.3).

Processing the B and T cell enriched libraries with `cellranger vdj` showed the total unique immune cell receptor count to be higher than the total number of detected cells. Moreover, after grouping the *patient1* and *patient2* samples based on cell type, clusters of naive B and T cells could not be detected. The initial cell count was only representative of large tumor cells, and highly active plasma B cells, whose expression levels were quite high.

To overcome this limitation, I initially fit a mixture model to the UMI count to filter out the background noise and also mark a low RNA barcode as a cell if it had a productive BCR or TCR, as identified by `cellranger vdj`. I later discarded the falsely called cells during quality check and clustering for cell-type identification.

sample, the loaded cell count was reported as 14,000 by the sequencing facility [Katarzyna Kania, personal communication, June 20, 2018].

[4]The fraction of confidently-mapped-to-transcriptome reads with cell-associated barcodes were 57% for the *patient1* sample and 64% for the *patient2* sample



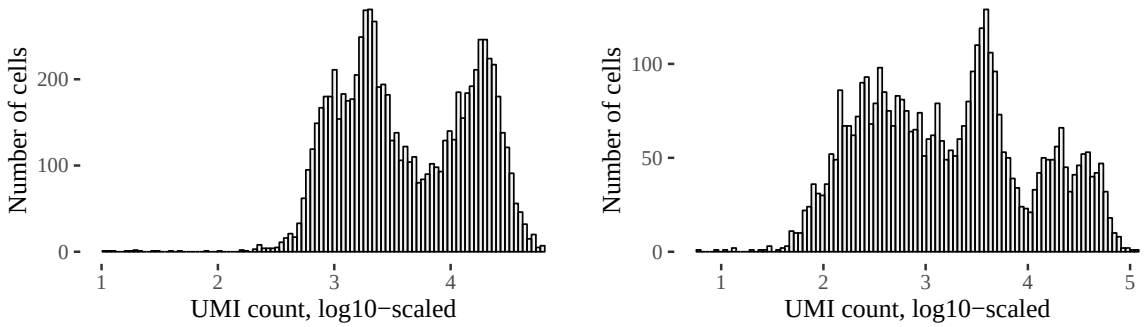
(a) *patient1* sample **(b)** *patient2* sample

Fig. 2.3 UMI count distributions of samples *patient1* and *patient2*. Red dashed lines show the UMI count threshold determined by the Cell Ranger version 2.2.0 cell calling algorithm. Note how only the very large cell population is kept while majority of cell candidates are discarded. Here, only barcodes with a UMI count of at least 10 are plotted.

However, a more elegant approach is presented in [150], where an expression profile for GEMs containing extracellular RNA is identified, and each barcode is tested against this profile and called as a valid cell if it shows a divergent expression profile. This method is not based on the UMI count of a barcode but its expression profile, thus ensuring that small TILs are kept while GEMs full of high levels of extracellular RNA are discarded. Using this method as implemented in `DropletUtils::emptyDrops()`, I obtained 8070 cells for the *patient1* sample and 4241 cells for the *patient2* sample (see Figure 2.4), by specifying barcodes with UMI counts less than 150 and 60 to correspond to empty droplets for each sample respectively. In each sample, barcodes with total UMI counts less than the specified thresholds are used to estimate the profile of the ambient RNA pool, and each barcode is then tested against this profile for deviations in order to determine real cells. A fraction of these reads were later discarded during the QC step, as described in Section 2.5.1.

Incorporating V(D)J reference from the IMGT database

I processed the BCR and TCR enriched RNA-seq data with `cellranger vdj`. Looking at the reconstructed λ chains, we see that the J regions have a large deletion sequence. This is a limitation in the default reference `cellranger vdj` uses. Due to licensing restrictions, it is not able to distribute an IGMT-based reference. To this end, I built an IGMT-based reference downloading the V, D, J, and C sequences from IGMT and used it as a custom reference. Figure 2.5 shows the alignments using the default reference and the IMGT reference. In addition to fixing the deletions, using IMGT has also allowed



(a) *patient1* sample **(b)** *patient2* sample

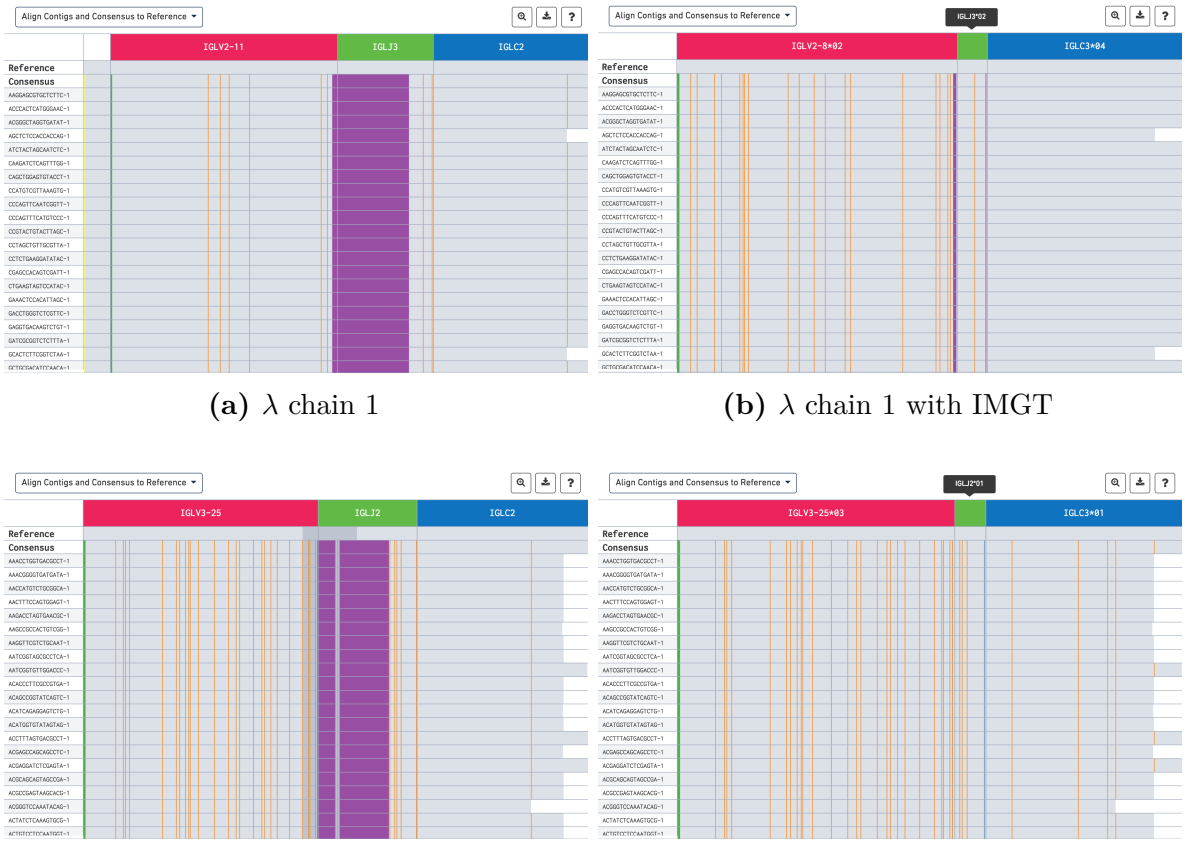
Fig. 2.4 UMI count distributions of the detected cells from the *patient1* and *patient2* samples.

for a better annotation of heavy and light V, D, J, and C segments, and revealed SHMs within the variable domains. However, do note that the use of the custom IMGT reference caused a discrepancy between the GEX and V(D)J gene annotations. In some cases, it is no longer possible to detect the gene expression of an AgR sequence's exact V, D, or J gene directly from the (gene x barcode) UMI count matrix as the annotations are not consistent. However, the gene family names are still the same.

Merging barcode information between GEX and V(D)J output

In version 3.0, Cell Ranger updated its cell-calling algorithm, which is heavily inspired by [150]. This algorithm is further described in Section 4.3.4. Note that, in this study, although I re-processed the samples using the updated pipelines (Cell Ranger Version 3.0.2[5]), I relied on `DropletUtils::emptyDrops()` to call cells. There is a discrepancy between the cells called by the Cell Ranger pipelines `cellranger count` and `cellranger vdj`. Barcodes accepted as cells via `cellranger count` with the updated algorithm are not reflected in the AgR repertoires, i.e., a significant number of V(D)J sequences are discarded claiming they are not reconstructed from real cells. In order to keep the cell detection consistent, I retrieved these productive V(D)J sequences. Also, the targeted cell calling algorithm implemented within `cellranger vdj` does not share information with `cellranger count`, resulting in certain B and T cells going undetected in the GEX analysis. I marked such barcodes as cells if the GEX profile was also consistent with that of a B or T cell. I later inspected the additional B cells in Section 2.5.6.

[5]Cell Ranger Version 3.1 was deployed after the completion of this study



(a) λ chain 1 **(b)** λ chain 1 with IMGT

(c) λ chain 2 **(d)** λ chain 2 with IMGT

Fig. 2.5 Screenshots from the Loupe VDJ browser showing alignments using the default reference (a and c) and the IMGT-based reference (b and d). Here each row represents a chain from a single cell, whose barcodes are given on the left. Consensus is the consensus of all the chain sequences, and the reference is the V(D)J sequence to which they aligned to best.

2.5 Gene Expression (GEX) Profiling

In this section, I present the downstream analysis of the *patient1* and *patient2* samples. Within the remainder of this chapter, *patient1* and *patient2* samples refer to cells harvested from each patient's melanoma metastasis in the lung and processed with the Chromium Single Cell Immune Profiling Solution.

2.5.1 Cell quality control and filtering

During the cell detection step, cells are separated from empty GEMs, however, there may still be cells of low quality, such as dying cells, within the population. The three main attributes used for cell quality check (QC) are (i) total UMI number per cell, also referred to as library size, (ii) the total number of expressive genes per cell, and (iii) the fraction of UMI count contributed by mitochondrial genes [105]. Outliers based on these attributes may be removed as low-quality cells. However, the underlying biology must be taken into consideration. For example, it is common in the literature to remove cells with high levels of mitochondrial gene expression based on the fact that this could be an indicator of dying or lysing cells. However, in tumor biopsies, this could be indicative of the underlying biology as metabolic activity increases mitochondrial gene expression. For example, mitochondrial activity may be indicative of activated B cells undergoing class-switch recombination [108]. Additionally, these QC attributes must be inspected together so that specific subpopulations are not removed. While it is common practice to remove cells with low UMIs or a low number of expressed genes, in the TME these can represent dysfunctional immune cells. Figures 2.6 and 2.7 inspect these attributes to determine low quality cells.

Figure 2.7b shows that there are cells in the *patient2* sample which are quite high in mitochondrial UMI count but low in total expressed genes and UMI count. High mitochondrial transcripts and a small library size or total detected genes is most likely indicative of dead or dying cells (see Figure B.1 for plots of the mitochondrial UMI count against the UMI count and uniquely detected genes).

I removed cells with total expressed gene count <100 in the *patient1* sample, and <50 in the *patient2* sample. Rather than placing a strict threshold on mitochondrial UMI percentage, I used the robust Z-score method to detect outliers. Robust Z-score method, instead of using the mean and standard deviation, uses the median and the deviation from the median - median absolute deviation (MAD) - as the median and MAD are robust measures of central tendency and dispersion. MAD is calculated by taking the absolute difference between each point and the median, then calculating the median of

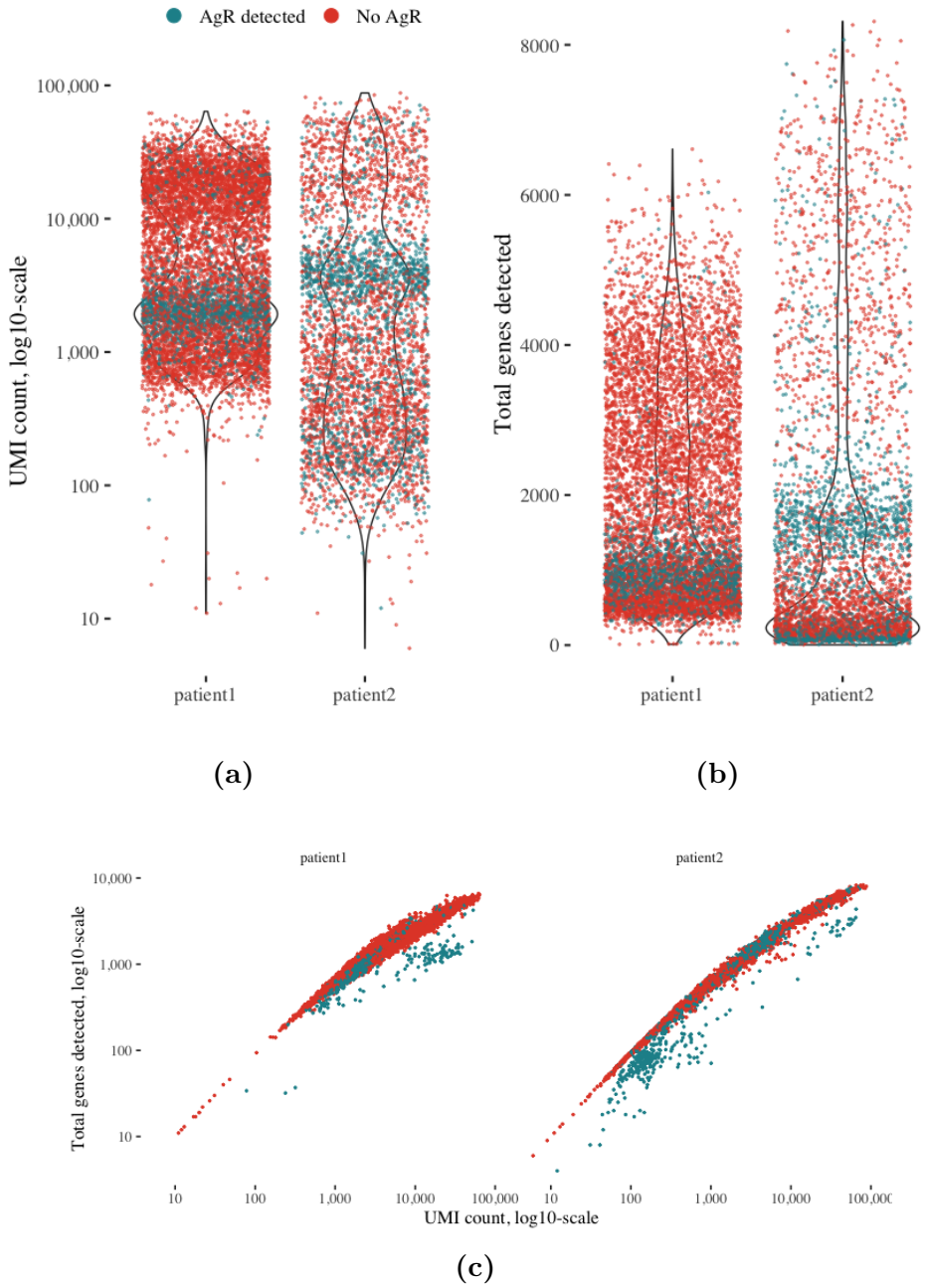
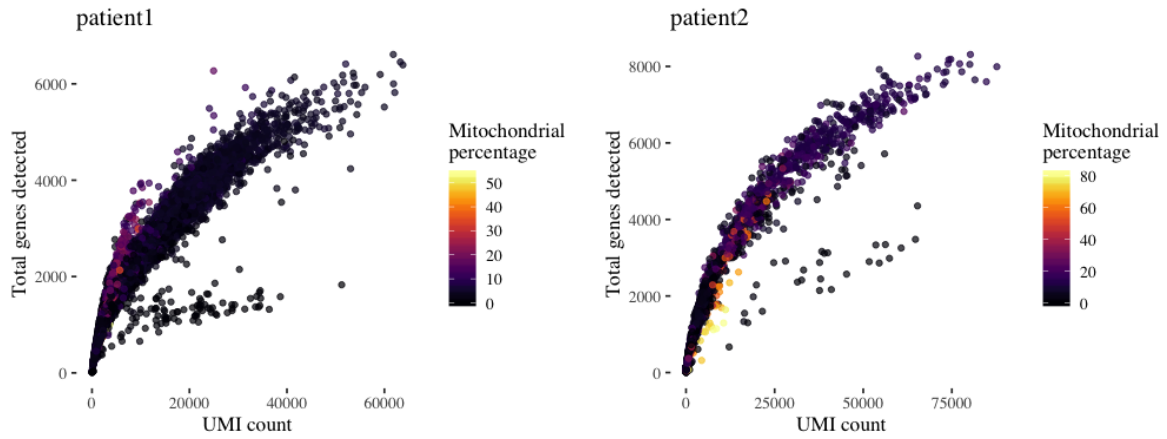


Fig. 2.6 (a) shows the library size distributions and (b) the detected gene distributions of the *patient1* and *patient2* samples, while (c) plots them against each other to show cell complexity. Each dot represents a cell. Color denotes whether an AgR was detected in a cell or not.

the differences. I determined outliers as cells having mitochondrial content more than 3 MADs away from the median. This removed cells with mitochondrial UMI percentage >10.4% and >21.5% from the *patient1* and *patient2* samples respectively, resulting in



(a) *patient1* sample **(b)** *patient2* sample

Fig. 2.7 For each sample, plots show the total number of detected genes versus the UMI count colored by the percentage of UMI originating from mitochondrial genes. Cells with extremely high mitochondrial percentage also have very low UMI counts and numbers of detected genes, meaning these are dead/lysing cells. Without placing a threshold on the UMI count or number of expressed genes, these are removed directly by filtering out high mitochondrial percentage.

an approximately 10% cell count decrease in each sample. High mitochondrial UMI percentages have been reported in tumor samples [160].

2.5.2 Normalization

The UMI count of a single gene in a single cell is dependent on the successful capture, lysis, reverse transcription, and sequencing of an mRNA molecule. When sufficient sequencing saturation[6] is realized, the UMI count can be a direct measure of the cDNA molecules associated with that gene [226]. However, this does not account for the technical variation that may have arisen during the mentioned sampling processes. Normalization methods such as counts per million, where UMI counts are normalized for each cell by the total UMI count and multiplied by a scale factor (1M in this case), aim to address this issue [244]. This method assumes that all cell types include a similar number of mRNA molecules and that count depth variation is only due to sampling. However, in a highly heterogeneous population like the TME, this assumption does not hold. The pooling-based method described in [148] accounts for this cellular heterogeneity. Cells are pooled together based on similar library sizes, and expression values are summed

[6]Sequencing saturation is the measure of the fraction of the total number of different transcripts in the final library that was sequenced. The inverse of this measure would be a proxy for the number of additional reads it would take to detect a new transcript. As sequencing depth increases, so does the number of observed genes, but then this reaches a saturation point where no new genes are detected.

across all cells in the pool, resulting in fewer UMI counts of zeros. The pools are then normalized against a reference which is obtained by averaging all cells in the dataset. This gives per-pool size factors. Repeating the same process for many different cell pools constructs a linear system of equations that can be solved to obtain the per-cell size factors. `scran::computeSumFactors` implements this method [149].

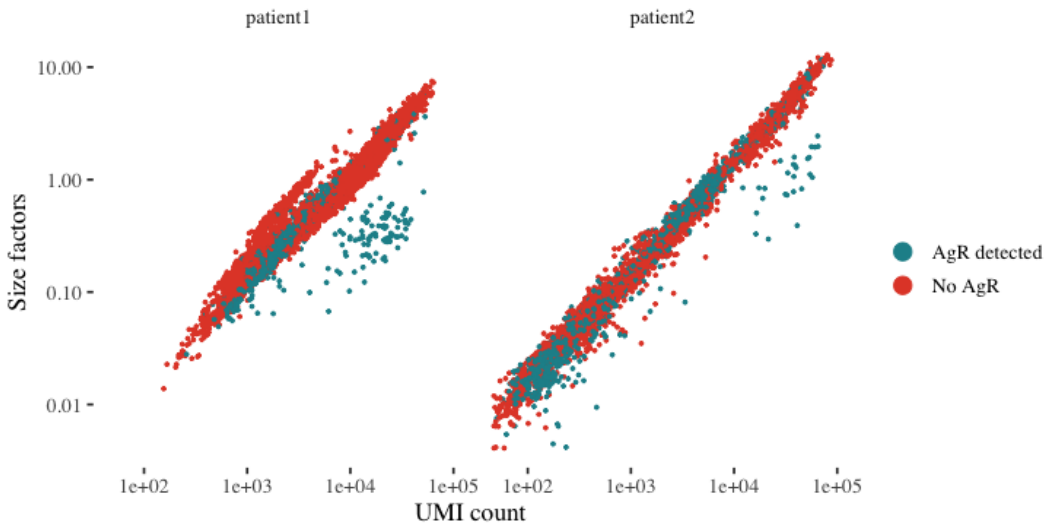


Fig. 2.8 Size factors plotted against the UMI count show that they are well correlated except for some of the AgR expressing cells in the *patient1* sample. This was also evident in Figure 2.7a.

I used `scran::quickCluster` to pre-cluster the samples so that cells in a cluster have similar expression profiles, and later used `scran::computeSumFactors` to normalize cells in each cluster separately. `scran::computeSumFactors` re-scales size factors for comparison between clusters. Finally, I used `scater::normalize` to normalize the count data using the computed size factors, and log2 transformed the normalized data.

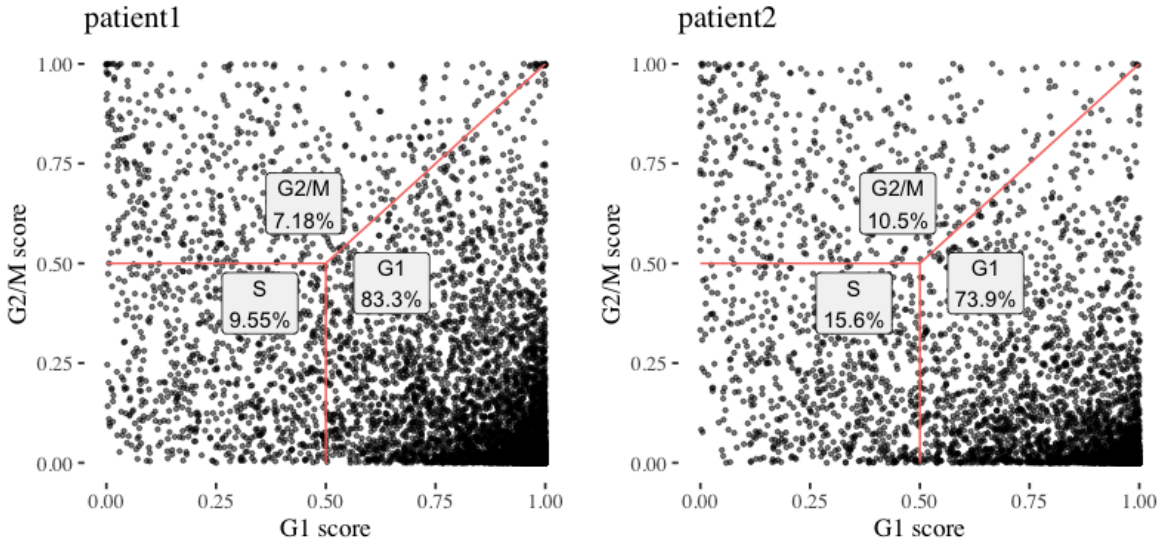
2.5.3 Determining cell-cycle stage

Cell-cycle is a biological effect which can introduce heterogeneity within the same cell type [37]. During the cell-cycle, in the G1 stage cells increase in size, replicate their DNA in the S stage, and in the G2 stage they continue to grow and get ready to enter the M (mitosis) stage in order to divide into daughter cells. Two cells of the same type may be at different cell cycle stages and thus have different expression profiles causing these cells to be clustered separately. It is thus essential to determine the cell cycle stage of each cell. To this end, [207] have developed the method *Pairs* to infer the cell cycle stage of a cell directly from its gene expression profile. Their classification algorithm, using a

training dataset, found pairs of genes such that their difference was positive in one phase while negative in others, i.e., one gene is more highly expressed than the other in one phase, whereas it is lowly expressed in the other phases. As these genes present a specific behavior across phases, they are selected as marker pairs, which can classify cells into phases. Then, using these pairs, cells in the test dataset are assigned probability scores for phases G1 and G2M.

For each cell, when either G1 or G2M score is greater than 0.5, it is assigned to the phase with the score higher than 0.5, when they are both less than 0.5 it is assigned to phase S, and when they are both greater than 0.5 no assignment is carried out.

A slightly altered version of this method is implemented in `scran::cyclone` with an existing pre-trained set of marker pairs for human data. Using this function, I assigned cell cycle stages to cells based on their normalized gene expression data (see Figure 2.9).



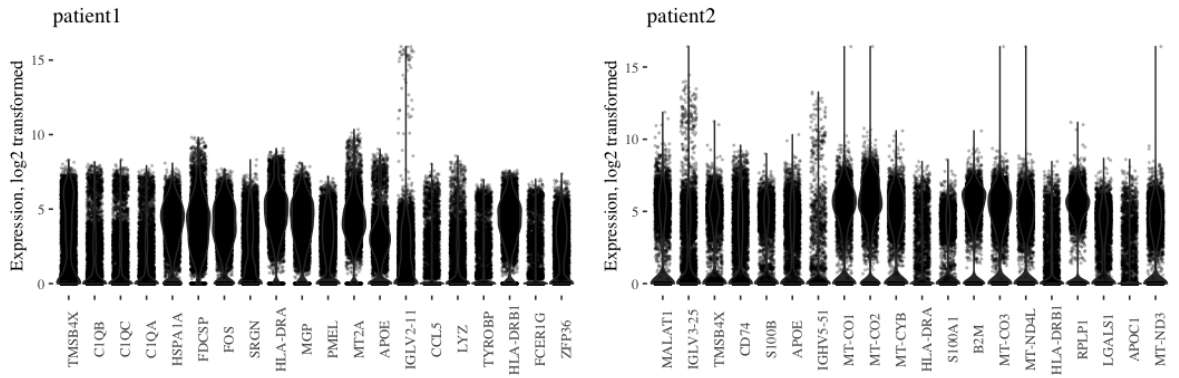
(a) *patient1* sample **(b)** *patient2* sample

Fig. 2.9 Cell cycle phase scores of samples are plotted. Each point represents a cell. Red lines separate cell phases and the proportion of cells in each phase is labelled.

It is possible to correct for the effects of cell-cycle stage using various methods [37], [38], however cell-cycle stage can be informative of the underlying biology. The choice should be made based on the research question and the required downstream analysis. For example, correcting for cell cycle effects can reveal further differentiation trajectories [37], but keeping the cell cycle information can help identify proliferating subpopulations. In this study, I did not remove the cell cycle effects, but having identified the cell cycle stage, I was able to identify separate clusters of the same cell type.

2.5.4 Dimensionality reduction using PCA

For each of the samples, there are >20,000 genes with a positive average expression across all cells. However, not all of these genes will be informative in the downstream analysis. Highly variable genes (HVGs) are genes that reflect the variability in the data. In order to select HVGs, I used the `scran::decomposeVar` function which decomposes the variance of a gene into biological and technical components. Based on [149], I assumed that the technical component follows a Poisson distribution. The biological component is then computed by subtracting this from the total variance. HVGs are determined as genes with positive biological components.



(a) *patient1* sample **(b)** *patient2* sample

Fig. 2.10 Normalized, log2-transformed expression values of the 20 HVGs with the highest biological components in each sample. Each point represents a cell.

In the *patient2* sample, mitochondrial (MT-*) and ribosomal genes (RPL*/RPS*) seem to have large biological components. The low expression of ribosomal protein mRNA can signify cells that are in atrophy or that the cell's cytoplasm is damaged, and usually in scRNA-seq data analysis these cells are discarded. However, in the context of cancer these ribosomal genes can be determinant of normal and malignant human cell types [88]. There are studies which have revealed that certain ribosomal genes act as tumor suppressors while some contribute to tumor growth [230], [116]. In this study I did not remove cells based on ribosomal expression content.

Gene selection reduced the gene count to half. However, the expression data can be further reduced. scRNA-seq data is inherently low dimensional owing primarily to the co-regulation of genes [97]. Hence, the high dimensional (gene x cell) matrix can be represented in a lower-dimensional space. To summarize the gene expression data, I used PCA as implemented in `scran::denoisePCA` where PCs linked only to biology are identified and kept while discarding those corresponding to technical noise.

I excluded the 13 protein-coding genes located on the mitochondrial chromosome (MT-ATP6, MT-ATP8, MT-CO1, MT-CO2, MT-CO3, MT-CYB, MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5, MT-ND6) the ribosomal genes (RPL*/RPS*), Ig genes IGH*/IGK*/IGL*, and the TCR coding TRA*/TRB* genes from the PCA as I did not want these genes driving the clustering directly.

2.5.5 Graph-based clustering using community detection

In order to determine cell identity, it is common practice to group cells together based on the similarity of their gene expression profile. The resolution of this grouping is based on how detailed the cell identity should be. For example, adaptive immune cells can be subgrouped as B and T cells, and the T cells can be further separated as $CD8^+$ cytotoxic T cells, $CD4^+$ helper T cells, and $CD4^+$ $FOXP3^+$ Tregs. In parallel, the same cell types can be in different states. For example, in the TME, $CD8^+$ T cells can be in a gradient of dysfunctionality states [140], which can be further partitioned into multiple groups. Furthermore, not all groups will represent true cell identity; partitioning may also depend on various biological or technical variations, such as cells with high mitochondrial content, cells in a specific cell cycle phase, or cells coming from the same batch, and so on. Therefore, the resolution and the drivers of separation should be determined based on the research question. In this study, I aimed to identify B and T cells, and more specifically, the differentiation of B cells into plasma cells, and the abundance of these plasma cells.

In order to characterize the heterogeneity of the cell population and group the cells based on their gene expression profile without any user input, various unsupervised clustering algorithms have been employed, including k-means clustering, hierarchical clustering, and graph-based algorithms [119]. In this study, I used a graph-based method. A graph is a collection of vertices (or nodes) that are connected by edges. The cell population can be represented as a weighted-graph where each node is a single cell, and cells similar in terms of their gene expression profile are connected with edges that convey the strength of this similarity with associated weights. When the cell population is represented as a similarity graph, clustering can be then achieved by finding highly intraconnected but well-separated subgraphs.

To build this graph, I used a shared nearest neighbor (SNN) method [255] which captures the similarity between two nodes in terms of their connectivity in the neighborhood. First, a similarity matrix between cells is computed using Euclidean distances in their gene expression profiles. Next, for each cell c_i, its k nearest neighbors are listed using the similarity matrix, with c_i as the first entry in the list. An edge $e(c_i, c_j)$ is drawn

between c_i and another cell c_j, only if they have at least one shared neighbor in their list. The weight of the edge e(c_i, c_j) is equal to k - r, where r is the minimum averaged sum of ranks in their nearest neighbor lists for any shared neighboring node. So cells that are genuinely from the same cluster will have smaller r and hence highly weighted edges, as the ranking of their shared neighbors are expected to be high. High-dimensional data is usually sparse, and therefore the similarities measured by conventional distance metrics tend to be low between nodes [19]. SNN takes the primary similarity, and using ranks preserves its meaning in high-dimensional space. I applied this method to each sample, computing the similarity matrices using Euclidean distances on the reduced PCA space obtained previously via `scran::denoisePCA`, and specifying `k=10` as the number of nearest neighbors to consider when building the graph.

Having constructed an SNN graph to represent the cell population, clustering is now about finding subgraphs. In graph theory, one approach in finding subgraphs is to look for groups of nodes that are heavily connected within, but sparsely connected to the rest of the network [69]. Such intraconnected nodes are referred to as "communities". A good number of community detection methods have been developed where the common goal is to better identify meaningful communities while maintaining a reasonable computational complexity. [69] and [256] provide a review and a comparative analysis of some of the well-known community detection algorithms to date. WalkTrap [182] is a node-similarity based algorithm where, at each time step, an agent moves from a node v_i to a linked node v_j by randomly picking a neighbor of v_i. The idea is that these random walkers tend to get "trapped" in a community. If v_i and v_j are in the same community, the probability of getting to a third node v_k in the same community with a random walk should be similar for both nodes. Another class of community detection algorithms, namely modularity-based community detection, aims to maximize the assignment quality score called *modularity*, which evaluates the connectedness of nodes in a community compared to their connectedness in a random network. The Louvain algorithm [78] is an efficient modularity-based algorithm that performs well with large networks. Initially, each node is considered a community on its own. Step by step, nodes are moved to other communities so that the highest contribution to modularity is achieved. When no further improvement is possible, the network is aggregated such that each community is considered to be a node on its own, and the same improvement process is repeated. The algorithm stops either when only a single node is left, or the modularity cannot be further increased.

I used both WalkTrap and Louvain community detection algorithms as implemented in the functions `igraph::cluster_walktrap` and `igraph::cluster_louvain` in order to

detect subgraphs within the cell network of the *patient1* and *patient2* samples. WalkTrap found more communities when compared to Louvain[7]. WalkTrap was more granular, and the communities it detected were in fact, subcommunities of the ones detected via Louvain, so it was detecting further communities within Louvain communities. As mentioned before, the resolution of clustering is a choice. For example, while Louvain did not separate $CD20^+$ B cells and T cells in the *patient1* sample, WalkTrap detected these as separate communities. However, by inspecting communities for gene expression, incremental subclustering (clustering within communities) can resolve such communities. I chose the Louvain algorithm as the resolution achieved with the algorithm was satisfactory for me to answer my research question, and there was no assignment difference between the algorithms other than the higher resolution (see Figure B.2 for the comparison of community detection on the *patient2* sample.).

Figures 2.11 and 2.12 show the cell frequency in each community as detected by Louvain, as well as the distribution of library sizes, and of mitochondrial and ribosomal UMI fractions. As expected, these vary noticeably among certain communities. This is discussed further in Section 2.8. In the remainder of this chapter, communities will be referred to as clusters or cell clusters.

2.5.6 2D visualizations with Uniform Manifold Approximation and Projection (UMAP)

In order to explore a scRNA-seq dataset, we reduce the data to 2D or 3D and use these reduced dimensions as coordinates on a scatter plot where each point represents a single cell. A straightforward approach is to use the first two reduced principal components, PC1 and PC2, and visualize the scRNA-seq data. However, PCA is a linear dimensionality reduction method meaning it can only capture linear structures in the features but not the complex polynomial relationship between them. There are a number of non-linear approaches to generate reduced dimensions, of which the most widely used for scRNA-seq data visualization are t-SNE [245] and, more recently, UMAP [156].

t-SNE is a probabilistic dimensionality reduction technique that efficiently reveals local similarity in the high-dimensional data and is the most widely used method in scRNA-seq data visualization. However, t-SNE does not capture the global structure of the data; while intraconnected points are projected onto lower dimension preserving their distances, intercluster relationships are not reflected in the lower-dimensional projection [18]. With

[7]WalkTrap detected 20 communities for the *patient1* sample and 27 for the *patient2* sample, whereas Louvain detected 14 and 15 communities for each sample, respectively.

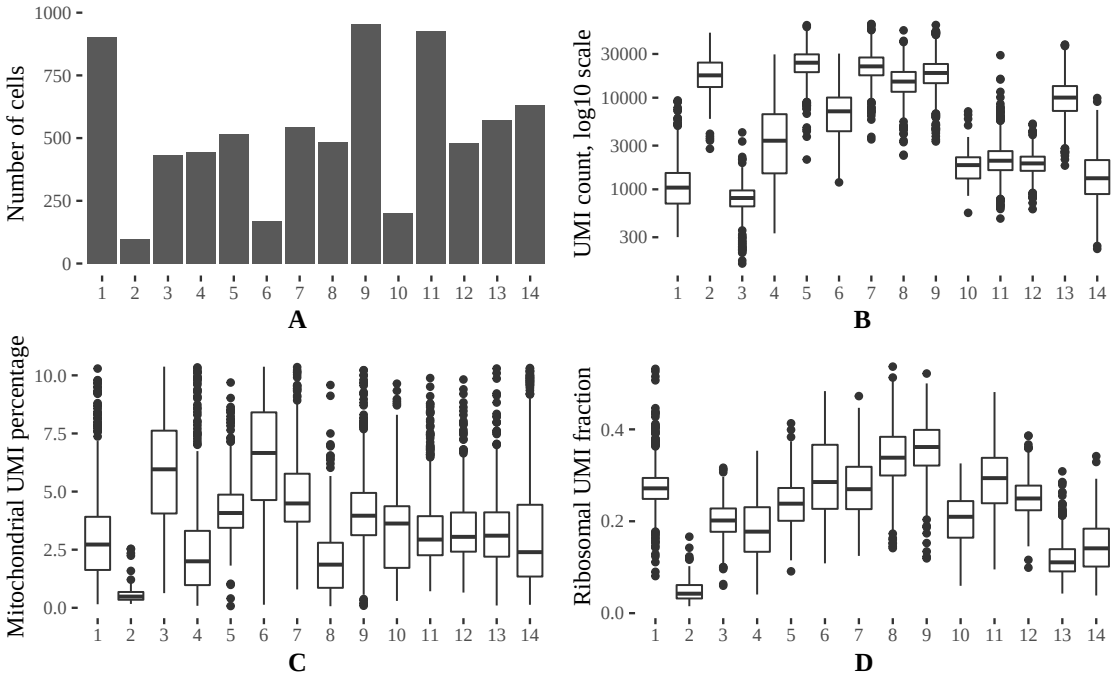


Fig. 2.11 For the *patient1* sample, (A) shows the cell frequency in each cluster, and (B-D) show the distribution of library sizes, and of mitochondrial and ribosomal UMI fractions.

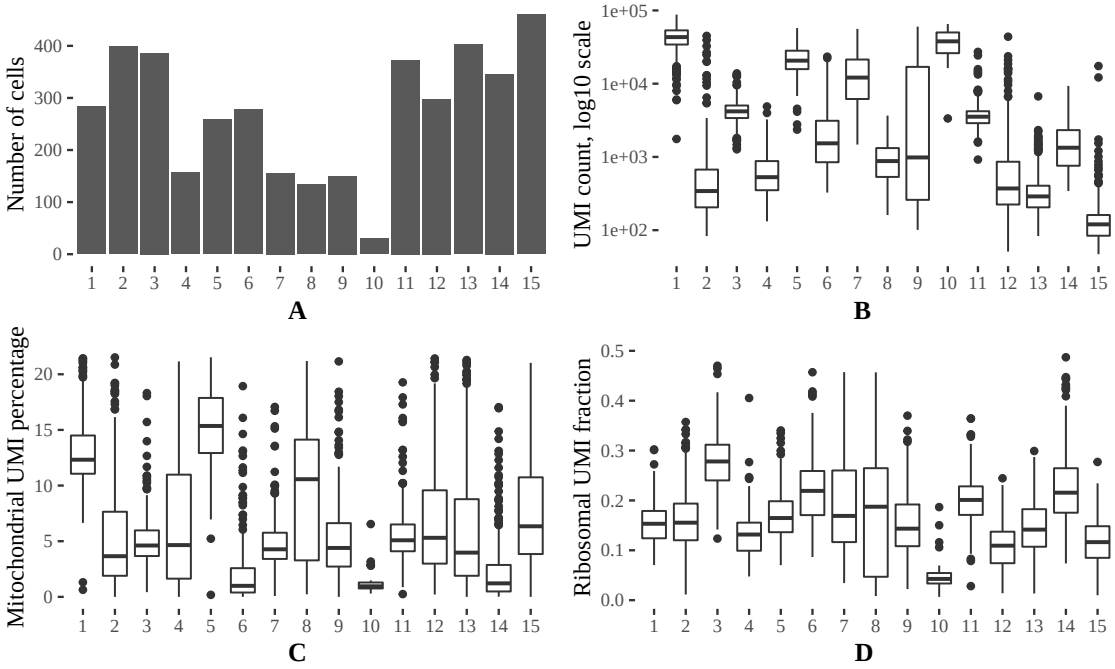


Fig. 2.12 For the *patient2* sample, (A) shows the cell frequency in each cluster, and (B-D) show the distribution of library sizes, and of mitochondrial and ribosomal UMI fractions.

t-SNE, the differences between different cell populations are no longer meaningful in the low-dimensional space.

UMAP is a computationally more efficient alternative that also preserves both the local and the global similarity present within the dataset. By preserving the global structure, UMAP can recapitulate the differentiation stage of different cell types, while t-SNE can only identify such trajectories within an individual cell type cluster [18]. Furthermore, UMAP helps in identifying less obvious cell cluster identities as similar cell clusters will be in closer proximity. In this study, I used UMAP with the prior reduced and denoised PCs in order to speed up the computations and reduce possible noise. Figures 2.13 and 2.14 show the scRNA-seq data projected onto two dimensions with UMAP where each point representing a cell is colored by cell metadata. In this thesis, unless otherwise stated, all UMAP visualizations plot the UMAP estimated first two dimensions.

By reflecting gene expressions onto the 2D projections, and observing the upregulation of known biomarkers in distinct clusters, I determined cluster identity. For less obvious clusters, I performed differential gene expression analysis between clusters with pairwise Welch t-tests as implemented in `scran::findMarkers`. Figures 2.15 and 2.16 show the expression of genes which distinguish clusters, and Figures 2.17 and 2.18 show the UMAP embedded scRNA-seq data where each point is colored by cluster membership, and clusters are annotated with cell identity.

Among some of the cells, I detected unexpectedly high levels of Ig gene expression (see figure B.10). I further inspected these cells for their gene expression profile, and a small number of them in the *patient2* sample showed expression of only Ig genes (see Figure B.11). These are ambient antibody transcripts rather than intact cells.

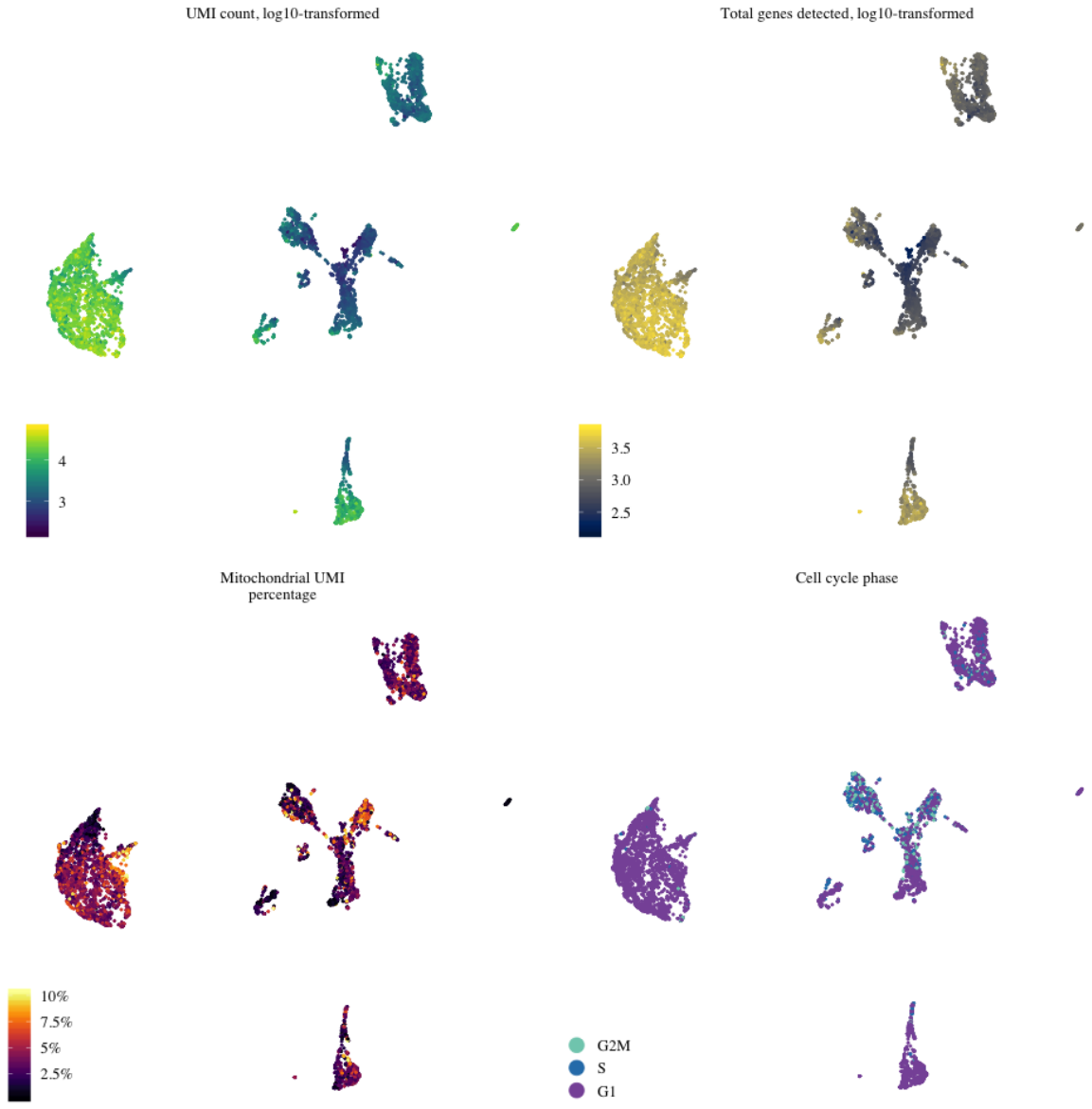


Fig. 2.13 UMAP of the *patient1* sample colored for UMI count, total detected gene count, mitochondrial UMI percentage, and cell cycle phase. Each dot represents a cell.

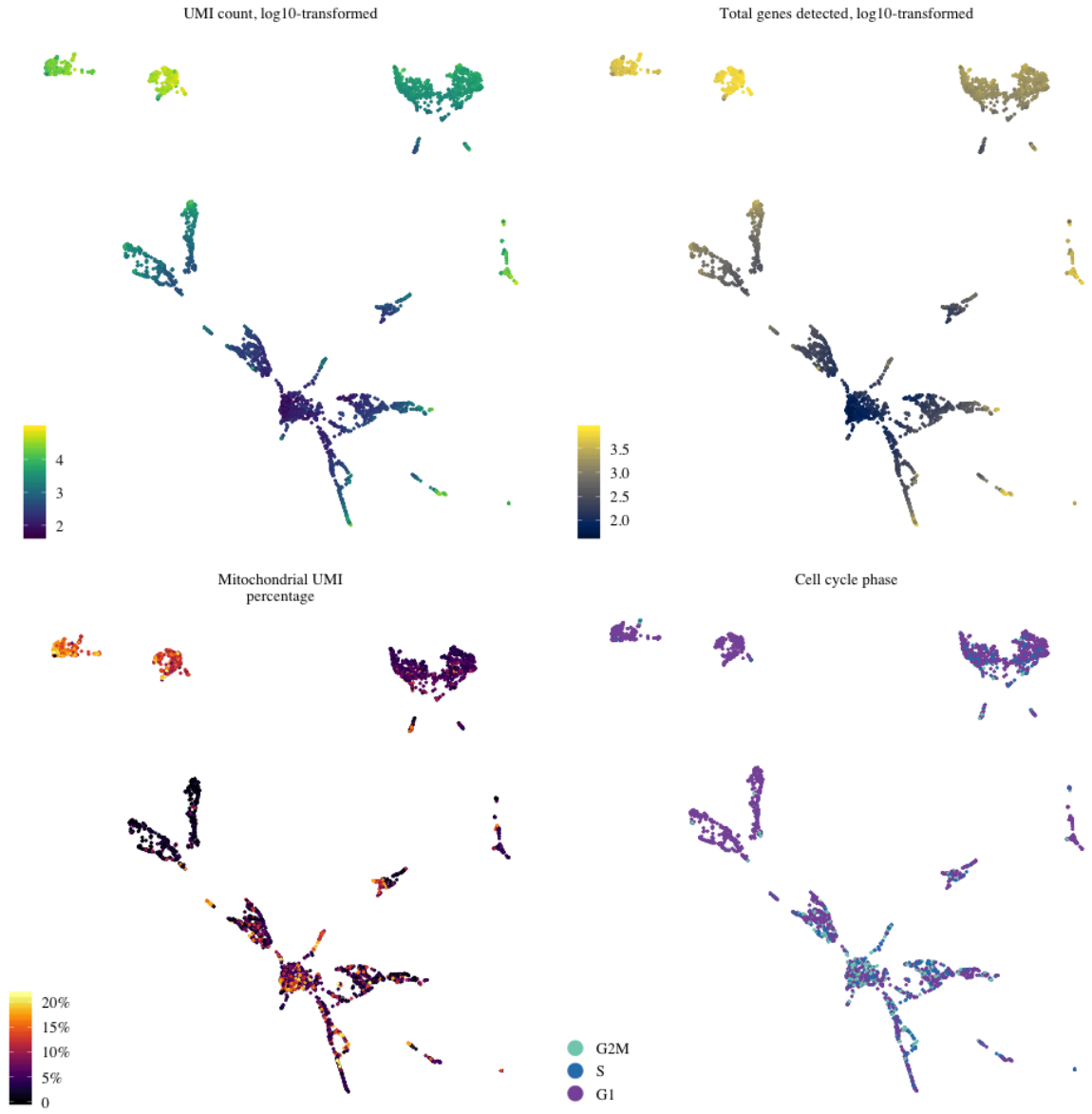


Fig. 2.14 UMAP of the *patient2* sample colored for UMI count, total detected gene count, mitochondrial UMI percentage, and cell cycle phase. Each dot represents a cell.

Fig. 2.15 Cell identification of the *patient1* sample. UMAP embedded data is colored by known biomarker expressions from low expression (gray) to high expression (red).

Fig. 2.16 Cell identification of the *patient2* sample. UMAP embedded data is colored by known biomarker expressions from low expression (gray) to high expression (red).

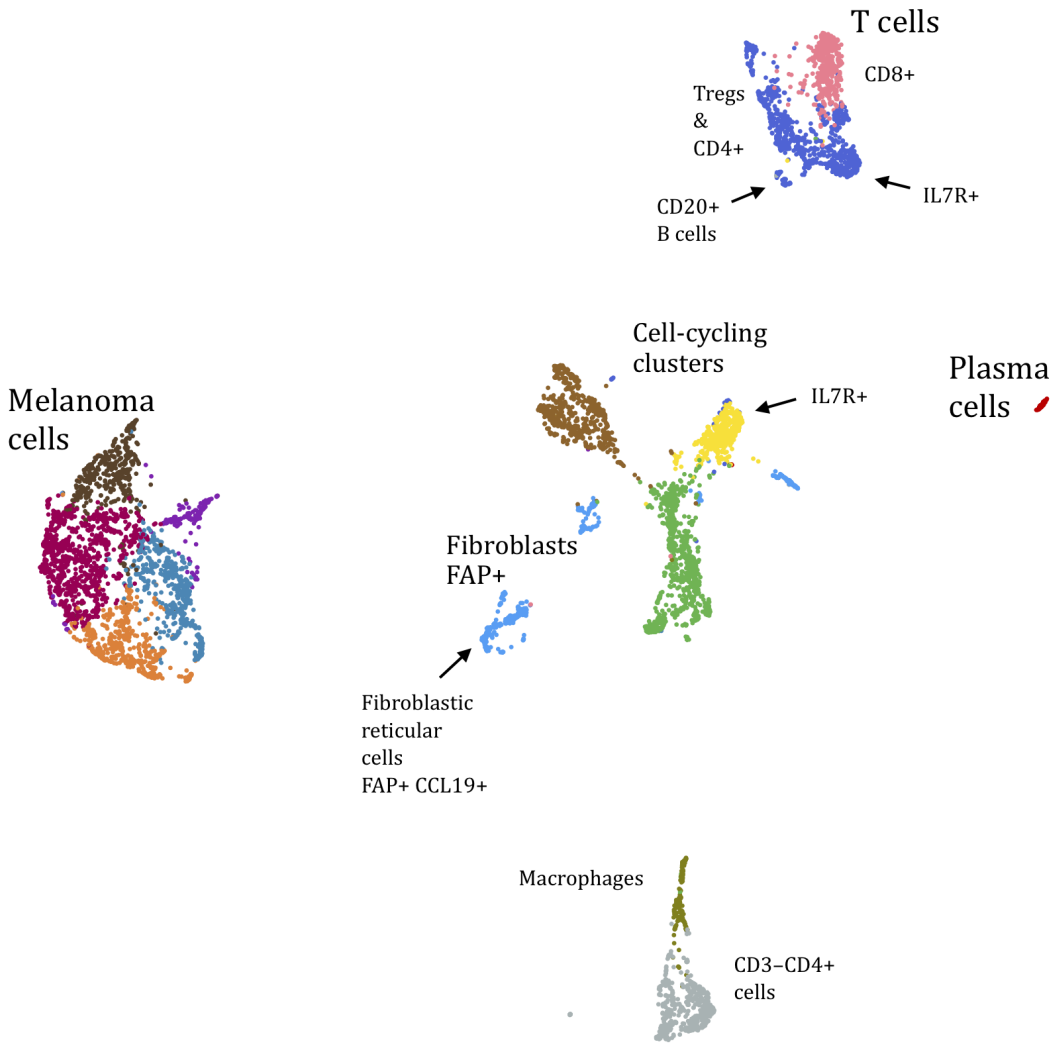


Fig. 2.17 Cell identification of the *patient1* sample. UMAP embedded data is colored by cluster membership. Each cluster is annotated for cell identity.

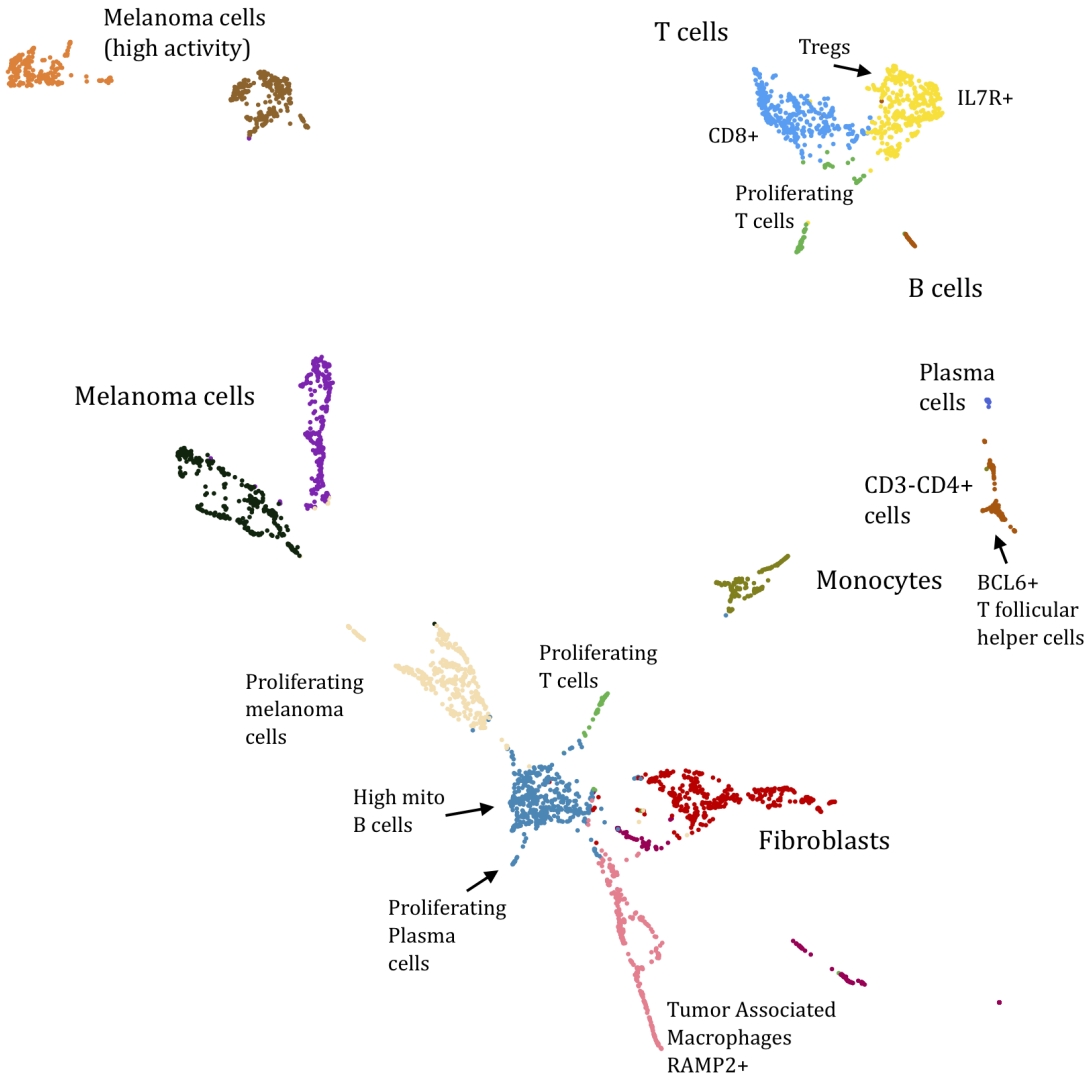


Fig. 2.18 Cell identification of the *patient2* sample. UMAP embedded data is colored by cluster membership. Each cluster is annotated for cell identity.

2.6 V(D)J Reconstruction

In the *patient1* sample, in 1,573 cells a full-length, productive BCR, and in 1,999 cells a full-length, productive TCR was detected, while in the *patient2* sample, 1,486 and 819 cells were found bearing a full-length, productive BCR and TCR, respectively.

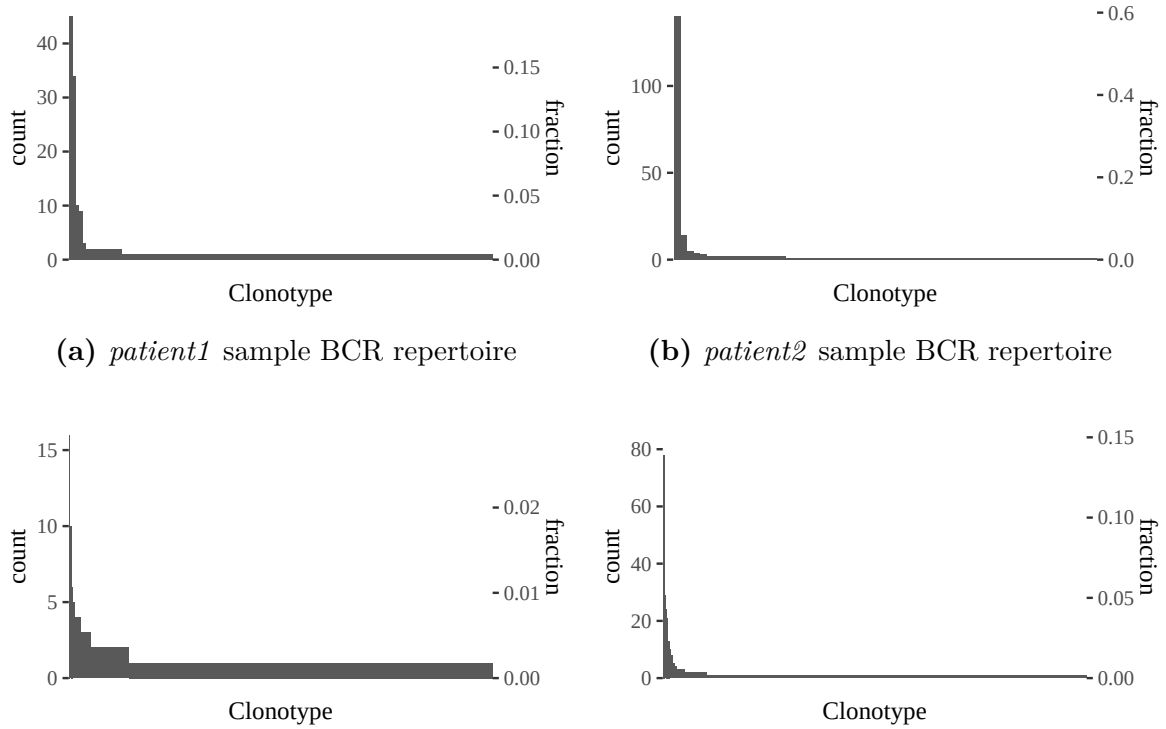
2.6.1 BCR repertoire

The high number of Ig gene expressing cells in the samples can be due to background noise introduced by plasma B cells and B cells with high levels of RNA, which might have leaked during sample preparation, and thus we are picking up transcripts present in the fluid rather than in intact cells. Further inspection of these cells for BCR expression revealed that some of them only express single chains, and these chains are, in fact, shared with a high-frequency clonotype with paired chains. Additionally, some of these chains have only one UMI, which also strengthens the possibility of the detected BCR coming from ambient RNA. To detect the likely B cells, I only kept cells with at least one heavy and one light chain having a minimum of two UMIs. I also removed cells that had an even number of multiple heavy and light chains that were shared with other, more frequent clones as these were most likely doublets. This resulted in 236 BCR expressing cells in the *patient1* sample, and 237 in the *patient2* sample. In the *patient1* sample, I detected 129 unique clonotypes, and 64 in the *patient2* sample. Recall that clonotypes are distinguished by the set of CDR3 amino acid sequences. In Figure 2.19b, observe that the *patient2* sample has a clonotype which makes up 59% of the BCR repertoire while the most frequent ten clonotypes in the *patient1* sample make up 46.55% of the sample's BCR repertoire.

2.6.2 TCR repertoire

Similar to the filtering with BCR expressing cells, I only kept the cells with at least one TRA and one TRB chain with a minimum of two UMIs per chain and removed possible doublets which had an even number of multiple TRA and TRB chains that were shared with other clones. In the *patient1* sample, this resulted in 810 TCR expressing cells and 631 unique TCR clonotypes, and with 561 TCR expressing cells and 338 unique TCR clonotypes in the *patient2* sample.

`cellranger vdj` pipeline detected more cells, especially T cells, in each sample and AgR repertoire, however, the `cellranger vdj` pipeline does not check for possible low-frequency chains present in clonotypes that share their other chains with a higher



(a) *patient1* sample BCR repertoire

(b) *patient2* sample BCR repertoire

(c) *patient1* sample TCR repertoire

(d) *patient2* sample TCR repertoire

Fig. 2.19 BCR and TCR clonotype repertoires of each patient are shown either as the number of occurrences or the fraction of clonotypes in the repertoires. The y-axes are free-scaled between samples.

frequency clonotype. The pipeline does not consider possible residual transcripts; the transcripts which are usually detected with a single UMI and create false clonotypes of multiple chains. Furthermore, the pipeline does not analyze single-chain clonotypes. They are kept as is without checking if they are, in reality, coming from a higher frequency clonotype. Figures B.3 and B.4 show the clonotypes called by the `cellranger vdj` pipeline. In Figure B.3, the 6$^{\text{th}}$ clonotype has a chain (IGKV3-11:IGKJ3:IGKC) which has five contigs made of six UMIs while the other chains have a median UMI to contig ratio >15. This chain is a residual transcript and is creating a false clonotype when, in fact, these cells belong to the 1$^{\text{st}}$ clonotype, which also happens to be the most clonally expanded. This is also the case for the 8$^{\text{th}}$ clonotype. Though not visible in the screenshots, this is, in fact, the case for 14 clonotypes; they all merge with the most dominant clonotype resulting in an additional 26 cells for this clonotype. Also, in Figure B.4, observe the abundance of single-chain TCR clonotypes. This is also present in Figure B.3.

2.7 Results

2.7.1 Clonal plasma cell expansion

Integration of GEX and V(D)J data

Recall that once the cells are barcoded with the 10x beads inside individual GEMs, they are broken, and all the transcripts mix. Some of these transcripts come from cells bearing a BCR or a TCR. `cellranger count` and `cellranger vdj` rely on different algorithms and assumptions when detecting cells, and therefore, a discrepancy can arise among the detected cells. I aimed to overcome this by merging cells as previously described. However, some BCR and TCR expressing cells may have been missed due to (i) low expression levels for one of the chains resulting in a single chain receptor, or (ii) low expression levels, in general, resulting in undetected chains by `cellranger vdj`, or (iii) a cell being discarded by QC. Therefore, the cells here which I integrate in terms of the GEX and V(D)J data are definitely B or T cells, but there may be more within the cell population, but a chain-pair (or more) with sufficient UMI count was not detected in those cells.

GEX profile of BCR bearing cells

I overlaid the clonotype information on the UMAP embedded scRNA-seq data in order to visualize the clustering and abundance of the detected clonotypes. Note that, from each AgR repertoire of each sample, I took the most clonal five clonotypes and assigned them a unique id while keeping all the other clonotypes labeled as "other" in order to prevent unnecessary noise. Figures 2.20 and 2.21 show the projections of each sample with the t-SNE reduction. t-SNE allows for a spread out visualization (since the global distances are not preserved) which is more helpful in this case. The UMAP visualizations are shown in Figures B.5 and B.6.

Sample	ID	CDR3.aa	Frequency	Fraction
Patient1	P1.BCR1	IGH:CASGAGEDAAMVTILFDYW;IGL:CSYAGGYTWVF	45	0.190
	P1.BCR2	IGH:CARGMDYDSSGYYFRFDYW;IGK:CQQRSNWPPWTF; IGL:CAAWDDSLNGGVF	34	0.140
	P1.BCR3	IGH:CASGAGEDAAMVTILFDYW;IGL:CQVWDSSSDHWVF; IGL:CSYAGGYTWVF	10	0.042
	P1.BCR4	IGH:CARGMDYDSSGYYFRFDYW;IGL:CAAWDDSLNGGVF	9	0.038
	P1.BCR5	IGH:CARDGGDKQYCSGGNCYFFDSW;IGK:CQQSYSIPYTF	3	0.013
Patient2	P2.BCR1	IGH:CARRVGATSAFDIW;IGL:CQSGDSRLTFVVF	140	0.590
	P2.BCR2	IGH:CARDIGVRGLFLKTYHYGLDVW;IGK:CQHCDDSNQAF	14	0.059
	P2.BCR3	IGH:CAKDIIKGRGDYAMDVW;IGL:CGTWDSSLSAVVF	5	0.021
	P2.BCR4	IGH:CATEISCSGDDCRDHW;IGK:CQQYDSWPRNTF	4	0.017
	P2.BCR5	IGH:CAVDFWNGFLRGFFDSW;IGK:CQQYFTTPRTF	3	0.013

Table 2.1 Details of the BCR clonotypes that are reflected onto the 2D projections.

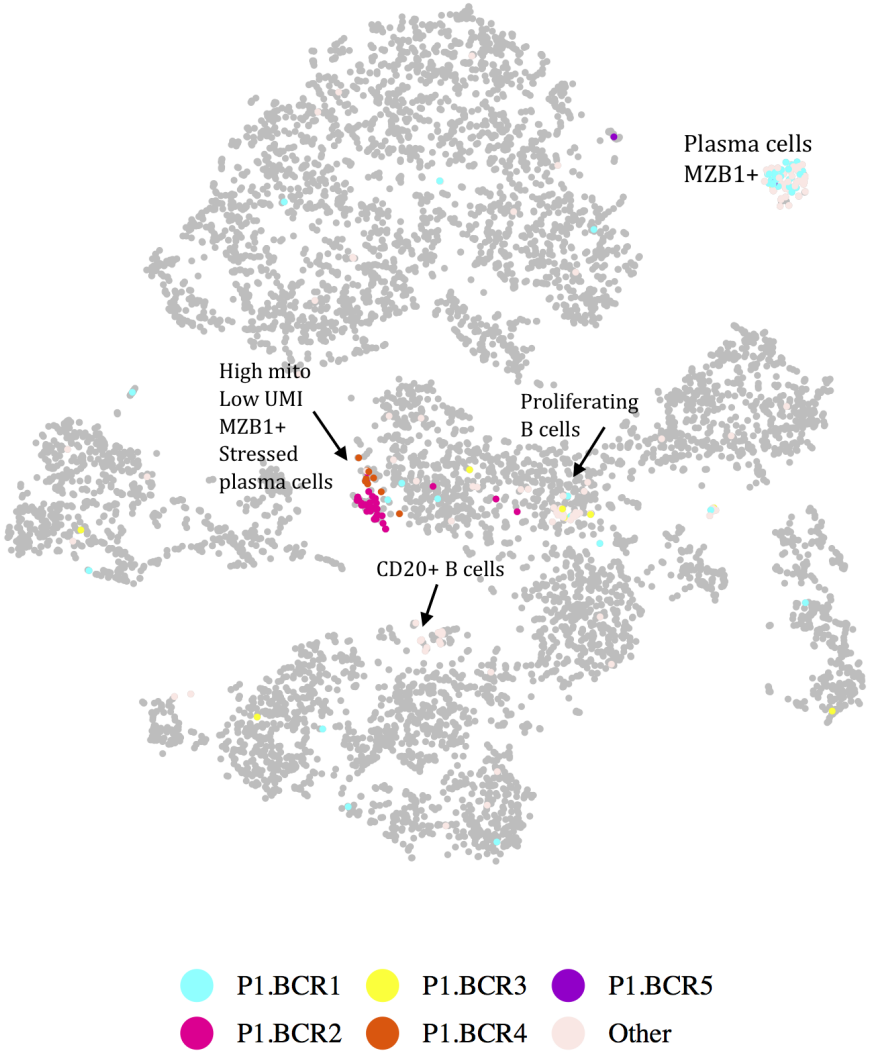
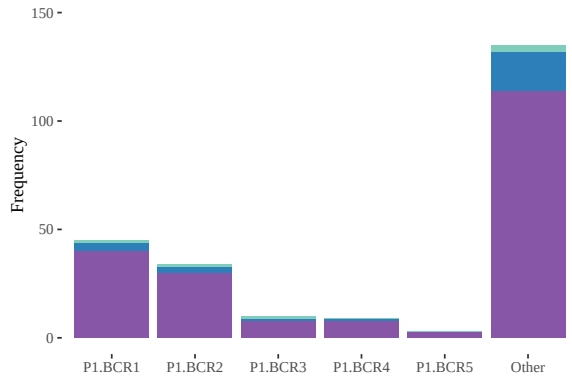
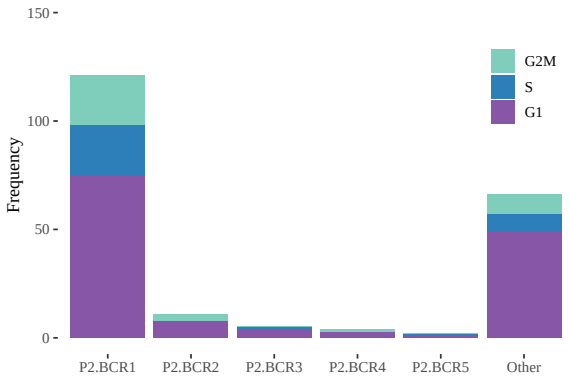


Fig. 2.20 Clonotype projection of the *patient1* sample. t-SNE projection is colored by clonotype membership. A BCR was not detected in cells colored with dark gray (7,127 cells). P1.BCR1:45, P1.BCR2:34, P1.BCR3:10, P1.BCR4:9, P1.BCR5:3 cells. Other:135 cells.

Fig. 2.21 Clonotype projection of the *patient2* sample. t-SNE projection is colored by clonotype membership. A BCR was not detected in cells colored with dark gray (3,875 cells). P2.BCR1:140, P2.BCR2:14, P2.BCR3:5, P2.BCR4:4, P2.BCR5:3 cells. Other:71 cells.

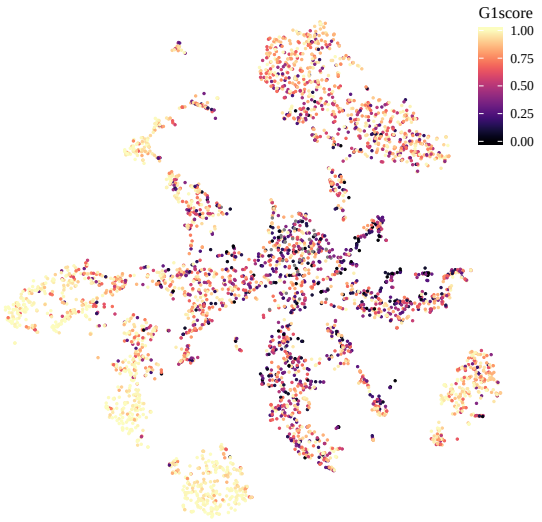


(a) *patient1* sample



(b) *patient2* sample

(c) *patient1* sample



(d) *patient2* sample

Fig. 2.22 Cell cycle stage distribution within the most dominant five clonotypes of each sample are given in (a-b). Cell counts may be less than the actual count as not all cells were assigned a phase. (c-d) overlays the G1 score on the t-SNE projections of each sample to further show the replicating cells. Gray colored points indicate the cells which were not assigned a cell cycle phase (52 cells in the *patient2* sample, none in *patient1*).

P2.BCR1 clonotype in the *patient2* sample displays a significant intratumoral clonal expansion. Figure 2.22b further shows that it is also proliferating as cells bearing this clonotype are in S and G2M cell cycling stages. Also, recall that the phase assignments are based on thresholds on the G1 and G2M scores, so it is possible that there are, in fact, more cells that are likely proliferating.

Class-switching in BCRs

In both samples, IgG isotypes were the most common. The *patient2* sample's BCR repertoire had very few IgM isotypes, and no IgD isotypes were detected. In the *patient2* sample, the most dominant clonotype expressed has the IGHG2 constant region gene.

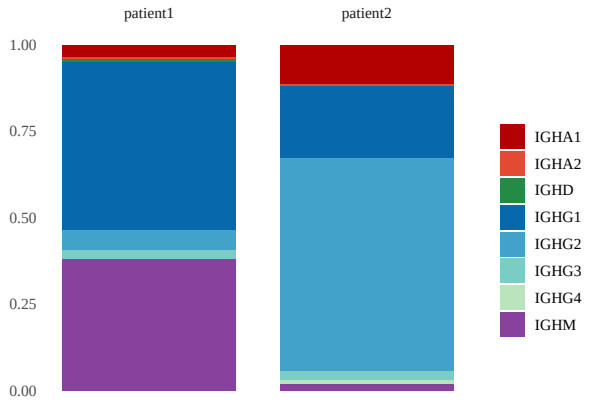


Fig. 2.23 Distribution of Ig isotypes in each sample's BCR repertoire.

SHM

Other than GEX profile and class-switching, SHM is also an indicator of antigen challenge. `cellranger vdj` detects mutations within the assembled contigs. With all the contigs that belong to the same clonotype chain, a consensus contig is generated. Figure 2.24 shows the consensus alignment to the reference IMGT sequence for the chosen P2.BCR1 clonotype's heavy chain. Here orange denotes the mutations, which are the possible SHMs. However, as I only have scRNA-seq data, I cannot say with certainty that these are SHMs; they could be point mutations in the patient's genome.

Fig. 2.24 The consensus contig alignment of the P2.BCR1 clonotype's heavy chain to the IMGT reference. Image is a cropped screenshot of the Loupe VDJ Browser, Version 3.0.0, with the `cellranger vdj` output of the *patient2* sample loaded. Orange lines denote mismatches, purple deletions, green start codon, dark gray is the CDR3 region, and light gray denotes the contig coverage.

However, the GEX profile and the class-switching is enough evidence to show an intratumoral expansion of the plasma cells within both samples, with the *patient2* sample demonstrating a highly clonal expanded one.

2.7.2 Sequences of clonal intratumoral antibodies

In the *patient2* sample, the `P2.BCR1` clonotype demonstrates high clonality. Furthermore, this clonotype has the GEX profile of mature and proliferating plasma cells. Class-switching and SHM is further evidence of antigen challenge. Sequences in Listing 2.1 are the nucleotide sequences of the heavy and light chains of this antibody, and Listing 2.2 is their IgBlast alignment to the IMGT reference which includes possible SHMs. This antibody is a candidate for finding a melanoma tumor-antigen. With these sequences, we can express the antibody, and further detect which antigen it binds. I discuss this in Section 2.8.

```
> P2.BCR1.H{IGHV5-51:IGHD1-26:IGHJ3:IGHG2}
TCTTGCATCATACTTCTTTTCTTATATGGGGGAGTCTCCCTCACTGCCCAGCTGGGATCTCAGGGCTTCA
TTTTCTGTCCTCCGCCATCATGGGGTCAACCGCCATCCTCGCCCTCCTCCTGGCTGTTCTCCAGGGAGTC
TGTGGCGAGGTGCAGCTGGAGCAGTCTGGAGGAGAGGTGAAGAAGCCGGGGGAGTCTCTGAAGATCTCCT
GTAAGGCTTCTGGATACAGATTTACCAGCTTTTGGATCGTCTGGGTGCGCCAGATGCCCGGAAAAGGCCT
GGAGTGGGTGGGGATCATCCATCCTGGTGACTCCGATATTAGTTACAGCCCGTCTTTTGAAGGCCACGTC
ACCCTGTCAGCCGACAGGTCCAGCACCACCGCCTACCTGCAGTGGGACAGCCTGAAGGCCTCGGACAGCG
CCATGTATTACTGTGCGAGAAGGGTGGGAGCTACCTCTGCTTTTGATATCTGGGGCCTAGGGACACTGGT
CACCGTCTCTTCAGCCTCCACCAAGGGCCCATCGGTCTTCCCCCTGGCGCCCTGCTCCAGGAGCACCTCC
GAGAGCACAGCGGCCCTGGGCTGCCTGGTCAAGGACTACTTCCCCGAACCGGTGACGGTGTCGTGGAACT
CAGGCGCTCTGACCAGCGGCGTGCACACCTTCCCAGCTGTCCTACAGTCCTCAGGACTCTACTCCCTCAG
CAGCGTGGTGACCGTGCCCTCCAGCAACTTCGGCACCCAGACC
> P2.BCR1.H{IGHV5-51:IGHD1-26:IGHJ3:IGHG2_CDR3}
TGTGCGAGAAGGGTGGGAGCTACCTCTGCTTTTGATATCTGG
> P2.BCR1.L{IGLV3-25:IGLJ2:IGLC3}
GATCCTGCTTCTTCTTCTTCCTCTTGTCGCCTGTGCTGCCCCCACAGCTGGTTTGGGTGACATCTCTCCA
GGAGGAGTCCCAGAGGAAGTAAATTTGCATAAACACCAAACACTGACTACCCTAAAAAGCCTGAGAGAGA
ATAAGAGAGGCCTGGGGAGCCTAGCTGTGCTGTGGGTCCAGGAGGCAGAACTCTGGGTGTCTCACCATGG
CCTGGATCCCTCTACTTCTCCCCCTCCTCACTCTCTGCACAGACTCTGAGGCCGCCCATGAGTTGACACA
GCCACCCTCGGTGTCAGTGTCCCCAGGACAGACGGCCAGAATCACCTGCTCTGGAGATGGATTGTCAAAG
CAGTATGTTCATTGGTACCAGGCGAAGCCAGGCCAGGCCCCTGTCTTGGTGATATATAAAGACACTGAGA
GGCCCCCAGGAATCCCTGAGCGATTCTCTGCCTCCAGCTCAGCGACGACAGTCACATTGACCATTAGTGG
AGTCCAGGCAGAGGACGAGGCTGACTATTATTGTCAATCGGGAGACAGCCGTCTTACTTTTGTGGTTTTT
GGCGGCGGGACCAAGCTGACCGTCCTACGTCAGCCCAAGGCTGCCCCCTCGGTCACTCTGTTCCCGCCCT
CCTCTGAGGAGCTTCAAGCCAACAAGGCCACACTGGTGTGTCTCATAAGTGACTTCTACCCGGGAGCCGT
GACAGTGGCCTGGAAGGCAGATAGCAGCCCCGTCAAGGCGGGAGTGGAGACCTCCACACCCTCCAAACAA
AGCAACAACAAGTACGCGGCCAGCAGCTATCTGAGCCTGACGCCTGAGCAGTGGAAGTCCCACA
> P2.BCR1.L{IGLV3-25:IGLJ2:IGLC3_CDR3}
TGTCAATCGGGAGACAGCCGTCTTACTTTTGTGGTTTTT
```

Listing 2.1 Sequences of tumor-antigen specific antibody candidate

```
                                      <--------------------------------FR1-IMGT---------------------------------><--------CDR1-IM
                                       E  V  Q  L  E  Q  S  G  G  E  V  K  K  P  G  E  S  L  K  I  S  C  K  A  S  G  Y  R  F  T
                   P2.BCR1.H     147  GAGGTGCAGCTGGAGCAGTCTGGAGGAGAGGTGAAGAAGCCGGGGGAGTCTCTGAAGATCTCCTGTAAGGCTTCTGGATACAGATTTACC  236
V  90.8% (267/294)  IGHV5-51*03  1    ............T...........C.........A........................................G............C......  90

                                      GT------><---------------------FR2-IMGT----------------------><--------CDR2-IMGT------><-----
                                       S  F  W  I  V  W  V  R  Q  M  P  G  K  G  L  E  W  V  G  I  I  H  P  G  D  S  D  I  S  Y
                   P2.BCR1.H     237  AGCTTTTGGATCGTCTGGGTGCGCCAGATGCCCGGAAAAGGCCTGGAGTGGGTGGGGATCATCCATCCTGGTGACTCCGATATTAGTTAC  326
V  90.8% (267/294)  IGHV5-51*03  91   ....AC.......G.....................G...............A...........T.............T....CC..A...  180

                                      ------------------------------------------------FR3-IMGT
      -------------------------------------
                                       S  P  S  F  E  G  H  V  T  L  S  A  D  R  S  S  T  T  A  Y  L  Q  W  D  S  L  K  A  S  D
                   P2.BCR1.H     327  AGCCCGTCTTTTGAAGGCCACGTCACCCTGTCAGCCGACAGGTCCAGCACCACCGCCTACCTGCAGTGGGACAGCCTGAAGGCCTCGGAC  416
V  90.8% (267/294)  IGHV5-51*03  181  ........C..CC.......G......A.C..........A.....T..G...................AG...................  270

                                      ------------------><---------------CDR3-IMGT------------->
                                       S  A  M  Y  Y  C  A  R  R  V  G  A  T  S  A  F  D  I  W  G  L  G  T  L  V  T  V  S  S
                   P2.BCR1.H     417  AGCGCCATGTATTACTGTGCGAGAAGGGTGGGAGCTACCTCTGCTTTTGATATCTGGGGCCTAGGGACACTGGTCACCGTCTCTTCAG  504
V  90.8% (267/294)  IGHV5-51*03  271  .C.....................----------------------------------------------------------------  294
D  100.0% (11/11)   IGHD1-26*01  7    ------------------------------...........------------------------------------------------  17
J  95.7% (45/47)    IGHJ3*02     4    ------------------------------------------...................A.......A.................  50

                                      <---------------------------------FR1-IMGT--------------------------------><----CDR1-IMGT--
                                         H  E  L  T  Q  P  P  S  V  S  V  S  P  G  Q  T  A  R  I  T  C  S  G  D  G  L  S  K  Q
                   P2.BCR1.L     265  CCCATGAGTTGACACAGCCACCCTCGGTGTCAGTGTCCCCAGGACAGACGGCCAGAATCACCTGCTCTGGAGATGGATTGTCAAAGCAGT  354
V  90.2% (257/285)  IGLV3-25*02  2    ..T.....C.........................................................G..................C....C.......A.  91

                                      -><---------------------FR2-IMGT----------------------><CDR2-IM><------------------------------
                                      Y  V  H  W  Y  Q  A  K  P  G  Q  A  P  V  L  V  I  Y  K  D  T  E  R  P  P  G  I  P  E  R
                   P2.BCR1.L     355  ATGTTCATTGGTACCAGGCGAAGCCAGGCCAGGCCCCTGTCTTGGTGATATATAAAGACACTGAGAGGCCCCCAGGAATCCCTGAGCGAT  444
V  90.2% (257/285)  IGLV3-25*02  92   ...C.T...........CA.....................GC..................G..........T....G.............  181

                                      ------------------------FR3-IMGT
      -------------------------------------------------------><----------
                                      F  S  A  S  S  S  A  T  T  V  T  L  T  I  S  G  V  Q  A  E  D  E  A  D  Y  Y  C  Q  S  G
                   P2.BCR1.L     445  TCTCTGCCTCCAGCTCAGCGACGACAGTCACATTGACCATTAGTGGAGTCCAGGCAGAGGACGAGGCTGACTATTATTGTCAATCGGGAG  534
V  90.2% (257/285)  IGLV3-25*02  182  ......G...........G...A........G........C.................A..T..............C........A.C..  271

                                      --CDR3-IMGT------------>
                                      D  S  R  L  T  F  V  V  F  G  G  G  T  K  L  T  V  L
                   P2.BCR1.L     535  ACAGCCGTCTTACTTTTGTGGTTTTTGGCGGCGGGACCAAGCTGACCGTCCTA  587
V  90.2% (257/285)  IGLV3-25*02  272  .....A..GG.....--------------------------------------  286
J  91.9% (34/37)    IGLJ2*01     1    ----------------......A..C.....A....................  37
```

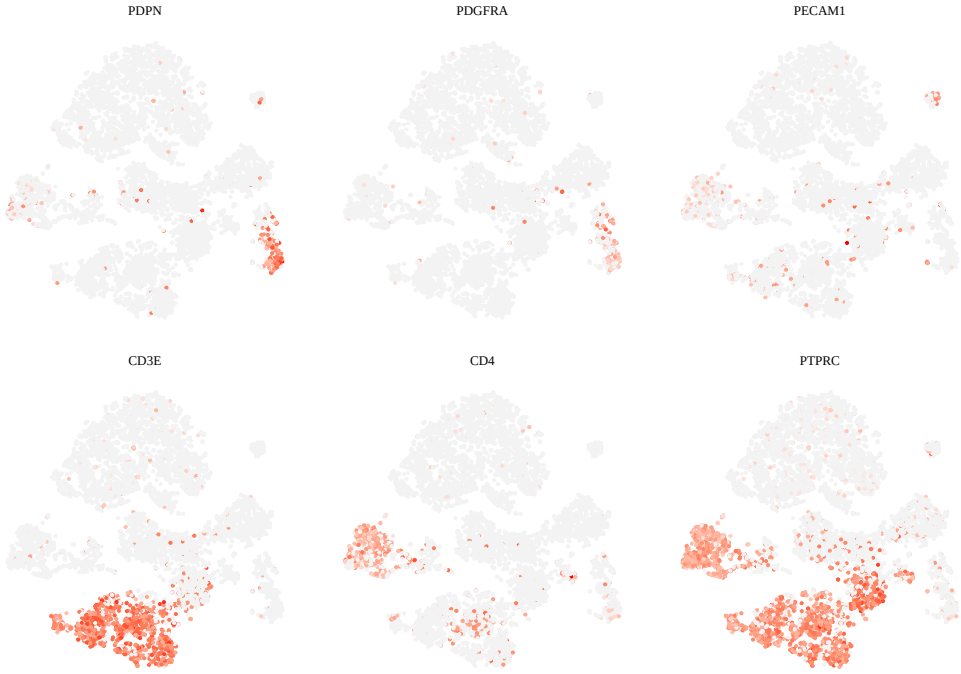
Listing 2.2 IgBlast alignments of P2.BCR1 heavy and light chains

2.7.3 Tertiary lymphoid structure detection

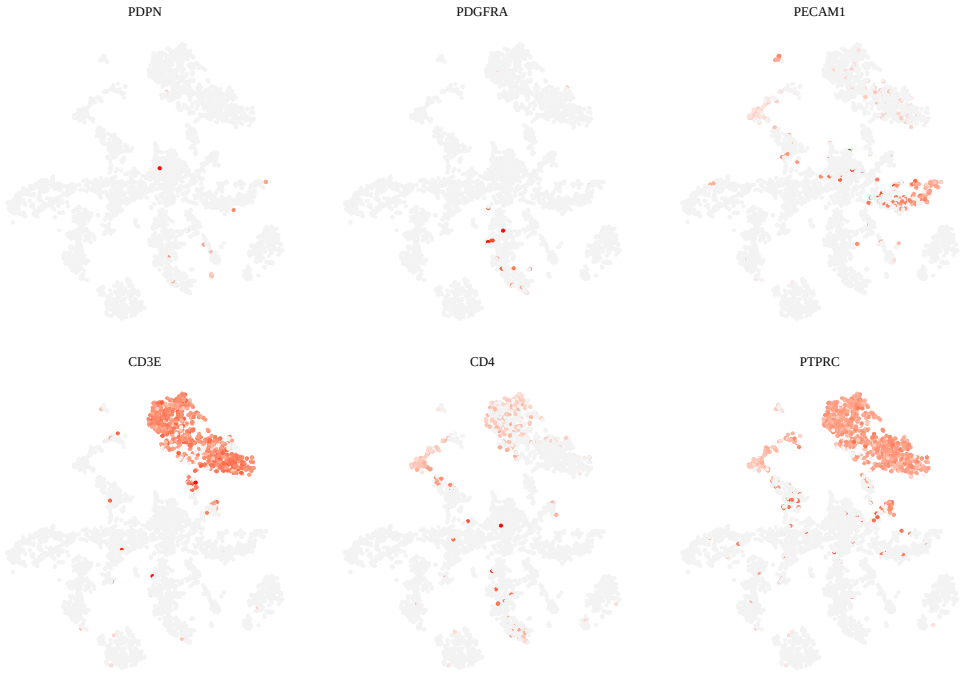
Lymph nodes are where immunological components meet: T cells, B cells, DCs, plasma cells, and macrophages gather inside a meshwork of connective tissue created by SLO-resident fibroblasts, also known as fibroblastic reticular cells (FRCs) [68]. Lymph nodes facilitate the initiation of an adaptive immune response by funneling the antigens and antigen-presenting cells towards B and T cells. Tertiary lymphoid structures (TLSs) are ectopic lymphoid-like structures, and they have been observed in various solid tumors, including melanomas [48].

Tumor-associated TLSs are composed of a T cell rich outer zone that contains mature DCs, surrounding a ring of naive B cells around a GC that contains mostly B cells, with a smaller number of T cells, follicular DCs (FDCs), macrophages, and FRC-like reticular cells [80], [36]. FRCs can be defined by the expression of the surface molecules podoplanin (PDPN) and platelet-derived growth factor receptor-α (PDGFRα) and the lack of platelet endothelial cell adhesion molecule (PECAM1) and CD45 (PTPRC) [36]. The FRCs and FDCs may be the source of the chemokines CCL19, CCL21, CXCL13, and CXCL12, which are known to regulate the SLO microenvironment and have also been observed in the intratumoral TLS microenvironment [63].

Various studies have proposed TLS gene signatures, and they are either chemokine or cell population based [201]. One geneset comprising 12 chemokines, including the aforementioned CCL19, CCL21, CXCL13, and CXCL12, have been used as a proxy for TLS presence in melanoma [158]. Also, a plasma cell specific signature comprising the expression of TNFRSF17 has been proposed to characterize the TLS in ovarian cancer [158]. For both samples I have previously demonstrated the presence of B lineage cells ($CD19^+$, $CD20^+$, $MZB1^+$/$TNFRSF17^+$ plasma cells), and T cells (including $CD8^+$, $CD4^+$, $CD4^+FOXP3^+$). In Figure 2.25a, observe the presence of FRCs which carry the gene expression profile $PDPN^+PDGFRA^+PECAM1^-PTPRC^-$. The *patient2* sample lacks such cells; recall however that this sample was obtained with a needle biopsy. In addition to the FRCs, observe the cells which are $CD45^+CD4^+CD3^-$. These cells show a hematopoietic lymphoid tissue inducer (LTi) cell profile [16] and are potentially important for lymph node formation. Besides these cell populations, I have also inspected both samples for chemokine expression. Figures 2.26 and 2.27 show the expression levels of the 12-chemokine signature geneset taken from [158], comprising CCL2, CCL3, CCL4, CCL5, CCL8, CCL18, CCL19, CCL21, CXCL9, CXCL10, CXCL11, and CXCL13. Figures 2.26a and 2.27a evaluate the expression of the the 12-chemokine signature by calculating the mean expression of the geneset on the normalized counts.



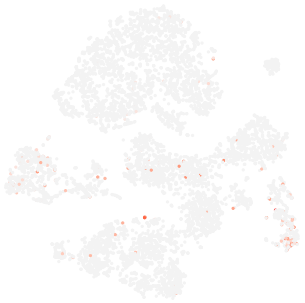
(a) *patient1* sample



(b) *patient2* sample

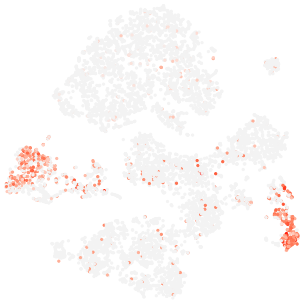
Fig. 2.25 t-SNE projections of biomarker expressions showing the presence of FRC-like reticular cells in the *patient1* sample, and LTi cells in both *patient1* and *patient2* samples. (Gray: low, red: high expression)

(a) 12-chemokine signature

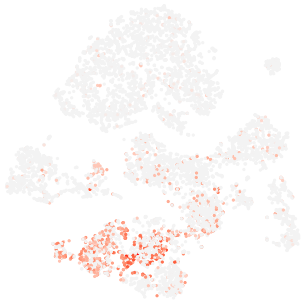


(b) CCL19

(c) CCL21

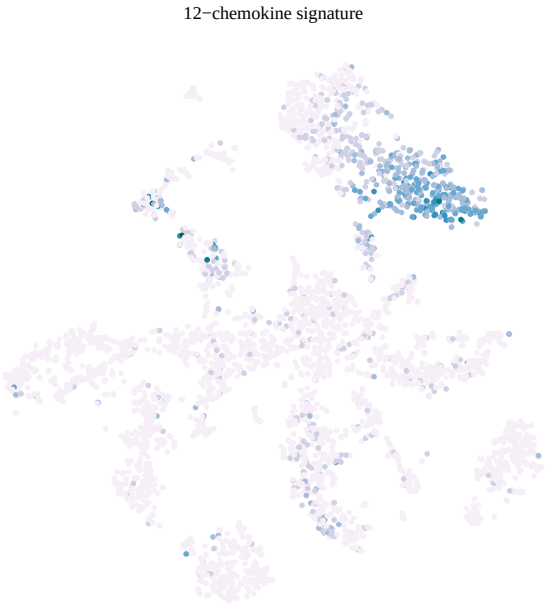


(d) CXCL12

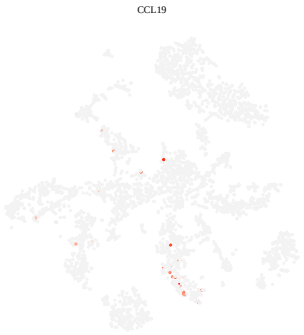


(e) CXCL13

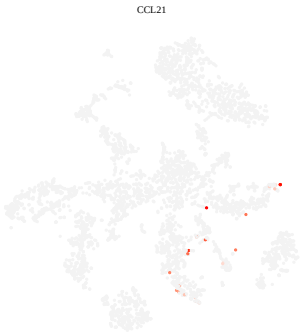
Fig. 2.26 t-SNE projections of TLS associated chemokine expressions in the *patient1* sample. (In (a), gray: low, blue: high expression. In (b-d) gray: low, red: high expression)



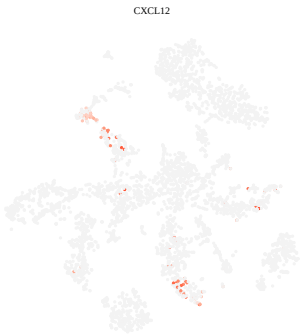
(a) 12-chemokine signature



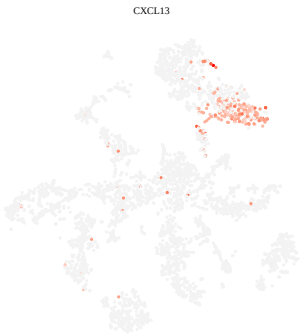
(b) CCL19



(c) CCL21



(d) CXCL12



(e) CXCL13

Fig. 2.27 t-SNE projections of TLS associated chemokine expressions in the *patient2* sample. (In (a), gray: low, blue: high expression. In (b-d) gray: low, red: high expression)

2.8 Discussion

The role of the effector T cells in anti-tumor immunity has been studied in depth, revealing that effective ICB therapies require the revival and clonal-proliferation of the pre-existing mass of antigen-experienced TILs in the TME [242], [214], [6], [26]. The role of B cells, however, is less clear and has only recently gained attention [241], [254]. In this study, I characterized TIL-Bs in melanomas, revealing $CD19^+$, $CD20^+$, and plasma B cells, as well as proliferating B cells at various differentiation stages. Furthermore, by coupling their gene expression profile with their BCR clonotype, I was able to detect clonally expanded plasma B cells.

The abundance of TIL-Bs and TILs show the cooperation of the two adaptive immune system lymphocytes. Also, in both samples, I detected $CD3^-CD4^+CD45^+$ cells which show an LTi cell profile [16]. Some of these cells also expressed BCL6, which could indicate the presence of $BCL6^+$ T follicular helper (Tfh) cells [168]. Tfh cells help B cells during GC reactions by mediating class-switching and SHMs, and facilitating affinity maturation. Additionally, I showed the presence of FRCs which carry the gene expression profile $PDPN^+PDGFRA^+PECAM1^-PTPRC^-$ and the 12-chemokine signature. The detection of such cells suggests the formation of an intratumoral TLS, and our findings are in agreement with previous studies [48]. The scRNA-seq results would need to be further confirmed by a visual analysis of the stained sections of these tumor biopsies.

While it is widely accepted that $CD8^+$ T cells drive anti-tumor immunity, studies of TLSs suggest that B cells may also play an essential role [63]. Additionally, an association between the presence of TLS and favorable prognosis has been reported for many solid tumors [201]. In our study, the presence of TLS along with proliferating and clonally expanded plasma B cells which express class-switched antibodies at baseline, may suggest that there is already an ongoing coordinated intratumoral immune response. The presence of a TLS is less obvious in the *patient2* sample, however, it is important to note that this sample was obtained with a needle biopsy, which usually recovers only a limited amount of tissue. It has been shown that TLSs are more abundant in the invasive margin or the stroma, and a wider sampling of the tumor is beneficial in detecting a TLS [201].

Both samples showed presence of plasma B cells where the *patient2* sample's Ig repertoire demonstrated higher clonality and clonal proliferation. By using single-cell sequencing coupled with targeted VDJ sequencing, we were able to detect the proliferation of this antibody-secreting plasma cell. The most dominant clonotype in this sample, P2.BCR1, expressed the heavy chain IGHV5-51:IGHD1-26:IGHJ3:IGHG2 and the light chain IGLV3-25:IGLJ2:IGLC3. The heavy chain shows both class-switching and possible SHMs. This intratumoral antibody is generating an immune response towards an

intratumoral antigen which is a candidate for a melanoma tumor-antigen. Using the sequences given in Listing 2.1, it could be possible to express this antibody in mammalian cells (i.e. HEK293 cells), and find the antigen it binds upon successful antigen-antibody recognition.

The TME is a highly heterogeneous microstructure [111]. Ideally, the clustering of cells should be driven mainly by the biological component of the gene expression profile. However, it can be reasoned whether other factors such as the differences in library sizes of different cell types or mitochondrial gene expression levels should be considered biological or technical in nature. For this analysis, I decided not not remove the effects of library size and mitochondrial UMI count proportions from the scRNA-seq data. The fact that clustering is exploratory data analysis is much more evident when immune components are in question. While it is possible to cluster based on high-level cell types (B, T, DC, NK, and so on), we can further zoom in to separate $CD8^+$ T cells from $CD4^+$ T cells, or mature plasma B cells from high-metabolic-activity proliferating plasma B cells. It should be noted that all of the cell QC and clustering outlined in this study was an incremental process which took many attempts. Another source of variation, the cell cycle heterogeneity in scRNA-seq data, may assign the same type of cells into multiple clusters. Therefore it is common practice to remove the cell cycle effect from the data. The effects of cell cycle can be most evident in homogeneous cell populations. However, in this study I was not interested in detecting a new cell type or trajectories, but was instead aiming to determine the clonality and proliferation of the components within the TME. Hence I only determined the cell cycle phase of each cell but did not remove this effect.

In this study, I characterized the TME of two melanoma metastases taken pre-immunotherapy. Using the same technique, we can study all the patients in the MelResist clinical trial for both pre- and post-immunotherapy in order to determine the effect of baseline TIL-B and TLS presence on patient response. Moreover, we can determine the fate of the clonally expanded antibodies post-immunotherapy and potentially find candidate antibodies for tumor-antigen detection.

Chapter 3

Intratumoral antibody sequences reconstructed from RNA-seq

3.1 Introduction

The TME is infiltrated by various immune cells, and their presence has been shown to be associated with clinical outcomes in some cancer types [206], [82], [264], [71]. The interplay between the tumor and the immune cells is a complex process that determines the fate of tumor immunity. To date, the majority of research has focused on tumor-infiltrating T lymphocytes [260]. In this chapter, I focused on the tumor infiltration capabilities of B lymphocytes instead, specifically in two highly immunotherapy resistant cancers, colorectal cancer (CRC) and pancreatic ductal adenocarcinoma (PDAC). Furthermore, I studied the effect of a particular compound called AMD3100 [65] on the B cell responses in CRC and PDAC. Ultimately in this study, using paired data from a clinical trial of AMD3100 [50], I aimed to identify antibody sequences expanded after this particular treatment in PDAC and CRC.

In Section 3.2, I explain immune privilege in PDAC and CRC and the ineffectiveness of current immunotherapies towards these cancers. Next in Section 3.3, I present the CAM-PLEX clinical trial [50], which assesses the efficacy of a new immunotherapy drug. In Section 3.4, I use RNA-seq data of each biopsy pair to reconstruct patient immune repertoires computationally. Later in Section 3.5, I introduce my findings, which reveal the clonal expansion of single antibody clonotypes in some of the patients after therapy and present the nucleotide sequences of these antibodies. Finally, in Section 3.6, I suggest a possible future direction that could result in finding tumor-antigens of CRC and PDAC.

3.2 Background

3.2.1 Immune privilege in CRC and PDAC

CRCs are malignant tumors of the colon and rectum arising primarily from neoplastic colon polyps, which are small clusters of cells that form on the lining of the colon. CRCs are the third most common cancers worldwide, with 1.8 million reported cases in 2018 and the second most common cause of cancer-related mortality, resulting in 862,000 deaths [250]. CRC has a reasonable chance of survival compared to other cancer types; the five-year overall survival rate is 64%, with 90% of patients with localized cancer surviving at least five years [9].

The pancreas has many functions, two of which are especially important: it harbors a pool of exocrine glands aiding digestion, but has also an endocrine role. In particular, the endocrine pancreas is composed of small islands of cells, namely the islets of Langerhans, which release hormones, such as insulin, into the bloodstream. On the other hand, the exocrine glands secrete various enzymes through the pancreatic duct into the duodenum [40, Chapter 1, p. 3-20]. Uncontrollable growth of the exocrine cells may result in pancreatic exocrine tumors, which make up 93% of pancreatic cancers [41]. PDAC, a type of exocrine pancreatic cancer [100], which, even though is only the eleventh most common cancer in women and the twelfth most common in men worldwide, is the seventh most common cause of cancer-related deaths with a five-year survival rate of only 9% [187]. The most effective treatment is surgery; however, only a small fraction is resectable at the time of diagnosis, and even after resection, the five-year survival rate is not particularly encouraging [267].

Although popular ICIs such as PD-1 and CTLA-4 inhibitors have delivered clinical benefits in various solid tumors [130], [189], PDAC was not found to be responsive to ICIs. In a Phase I trial, a total of 207 patients with advanced cancers, including melanoma, CRC, and PDAC, were administered anti–PD-L1 antibody [31]. While a complete or partial response was observed in nine of 52 patients with melanoma, no objective response was observed in patients with CRC[1] or PDAC. Similarly, in a Phase II trial evaluating the efficacy of anti-CTLA-4, none of the enrolled 27 patients with advanced or metastatic PDAC showed a response, and progression was rapid with a short survival [195]. It is crucial to investigate which mechanisms protect these specific cancer types from immune attack.

[1]It is important to note that, a number of clinical trials looking at the efficacy of ICIs in microsatellite-instability-high (MSI-H) CRC [90], reported the objective response rate to be between 30% and 52% [170].

In PDAC, unlike other solid tumors that respond positively to immune checkpoint blockade (ICB) therapies, intratumoral effector T-cells are rare [45], and certain immunosuppressive cells such as cancer-associated fibroblasts (CAFs) within the TME might restrict the function of intra-tumoral effector T cells [65]. According to one hypothesis, PDAC does not respond to anti-PD-L1 because the T cells may not be getting enough contact with the cancer cells, as they are coated with high densities of chemokine (C-X-C motif) ligand 12 (CXCL12), which is known to repel T cells [65]. In one well known study, fibroblast activation protein (FAP) expressing CAFs have been shown to induce CXCL12 mediated immunosuppression on the basis of three observations: intra-tumoral T cells were not in the vicinity of cancer cells, cancer cells were coated with CXCL12, and the FAP^+ CAF was the primary source of CXCL12 in the tumor [65]. In order to test this hypothesis, [65] used two different approaches. First, they demonstrated that the depletion of FAP^+ CAFs results in the immune control of tumor growth, and recover the efficacy of anti-PD-L1 in PDAC bearing mice. Moreover, administering *AMD3100*, a molecule known to disrupt the interaction between CXCL12 and its receptor on immune cells CXCR4, increased the accumulation of effector T cells among cancer cells. Finally, the combination of AMD3100 with anti-PD-L1 significantly diminished the number of detectable cancer cells; however, anti-CTLA-4 did not augment the anti-tumor effect of AMD3100. These findings led to a Phase I clinical trial of AMD3100, enrolling patients with PDAC or microsatellite stable CRC (MSS-CRC) [90], which is discussed in Section 3.3.

The immune repertoire in pancreatic cancer has been studied in the past. For instance, [14] analyzed the TCR repertoire in pancreatic cancer. T cells were obtained from both the primary pancreatic tissues and matched blood samples of 16 cancer patients, and only from blood in the case of 8 healthy donors. Using deep sequencing of the TCRs, they found no significant differences in the TCR repertoires between cancer patients and controls. They did not identify any significantly expanded T cell clones. However, TCR clonotypes with low frequencies were relatively more abundant in tumors when compared to the matching blood samples.

In a different study, [99] investigated whether TCR repertoires could provide insights into the mechanisms of immunotherapy and predict outcome in pancreatic cancers. TCR repertoires were recovered using targeted deep sequencing. They examined the TCR repertoires recovered from the peripheral blood of 25 metastatic pancreatic cancer patients treated with ipilimumab (targeting CTLA-4) with or without GVAX (a pancreatic cancer vaccine), and from the peripheral blood and tumor biopsies from 32 patients treated with GVAX and mesothelin-expressing *Listeria monocytogenes* with or without nivolumab

(targeting PD-1). Their results demonstrate a change of diversity in the peripheral TCR repertoires of patients who received treatment, especially those receiving ipilimumab in combination with GVAX. They determined the combination of low baseline clonality and a high number of expanded clones post-therapy associated with significantly longer survival in patients who received ipilimumab but not in patients receiving nivolumab. Importantly, their study shows that TCR repertoire profiling can serve as a biomarker of clinical response in pancreatic cancer patients receiving immunotherapy. It should be noted that, neither study investigated the B cells and the BCR repertoire, and by solely focusing on TCRs, they captured only one aspect of the adaptive immune response. Analyzing the BCRs would give a more complete picture of the immunological status of cancer patients and a better understanding of the anti-tumor immune response in PDAC and CRC.

3.3 CAM-PLEX clinical trial

As already mentioned, [65] demonstrated that the blockade of the CXCL12/CXCR4 axis in a mouse model of PDAC resulted in the effector T cells and cancer cells to intermingle, and reduced tumor growth when combined with anti–PD-L1 immunotherapy. Consequently, a phase I clinical trial investigating plerixafor (AMD3100) in the same cancer type was initiated [50].

3.3.1 Setup

CAM-PLEX clinical trial enrolled patients with MSS-CRC, PDAC, and ovarian cancers. The aim of the phase I trial was two-fold: (i) study the molecular profiling of the underlying mechanisms to determine if AMD3100 helps the immune cells to intermingle with the cells of these cancers in humans, and (ii) find the highest safe dose of AMD3100 to administer. Cancer tissue biopsies and blood samples were taken from each enrolled patient before (day 0) and after (day 8) AMD3100 treatment. It should be noted that CAM-PLEX was a phase I single-agent only trial, i.e., patients were only administered AMD3100 and no other ICB therapies. The main scientific aim of this investigation was to assess whether the blockade of the CXCL12/CXCR4 axis would result in increased immune cell infiltration and activation in humans.

3.3.2 Dataset

In this thesis, I studied the RNA-seq data of the paired biopsies collected pre and post AMD3100 therapy from metastatic sites. Table 3.1 shows the dataset. I had paired-end RNA-seq reads from 21 patients, of which five had PDAC and 16 MSS-CRC. There were a total of 37 samples, as five MSS-CRC patients (pat06, pat07, pat09, pat13, patX) did not have post-therapy RNA-seq data.

Patient	Time	Pool	Barcode	Cancer	Read count
pat01	day0	SLX-13655	D710_D503	crc	28,607,517
pat01	day8	SLX-13655	D710_D502	crc	31,681,242
pat02	day0	SLX-13655	D711_D501	crc	32,135,594
pat02	day8	SLX-13655	D710_D508	crc	27,091,124
pat03	day0	SLX-13655	D711_D505	crc	28,080,365
pat03	day8	SLX-13655	D711_D507	crc	25,628,945
pat04	day0	SLX-13655	D712_D505	crc	29,641,275
pat04	day8	SLX-13655	D712_D508	crc	26,712,999
pat05	day0	SLX-13655	D712_D502	crc	29,390,888
pat05	day8	SLX-13655	D711_D508	crc	26,519,678
pat06	day0	SLX-13655	D701_D501	crc	25,628,169
pat07	day0	SLX-13655	D701_D502	crc	32,607,604
pat08	day0	SLX-13655	D712_D506	crc	27,660,974
pat08	day8	SLX-13655	pat08day8	crc	51,126,382
pat09	day0	SLX-13655	D702_D501	crc	28,176,280
pat10	day0	SLX-13655	D710_D506	crc	26,534,718
pat10	day8	SLX-13655	D710_D507	crc	28,231,905
pat11	day0	SLX-13655	D710_D501	crc	31,990,182
pat11	day8	SLX-13655	D709_D508	crc	31,006,244
pat12	day0	SLX-13655	D711_D503	crc	27,994,609
pat12	day8	SLX-13655	pat12day8	crc	54,299,083
pat13	day0	SLX-13655	D702_D502	crc	26,357,112
pat14	day0	SLX-13655	D711_D504	crc	27,177,453
pat14	day8	SLX-13655	D712_D501	crc	29,629,240
pat15	day0	SLX-13655	D711_D502	crc	29,361,589
pat15	day8	SLX-13655	D710_D505	crc	28,333,433
patX	day0	SLX-13655	D711_D506	crc	30,096,987
pat19	day0	SLX-15110	D701_D503	pdac	88,177,787
pat19	day8	SLX-15110	D702_D503	pdac	93,452,302
pat20	day0	SLX-15110	D701_D504	pdac	95,515,600
pat20	day8	SLX-15110	D702_D504	pdac	93,628,229
pat23	day0	SLX-16181	UDI0006	pdac	87,681,016
pat23	day8	SLX-16181	UDI0001	pdac	109,405,081
pat24	day0	SLX-16181	UDI0005	pdac	81,298,367
pat24	day8	SLX-16181	UDI0008	pdac	39,288,930
pat25	day0	SLX-16181	UDI0004	pdac	77,410,241
pat25	day8	SLX-16181	UDI0002	pdac	103,516,169

Table 3.1 CAM-PLEX trial dataset

3.4 BCR repertoire recovery from paired data

In Section 1.3, I separated the current strategies of AgR reconstruction into two: targeted repertoire sequencing-based protocols and computational reconstruction of unselected RNA-seq reads. Here I will present the justifications of the chosen method and describe the workflow.

3.4.1 Computational reconstruction of BCRs from RNA-sequencing reads

The CAM-PLEX samples were sequenced with whole-exome sequencing, RNA-sequencing, and AgR targeted sequencing, however only for TCRs. As I wanted to study the dynamic BCR repertoire, I opted to reconstruct the BCR clonotypes directly from the available unselected RNA-seq. Section 1.3.2 describes this approach and Section 1.3.5 presents a review of the state of the art in computational reconstruction tools.

I investigated whether BCR repertoire clonality, class-switching, and SHM load and location had an association with patient survival in melanoma. To this end, I used the TCGA SKCM (Skin Cutaneous Melanoma) RNA-seq dataset comprising 468 cases [165]. I reconstructed BCRs from this dataset using MiXCR and V'DJer and observed that MiXCR showed better computing performance. V'DJer requires the STAR aligner [57] to align unselected RNA-seq reads prior to the assembly of V-J contigs. With STAR, the alignment of a large set of 400+ data samples requires significant computation power. Additionally, V'DJer has to be run three times in order to analyze all IGH, IGK, and IGL chains, as tool execution expects a chain name as a parameter. Finally, V'DJer does not analyze TCR chains. I did not evaluate TRUST on the TCGA dataset as it was initially developed for TCR reconstruction and only recently provided BCR support. A benchmarking of MiXCR, TRUST, and V'DJer can be found in [24], with a follow-up in [101].

Based on computational performance and MiXCR specific features, I opted to use a pipeline running sequential MiXCR tools to reconstruct the BCR clonotypes from the CAM-PLEX unselected RNA-seq data. MiXCR can assemble full BCR and TCR CDR3 sequences from unselected/non-enriched RNA-seq data and identify SHMs. The developers of MiXCR added several enhancements in order to identify AgR clonotypes from unselected RNA-seq data. The original tool is described in [25], whereas the enhanced version is presented in [24].

MiXCR starts with aligning RNA-seq reads against reference V, D, J, and C genes. It uses an altered version of a multi-seed strategy, called seed-and-vote, described in [142].

With this strategy, from each read, many short, equal distanced subreads (seeds) that are mapped without mismatches are found, and the number of consensus subreads (votes) determines the best location for the read being aligned. Then using dynamic programming, the alignment is completed by filling in the mismatch and indel information between the subreads that make up the winning voting block. MiXCR has made further enhancements: (i) Modified the seed size setting default as 5 bp, in order to handle short alignments. (ii) In the voting step, a scoring function, calculated from the number of matching and absent seeds and offsets of seeds relative to the optimal position in the sequencing read, is maximized. Offsets are introduced to account for indels. After the selection of single or multiple candidates, alignments for between seeds (using the Needleman–Wunsch algorithm) and for outside seeds (using the Smith-Waterman algorithm) are built. Next, alignment scores (scoring matrix and penalties for indels) are calculated for all candidates to filter the best alignment. In [24], they further enhanced their alignment algorithm to account for cases where V or J segment alignment is ambiguous by switching to Smith-Waterman/Needleman-Wunsch algorithms directly in such cases. Performance is only hindered for these specific cases, but they achieve higher sensitivity. After alignment, two rounds of partial-contig-assembly are run. Here, aligned reads which cover only the left boundary of CDR3 are merged with reads which do not cover the left CDR3 boundary and cover part of the J gene. Merging is based on rules such that the minimum overlap region is 12 bp, the overlap covers at least 7 non-germline-derived bases, and sequences are a perfect match inside the overlap. This results in assembly of short, fragmented sequencing reads into longer reads/contigs which contain a CDR3. After these steps, MiXCR runs the assembly of aligned reads with clustering. Finally, the assembled contigs are exported. The code snippet in Listing 3.1 is from this study's pipeline and shows the MiXCR component assembling the BCR and TCR contigs:

```
# align RNA-Seq reads
mixcr align --library imgt -p rna-seq -s hsa --report report.align.txt
    -t 24 --save-reads -OallowPartialAlignments=true $read1 $read2
    alignments.vdjca

# assembly of short reads
mixcr assemblePartial alignments.vdjca alignments_rescued_1.vdjca
mixcr assemblePartial alignments_rescued_1.vdjca alignments_rescued_2.
    vdjca
mixcr extend alignments_rescued_2.vdjca alignments_rescued_2_extended.
    vdjca

```

```
# assemble default CDR3 clonotypes (--write-alignments is required for further full contig assembly)
mixcr assemble --report report.assemble.txt -t 24 --write-alignments alignments_rescued_2_extended.vdjca clones.clna

# assemble full BCR receptor sequences
mixcr assembleContigs --report report.assembleContigs.txt clones.clna full_clones.clns

# export
mixcr exportClones -c IG -p fullImputed full_clones.clns full_IG_clones.txt # full BCR receptors
mixcr exportClones clones.clna clones.txt # BCR and TCR assemblies
mixcr exportClones -c IG --filter-out-of-frames --filter-stops -count -fraction -vGene -dGene -jGene -cGene -nFeature CDR3 -aaFeature CDR3 clones.clna clones.IG.txt # custom BCR dataframe output
mixcr exportAlignments clones.clna alignments.txt # alignments
```

Listing 3.1 Code snippet for assembling AgR contigs from paired-end RNA-seq reads

where `$read1` and `$read2` are the paired-end reads coming from a single sample. I have passed `imgt` to the `library` parameter in `mixcr::align` in order to specify annotation with the IMGT database.

3.4.2 Analysis of reconstructed clonotypes

Table 3.2 is an example of an output from MiXCR exported via `mixcr::exportClones` passing the filtering parameters `-c IG --filter-out-of-frames --filter-stops` in order to write only Ig chains that are not out-of-frame and do not contain stop codons. Also, `-count -fraction -vGene -dGene -jGene -cGene -aaFeature CDR3` were specified in order to customize the list of fields that were exported. Table 3.2 shows the top six most clonal CDR3 Ig chains. `cloneCount` is the number of reads that make up a clonotype, `best<V,D,J,C>gene` is the best `<V,D,J,C>` hit gene name.

cloneCount	cloneFraction	bestVGene	bestDGene	bestJGene	bestCGene	aaSeqCDR3
115929	0.1455473	IGLV2-14		IGLJ3	IGLC2	CSSYAITNSLVF
57129	0.0717247	IGHV3-30	IGHD6-13	IGHJ4	IGHG2	CAKDLRGNSWSFDYW
35502	0.0445723	IGKV1-12		IGKJ5	IGKC	CQQAKSFPITF
30524	0.0383225	IGKV1-39		IGKJ4	IGKC	CQQSHTNPLTF
25389	0.0318755	IGKV1-39		IGKJ2	IGKC	CQQGYSSPYTF
17282	0.0216973	IGHV4-59	IGHD6-13	IGHJ2	IGHA1	CARGVSSSWYGPNWYFDLW

Table 3.2 Example of customized mixcr::exportClones output

MiXCR assembled a total of 1,522,531 BCR and TCR sequences, detecting 33,591 unique clonotypes. In all samples, the number of unique BCR assemblies was significantly

higher than TCR assemblies, which is indicative of intratumoral antibody-secreting plasma cells (see Figure 3.1).

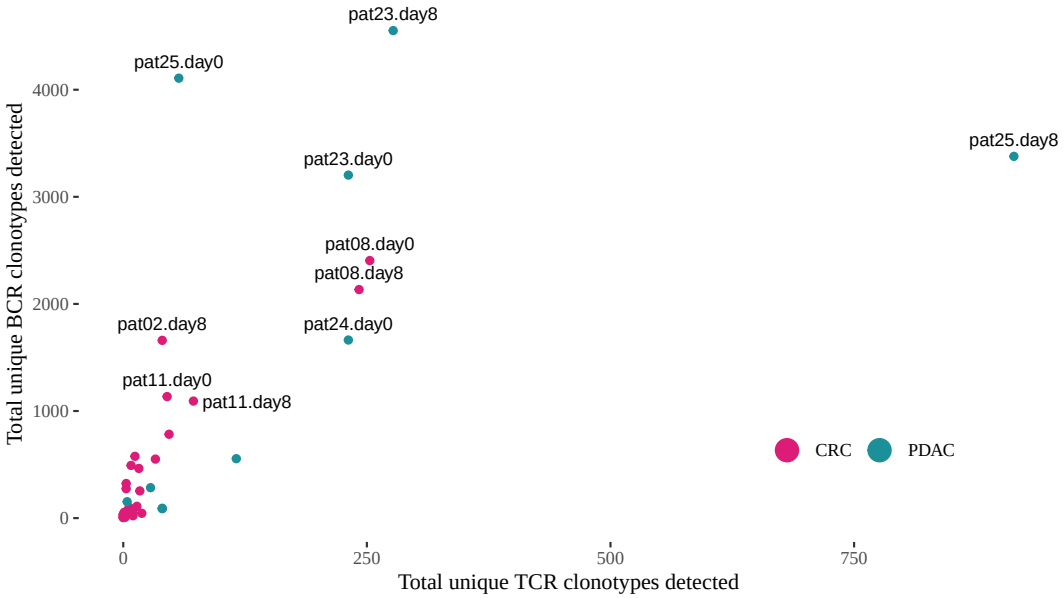


Fig. 3.1 Total number of uniquely detected BCR clonotypes plotted against the uniquely detected TCR clonotypes. These clonotypes include both heavy and light Igs, and α and β TCRs. Colors denote which type of cancer the sample belongs to. Annotated points are samples with >1000 detected BCR clonotypes. This threshold is chosen merely for ease of readability.

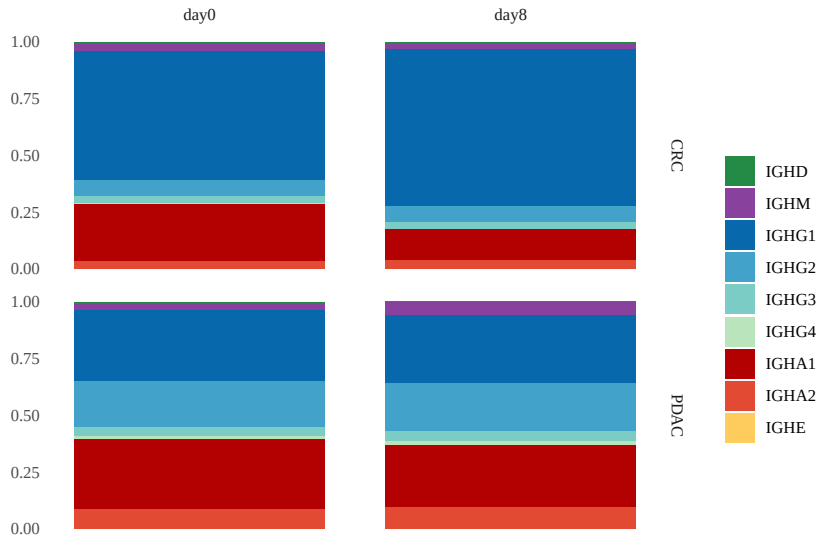


Fig. 3.2 Distribution of Ig isotypes across cancer types and time of sampling.

Of all the Ig isotypes, IgG was the most abundant class, followed by IgA, and IgM (see Figure 3.2). This shows that the majority of the detected Ig clonotypes are from the secondary immune response, where B cells have encountered an antigen and differentiated into plasma B and memory cells.

3.4.3 Normalizing clonotype count across samples

Looking at Table 3.1, we can see that some samples have relatively higher read counts. This is because they are a combination of multiple re-sequencing runs. I merged all sequencing reads per sample into a single read fastq file. Therefore we need to normalize for read count when calculating the clonotype count (CC).

To normalize clonotype counts across samples I simply defined *clone counts per million (CCPM)* reads which is clone counts scaled by the number of reads we sequenced, times one million. The CCPM of the i^{th} clonotype in the j^{th} sample is calculated by:

$$CCPM_i = \frac{CC_i}{N_j} 10^6 \tag{3.1}$$

where CC_i is the clone count of the i^{th} clonotype, and N_j is the number of fragments sequenced for sample j.

However, it should be obvious that using this scaling, we cannot compare repertoires across samples with highly variant read counts. Recall that MiXCR, and similar reconstruction tools, rely on the number of AgR reads in a sample. It is more likely to detect more unique AgR clonotypes the more RNA-seq reads a sample has. That is, the chance of detecting an AgR increases with the number of reads processed. Furthermore, MiXCR removes low-quality reads and performs error correction with clustering. The counts reflected as CC are not all the reads, but the reads selected after QC and merged after error correction. In order to accurately scale CCs to a comparable measure, one needs to downsample the aligned reads to an equal amount, and better yet, use bootstrapping [2] to obtain CC confidence intervals per single-chain clonotype. However, on inspection of Table 3.1 more closely, we see that the read count variation is between different cancer type samples (*crc* vs. *pdac*), whereas reads coming from the same cancer tissue are more or less the same. Although not 100% accurate, we can use CCPM when comparing samples within tumor types, but not between. Furthermore, it is not the focus of this thesis to compare different patient samples. I am interested in detecting an adaptive immune response and recovering expanded clonotypes within a single sample. Also,

[2] In statistics, bootstrapping is a resampling method used to estimate statistics of a population by sampling a dataset with replacement [60]. A confidence interval can be calculated to bound the estimate.

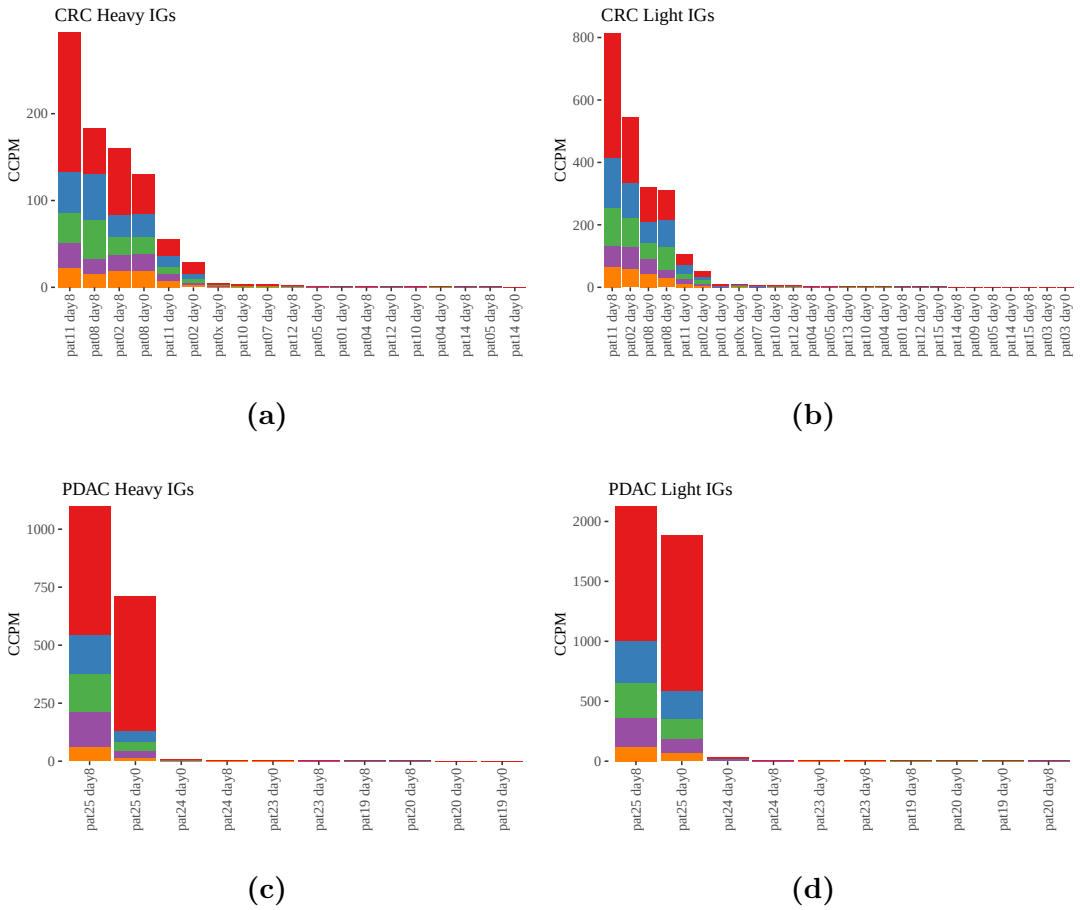


Fig. 3.3 CRC heavy (a) and light (b), and PDAC heavy (c) and light (d) clonotype frequency distributions. Here, for each sample, the CCPM of top five most clonal clonotypes are shown. Each color denotes a unique clonotype.

note that I am not calculating repertoire diversity/clonality. Though common in AgR repertoire analyses, it is not the focus of this thesis. In such cases, we would need to set all sample sizes to the size of the smallest sample to interpolate the diversity estimates with a number of resamplings.

3.4.4 Aggregating clonotypes

In the MiXCR output, there are clonotypes with the same CDR3 amino acid sequence, but different V, D, or J genes (see Table 3.3). The most likely interpretation of this result is that MiXCR is mis-annotating some of the genes, as it is very unlikely for an entirely different VDJ combination to have the same CDR3 amino acid sequence; this would require an incredibly strong evolutionary convergence. Moreover, as we see this

happen in almost every sample, for many different Ig chains, I think that a technical artifact remains the most likely explanation.

cloneCount	cloneFraction	Best V;D;J;C gene	aaSeqCDR3
2432	23.493	IGHV4-61;IGHD3-3;IGHJ3;IGHG2	CAREPLEFEGDFGPYDIW
78	0.753	IGHV4-34;IGHD3-16;IGHJ3;IGHG2	CAREPLEFEGDYGPYDVW
23	0.222	IGHV4-31;IGHD1-26;IGHJ3;IGHG2	CAREPLRFEGDYGAFDVW
16	0.154	IGHV4-4;IGHD3-16;IGHJ3;IGHG2	CAREPLEFEGDYGPYDVW
12	0.116	IGHV4-31;IGHD3-16;IGHJ3;IGHG2	CAREPLEFEGDFGPYDIW
8	0.077	IGHV4-28;IGHD3-16;IGHJ3;IGHG2	CAREPLEFEGDYGPYDVW
2	0.019	IGHV4-31;IGHD1-14;IGHJ3;IGHG2	CAREPLRFEGDYGAFDVW

Table 3.3 Example of incorrect clonotype annotation. Clonotypes with the same row color have the same CDR3 amino acid sequence but differing VDJ gene combinations. Note that *cloneFraction* is the fraction of the chain in the whole repertoire.

MiXCR determines clonotype uniqueness based on CDR3 nucleotide sequences. Instead, I define two clonotypes to be the same if they have identical CDR3 amino acid sequences. Hence, I further merged all clonotypes by CDR3 amino acid sequences. More abundant $clone_i$ will absorb less abundant $clone_j$ if clone count $N_j <$ clone count N_i, and both $clone_j$ and $clone_i$ have the same CDR3 region. This way we will be creating a more compact clonotype list for each sample.

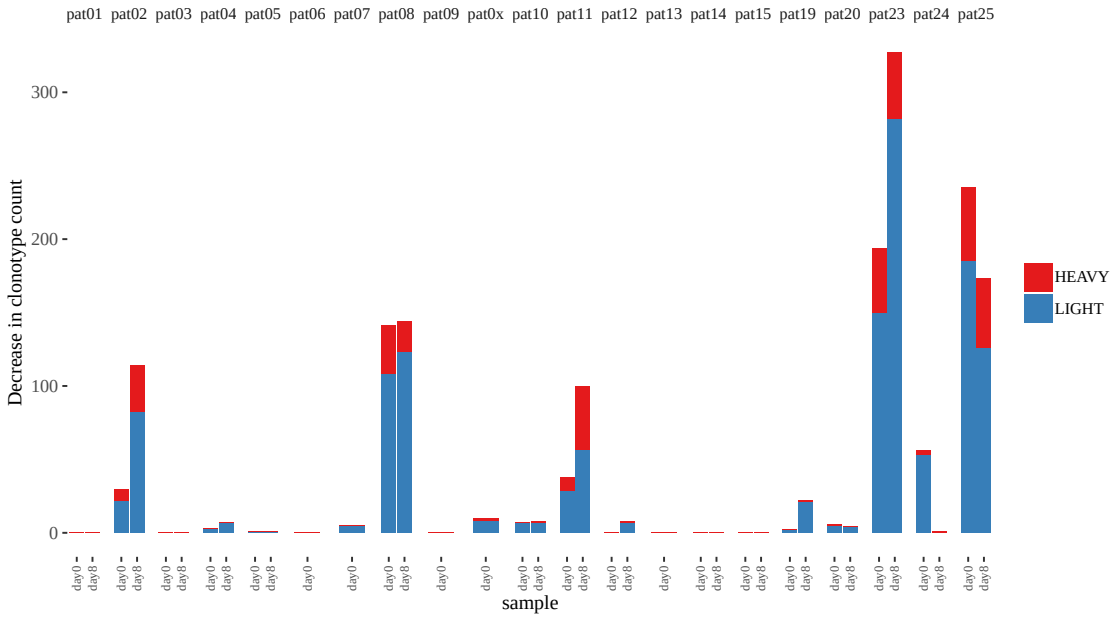


Fig. 3.4 Decrease in clonotype count across samples after aggregation by CDR3 amino acid sequence. Color separation is based on chain type where heavy chains are colored red and light chains blue. Certain samples did not experience any decrease.

3.5 Results

I define clonal expansion as the difference in the CCPM (3.1) of a clonotype between post (day8) and pre (day0) immunotherapy. For this analysis, I am interested in detecting Ig repertoires showing high B cell infiltration and clonal expansion, both of which are summarised in Figure 3.5. From the plot we can immediately observe that patients 2, 8, and 11 from the CRC cohort, and 25 from the PDAC cohort show the highest B cell infiltration, as well as the most clonally expanded clonotypes.

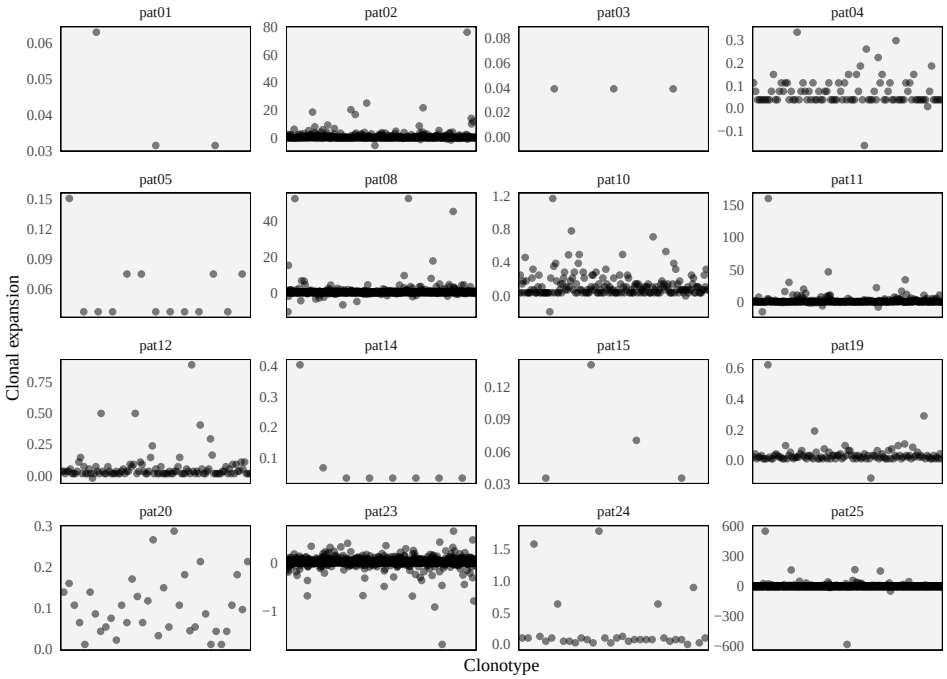


Fig. 3.5 Clonal expansion of all Ig heavy clonotypes in a patient's pre- and post-immunotherapy repertoire. Each point signifies the difference between post- and pre-immunotherapy CCPM of a clonotype. Note that the y-axes vary across patients.

Here, observe, for example, why patients 23 and 24 were not chosen. Both patients had a high count of uniquely detected BCRs (see Figure 3.1). However, when we further inspect the sum of the most clonally expanded clonotypes' CCPM, we see that these patients had a diverse BCR repertoire, i.e., even though they have a significant amount of BCR clonotypes, none of them have expanded, as we can clearly see in Figures 3.5 and 3.6. We can observe that patient 25 has a high count of uniquely detected IgH clonotypes as well as a high sum of CCPM, meaning its top clonal clonotypes expanded significantly. This is not true for patients 23 and 24.

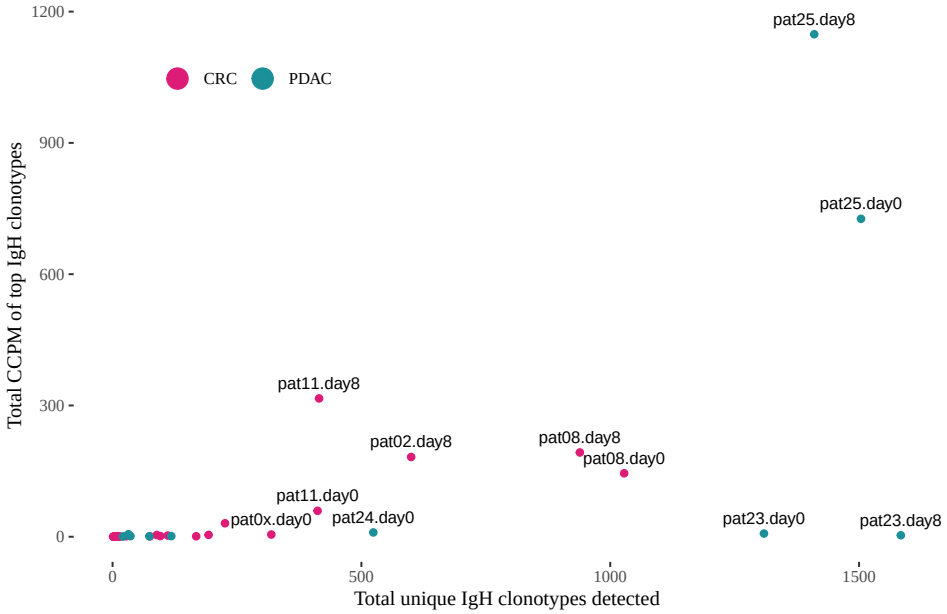
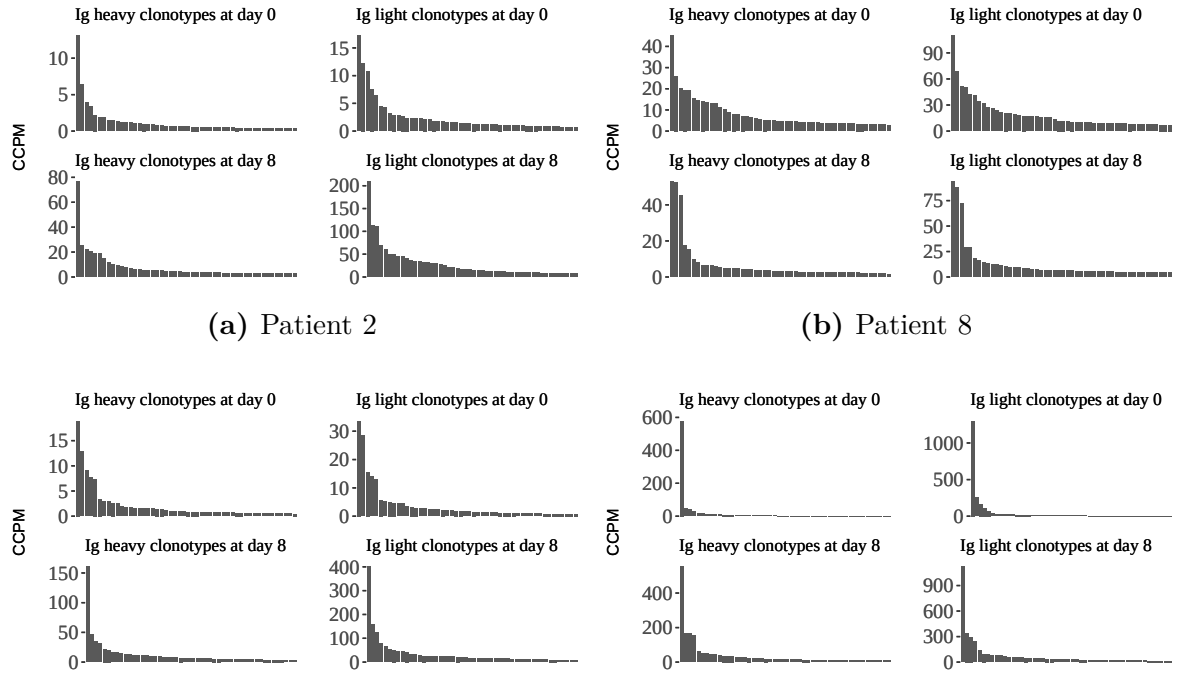


Fig. 3.6 Sum of CCPM of top IgH clonotypes plotted against total uniquely detected IgH clonotypes. Colors denote which type of cancer the sample belongs to. Annotated points are samples with >350 detected IgH clonotypes. This threshold is chosen merely for ease of readability.

Furthermore, although patient 23 had an increase in uniquely detected BCRs, the clonality of these clonotypes did not increase. However, do note that this could be due to the increase in sequencing read count (see Table 3.1). The `pat08.day8` sample has roughly two-fold read count compared to most other CRC samples, due to multiple re-sequencing. However, the same patient at day 0 also has a high CCPM. Hence it is justifiable to select patient 8 as a strong immune response candidate.

3.5.1 BCR distributions

Here I present the Ig heavy and light chain CCPM distributions of the selected patients. In Figure 3.7, I have only plotted the most clonal 50 clonotypes for each sample.



(a) Patient 2 **(b)** Patient 8

(c) Patient 11 **(d)** Patient 25

Fig. 3.7 Ig heavy and light chain CCPM distributions of the selected patients. Repertoires are of pre-treatment (day 0) and post-treatment (day 8). For each repertoire, the most clonal 50 clonotypes are plotted. Note that the y-axes vary across patients, days, and chain types.

3.5.2 Pre and post-therapy BCR repertoires

Here, for the selected patients, I present the possible tumor-antigen specific antibody sequences, and investigate the clonal expansion of the most clonal Igs in the post-immunotherapy Ig repertoire. In some cases, the most clonally expanded clonotype of

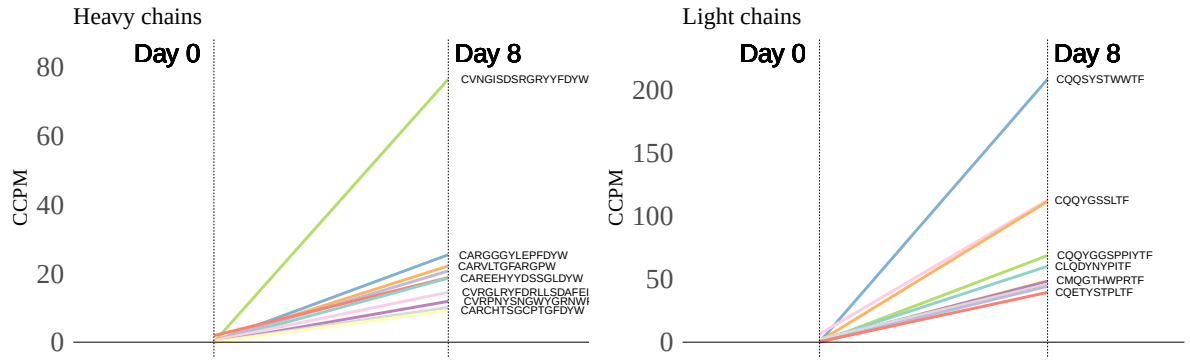
the post-immunotherapy repertoire is not present in the pre-immunotherapy repertoire. This could be due to a lack of sequencing reads in the pre-treatment repertoire. We could speculate that, although the Ig was expressed, it could not be picked up by MiXCR as there were not many cells expressing that particular Ig mRNAs. We might recall that MiXCR needs a certain number of assemblies in order to error correct for PCR and sequencing noise. Furthermore, the low signal in the pre-immunotherapy repertoire could also be due to read coverage. Observe that for each of the samples there are other dominant clonotypes, and these could be dominating the read space. Hence the first explanation is that the clonal Igs of post-immunotherapy are, in fact, present pre-immunotherapy as well, but either the RNA-seq or MiXCR could not detect them. The second explanation is that the plasma cells generating these antibodies were not in the tumor pre-immunotherapy, or they were present in minute amounts. I discuss this further in Section 3.6.

Sequences of reconstructed antibody chains

Below are the CDR3 nucleotide sequences of the possible pairings of Ig heavy and light chains. It is conceivable that these antibodies, if expressed, could help us identify PDAC and CRC specific antigens. The full nucleotide sequences are given in Appendix A.

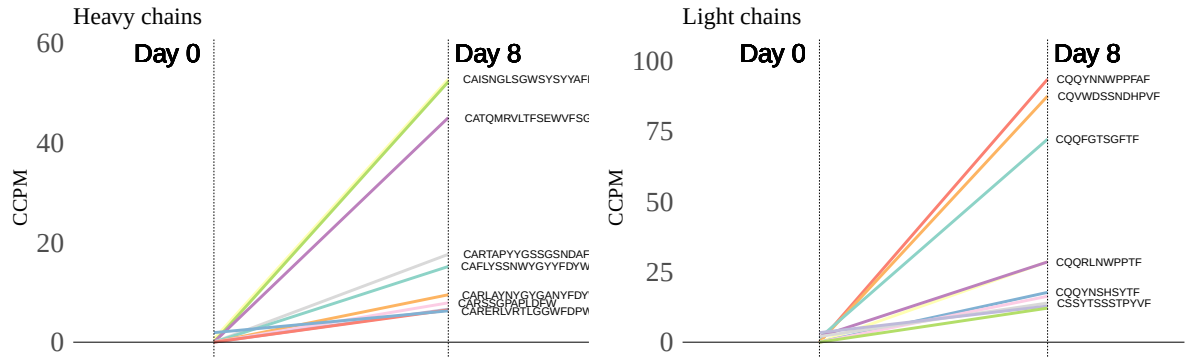
```
> pat02 Heavy chain CDR3
TGTGTGAACGGCATATCAGATAGTCGTGGTCGCTACTACTTTGACTACTGG
> pat02 Light chain CDR3
TGTCAACAGAGTTACAGTACCTGGTGGACGTTC
> pat11 Heavy chain CDR3
TGTGCGAAAGAAGAACTACGACGAGGGCCCCATGACTACTGG
> pat11 Light chain CDR3
TGTCAACAGAGTTACATTACCCCTCGGACGTTC
> pat25 Heavy chain CDR3
TGTGCGAAGGATTTGAGAGGCAACAGCTGGTCTTTTGACTACTGG
> pat25 Light chain CDR3
TGCAGCTCATATGCAATCACTAATTCTCTCGTTTTC
```

Listing 3.2 Sequences of clonally expanded Igs



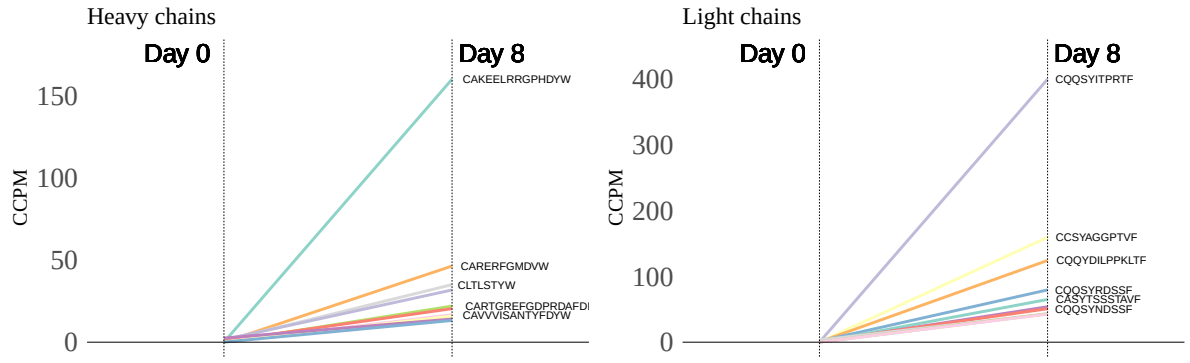
	CDR3	CCPM day0	CCPM day8	CC day0	CC day8
Ig Heavy	CVNGISDSRGRYYFDYW	0.00	76.48	0	2072
	CARGGGYLEPFDYW	0.31	25.43	10	689
	CARVLTGFARGPW	0.47	22.18	15	601
	CARDSFRGVFDYW	0.37	20.71	12	561
	CAREEHYYDSSGLDYW	1.90	18.79	61	509
	CAKSASFDNW	0.03	18.64	1	505
	CVRGLRYFDRLLSDAFEIW	0.46	14.54	15	394
	CVRPNYSNGWYGRNWFDPW	0.00	11.88	0	322
	CVRGPYSAGWFDPW	0.09	10.11	3	274
	CARCHTSGCPTGFDYW	0.00	9.23	0	250
Ig Light	CQQSYSTWWTF	0.00	208.74	0	5655
	CQQYGSSLTF	6.44	112.80	207	3056
	CQQYDNLPPFTF	0.81	111.71	26	3026
	CQQYGGSPPIYTF	0.00	68.80	0	1864
	CLQDYNYPITF	1.09	60.28	35	1633
	CMQGTHWPRTF	0.28	49.65	9	1345
	CQQYNSYPRTF	1.27	48.43	41	1312
	CQQYNNWPPLTF	2.37	45.96	76	1245
	CNSRDSGGNLVVF	0.00	44.37	0	1202
	CQETYSTPLTF	0.34	39.53	11	1071

Fig. 3.8 The change in CCPM and CC of the most clonally expanded clonotypes detected in the post-treatment sample of patient 2.



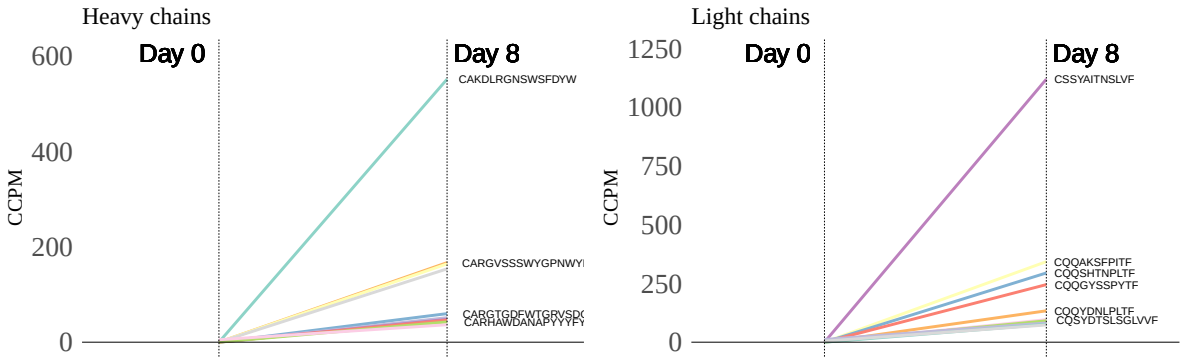
	CDR3	CCPM day0	CCPM day8	CC day0	CC day8
Ig Heavy	CAISNGLSGWSYSYYAFDVW	0.61	52.81	17	2700
	CARLRVSSVRGVPYFDSW	0.00	52.28	0	2673
	CATQMRVLTFSEWVFSGAYDVW	0.00	45.08	0	2305
	CARTAPYYGSSGSNDAFDLW	0.00	17.64	0	902
	CAFLYSSNWYGYYFDYW	0.00	15.19	0	777
	CARLAYNYGYGANYFDYW	0.00	9.52	0	487
	CARSSGPAPLDFW	0.00	7.90	0	404
	CAKDLWFGELYEMPFESW	0.00	6.63	0	339
	CAKDRGYSSGWKYFNYGLDVW	0.00	6.47	0	331
	CARERLVRTLGGWFDPW	1.92	6.29	53	322
Ig Light	CQQYNNWPPFAF	0.00	93.53	0	4782
	CQVWDSSNDHPVF	0.000000	87.52820	0	4475
	CQQFGTSGFTF	1.482233	72.21321	41	3692
	CQQRLNWPPTF	0.000000	28.61536	0	1463
	CVLYMGSGIVVF	1.988361	28.55669	55	1460
	CQQYNSHSYTF	0.000000	17.72079	0	906
	CSSYSGTSSVLF	0.000000	16.33208	0	835
	CSSYTSSSTPYVF	2.024513	13.92627	56	712
	CQQYDNLITF	3.398290	12.85051	94	657
	CQVWDSTSDRPFF	0.000000	12.10725	0	619

Fig. 3.9 The change in CCPM and CC of the most clonally expanded clonotypes detected in the post-treatment sample of patient 8.



	CDR3	CCPM day0	CCPM day8	CC day0	CC day8
Ig Heavy	CAKEELRRGPHDYW	0.00	160.29	0	4970
	CARERFGMDVW	0.00	46.41	0	1439
	CLTLSTYW	0.94	35.09	30	1088
	CARDFRHFDYW	1.56	31.83	50	987
	CARTGREFGDPRDAFDIW	0.06	21.99	2	682
	CARDLSSTYGMDVW	0.68	20.44	22	634
	CAVVVISANTYFDYW	0.18	16.70	6	518
	CARARAYSGYGPVDYW	0.00	15.93	0	494
	CRSGGDADDFDIW	2.50	14.12	80	438
	CARDLWRGGGSSFGYW	0.06	13.25	2	411
Ig Light	CQQSYITPRTF	0.032	399.92	1	12400
	CCSYAGGPTVF	0.00	159.49	0	4945
	CQQYDILPPKLTF	0.50	124.26	16	3853
	CQQSYRDSSF	1.53	79.53	49	2466
	CASYTSSSTAVF	0.00	64.95	0	2014
	CQVWDDSSDCVF	0.25	53.82	8	1669
	CQQSYNDSSF	1.15	50.90	37	1578
	CQSYDSSLSGYVF	3.28	43.25	105	1341
	CQQYYKIPWTF	0.09	43.18	3	1339
	CQQYYSTPWTF	0.00	42.70	0	1324

Fig. 3.10 The change in CCPM and CC of the most clonally expanded clonotypes detected in the post-treatment sample of patient 11.



	CDR3	CCPM day0	CCPM day8	CC day0	CC day8
Ig Heavy	CAKDLRGNSWSFDYW	0.00	551.89	0	57129
	CARGVSSSWYGPNWYFDLW	0.00	166.95	0	17282
	CARAGRVDGLPALFDPW	0.00	164.26	0	17004
	CARSPVAVAGTPDVFDIW	0.00	153.84	0	15925
	CARGTGDFWTGRVSDGFDVW	0.00	59.95	0	6206
	CARDGLGDDVTRFSWDAYDIW	0.00	51.38	0	5319
	CSGWWTLNWFDPW	0.00	47.89	0	4958
	CARENWKVVDVW	0.29	46.16	23	4779
	CARHAWDANAPYYYFYYPMDVW	0.00	42.46	0	4396
	CARHVLEGVPGNVDFW	5.15	36.64	399	3793
Ig Light	CSSYAITNSLVF	0.00	1119.91	0	115929
	CQQAKSFPITF	0.00	342.96	0	35502
	CQQSHTNPLTF	0.00	294.87	0	30524
	CQQGYSSPYTF	0.35	245.26	27	25389
	CQQYDNLPLTF	0.14	133.61	11	13831
	CQSYDTSLSGLVVF	0.00	95.83	0	9920
	CQQYGSSSWTF	0.00	91.09	0	9430
	CQQFGASPHSF	11.64	82.01	901	8489
	CGTWDSSLSAGWVF	0.00	76.15	0	7883
	CSSFTTSTTLDVVF	4.51	73.55	349	7614

Fig. 3.11 The change in CCPM and CC of the most clonally expanded clonotypes detected in the post-treatment sample of patient 25.

3.6 Discussion

Intratumoral B cells are a critical component of the TME, and they have been observed in multiple tumor tissues, including breast, colorectal cancers, and melanomas [128], [151], [216]. B cells recognize tumor antigens mainly via the CDR3 region on their antigen receptors. Upon antigen challenge, B cells undergo SHMs and class-switching in order to differentiate into plasma B cells, which begin secreting vast quantities of high-affinity antibodies in order to eliminate the tumor cells.

CRC and PDAC are two cancer types that are resistant to popular ICIs. However, it has been shown that this immunosuppression can be overcome by targeting other molecular targets prior to standard ICIs in mouse. As previously mentioned, [65] demonstrated that administering PDAC-bearing mice with AMD3100, an inhibitor of the CXCL12 receptor, CXCR4, revealed the anti-tumor effects of the anti-PD-L1 antibody and significantly reduced cancer cells.

Efforts have been made to study the TCR repertoire in PDAC [14], [99], however, BCR repertoire recovery in PDAC tumors after immune treatment, to the best of my knowledge, has not been undertaken before.

With tumor tissue samples collected from CRC and PDAC patients before and after AMD3100 treatment as part of a phase I clinical trial [50], I reconstructed BCR sequences directly from unselected RNA-seq data. Analyzing the BCR repertoire, I observed that the number of unique BCR assemblies was significantly higher than that of TCR assemblies, and that the detected BCRs had undergone class switching, which suggested the presence of intratumoral antibody-secreting plasma cells. Furthermore, in some patients, I detected clonal expansion of individual antibodies after treatment with AMD3100. This finding is in alignment with a recent study where, before and after PD-1 inhibitor therapy, RNA-seq data of seven patients with glioblastomas were processed with MiXCR to recover TCR and BCR repertoires [266]. Patients who did not show an objective response to therapy had a greater increase in clonal diversity among B and T cells relative to patients who responded to therapy [266].

Based on our findings, we can hypothesize that administering AMD3100, which blocks the CXCL12-CXCR4 axis, to PDAC patients may have, directly or indirectly, increased the level of intratumoral immune response, thus allowing for the effective differentiation of plasma B cells post-treatment. Furthermore, AMD3100 may have also allowed for better homing of plasma cells into the TME and let plasma cells persist longer in the TME and patient bone marrow. Another possibility is that both quantitative and qualitative improvement of B cell function, i.e., increased Ig production levels and the ability to produce more specific antibodies, may have been achieved as AMD3100

recovered T helper cells, and through them, APC function. Thus, B cells were called to the TME more efficiently, and they were presented with antigens by APCs more efficiently. This would need to be supported by looking at the effects of AMD3100 treated CXCL12/CXCR4 axis on APCs. Furthermore, it has been shown that, in mice, AMD3100 redistributes B cells from the bone marrow into the circulation [146]. It is possible that the administration of AMD3100 induced a redistribution of B cells from the patients' bone marrow into the tumor. These theories could potentially explain the abundance of antibodies post-treatment, but more functional studies are required to determine the mechanisms behind my results.

The absence of abundant post-treatment clonotypes in the pre-treatment repertoire can be due to these clonotypes having shallow signals falling below detection limits. These Igs may have existed pre-treatment, but it could not be picked up by the reconstruction algorithm as there were not many cells expressing those Igs. Furthermore, the low signal in the pre-immunotherapy repertoire could have also gone undetected due to read coverage. Recall that for each of the samples, there are other dominant clonotypes. These clonotypes could be dominating the read space. Hence the clonally expanded Igs of post-immunotherapy are, in fact, present pre-immunotherapy as well, but either the RNA-seq or reconstruction algorithm could not detect them. This explains the technical reason why we could not pick up the BCRs. But why are these BCRs so rare before therapy? Low counts of plasma cells could perhaps explain the reason behind the low signal in the TME, which could again be due to obstructed homing, hindered T cell calling, and APC presentation, or impaired class-switching or affinity maturation, possibly due to the immunosuppressive action of the CXCL12/CXCR4 axis in this context.

In the absence of an abundant post-treatment clonotype in the pre-treatment repertoire, as future work we can search for such BCRs in the pre-treatment dataset by direct alignment. As we have the full recombined and affinity-matured antibody sequence, the task at hand would be reduced to a straightforward alignment of reads to a reference sequence. For this, we would append the reconstructed post-treatment BCR sequences to the reference genome, as reads, in general, tend to align to non-specific regions as well. Another approach could be to pool together reads obtained pre- and post-AMD3100 treatment for each patient in order to intensify the signal of lowly observed clonotypes within each collection time. However, using this approach we would be making the assumption that an antibody present pre treatment is also to be present post treatment, and vice versa, which may or may not be true. Therefore we need to take into account that certain erroneous sequences might be regarded as true clonotypes as the true and high signal in one collection can increase the signal of the not only lowly present clonotypes

but also of the sequences reconstructed from erroneous reads. This can be mitigated by setting a reasonable base threshold on true clonotypes. Nevertheless, this is an approach that we can further investigate.

One of the ultimate goals in personalized medicine in cancer is to develop patient-tailored targeted cancer immunotherapy treatments. This requires finding the target which will distinguish cancer cells from normal cells. In this study, I was able to detect certain antibodies that were significantly clonally expanded after treatment. In the post-immunotherapy BCR repertoire of patients 2, 11, and especially 25, I detected highly clonal heavy and light Ig chains, and these chains are clonal enough to pair. With the sequences of these paired heavy and light Ig chains, we can have the specific antibody expressed in mammalian cells. These antibodies could then be used to detect cancer antigens targeted by the immune response after immunotherapy.

Although reconstructing BCR repertoires from unselected RNA-seq allowed for the detection of clonally expanded heavy and light Ig chains, we cannot always pair them with certainty. While highly clonal heavy and light chains may indicate a possible pairing, in other cases, the pairing may not be as obvious. Furthermore, using RNA-seq approaches, we cannot determine the activation status of cells that express a certain BCR. Based on these limitations of RNA-seq, it would be more beneficial to use 10x Genomics Chromium Single Cell Immune Profiling Solution in order to retrieve the gene expression profiles and VDJ sequences simultaneously from single B cells before and after AMD3100, and possibly anti-PD-L1, immunotherapy. With the gene expression data, we can obtain the activation and differentiation status of each cell by investigating specific biomarkers, and therefore identify Igs which recognize specific antigens in the tumor. These antibodies would, in turn, allow for the identification of antigen candidates likely to drive the intratumoral adaptive immune response.

Chapter 4

Tumor-reactive immune cell clonality

4.1 Introduction

In the previous two chapters, my work mainly focused on B lymphocytes. The other major actor of the adaptive immune system is the *T lymphocyte*. T (thymus-derived) lymphocytes defend the host against unwanted intracellular events; they recognize and attack infected cells such as those that have been penetrated by a virus or bacteria, or cancerous cells that deviate from healthy self cells, or non-self cells from transplants. T cells recognize the antigens on the surface of such cells using T cell receptors (TCRs) [117, Chapter 1, p.10].

In the present study, I link the TCR repertoire with whole-transcriptome scRNA-seq data. Combining the two at a single cell level allowed me to investigate the relationship between T cell transcriptional phenotypes and TCR repertoires. Furthermore, by examining the transcriptional profile of both intratumoral and splenic T cells that could be linked via shared TCR clonotypes, I was able to study the differences between tumor-infiltrating T cells and the T cells in the spleen.

In Section 4.2, I give an overview of how the T cells generate and develop in the primary lymphoid organs, describe their maturation and differentiation process, and briefly explain their activation. I then discuss T cell diversity and TCRs focusing on $CD8^+$ cytotoxic T lymphocytes. I explain how tumor cells can impair $CD8^+$ tumor-infiltrating lymphocytes in order to suppress immune responses, and how current immunotherapy methods are trying to reverse this process. Later in Section 4.3, I describe our approach that combines single-cell gene expression profiles with TCR information at the cellular level in order to uncover the links and transitions between transcriptional states. I then

present a novel transgenic mouse model, developed by Dr. James Thaventhiran, which can help track the clonal $CD8^+$ T cell response to tumor antigens. I describe how, using this mouse model, Dr. James Thaventhiran and Mr. Ty So were able to harvest and sequence tumor-reactive T cells, and how I obtained their gene expression profile and TCR repertoires from the scRNA-seq data. In Section 4.4, I give a detailed account of my analysis on the obtained dataset from the single-cell RNA sequencing of the mouse model, focusing on investigating the clonality and diversity of T lymphocytes. In Section 4.5, I present our findings and end this chapter with a discussion in Section 4.6.

4.2 Background

4.2.1 T cell development

Before starting, it is worth noting that the development of T cells is a complex and fascinating process, yet not completely understood. In the following paragraphs I will attempt a summary of the most consolidated knowledge on this topic. However, a complete review of the matter would be beyond the scope of this thesis, and is available elsewhere [125], [122], [117, Chapter 10, p. 221-229], [3, Chapter 4, p. 42-58].

T lymphocytes originate from ancestral lymphoid progenitor cells which differentiate from hematopoietic stem cells found in the bone marrow [3, Chapter 4, p. 42-58]. Progenitors of T cells migrate to the thymus. The thymus is a primary lymphoid organ of the immune system which "trains" the progenitor T cells into mature T cells. Thymocytes, the hematopoietic progenitor cells present in the thymus, go through multiple stages of development. These stages are characterized by the expression of specific cell surface markers. Early-stage thymocytes generally lack the cell-surface proteins CD4 and CD8, and are called *double negative* (DN) cells. Note that CD4 and CD8 are the two essential markers when discussing T cell development, but various other markers are also crucial in distinguishing between stages: CD44 and CD25 markers further characterize the DN population [212]. I have briefly illustrated the T cell developmental stages that take place in the thymus in Figure 4.1. The somatic recombination of the genes that encode the receptor proteins take place during the DN stage. First, TCRδ and TCRγ, as well as the TCRβ gene segments rearrange: genes that encode the variable domain of TCRs are in an array of separate germline DNA gene segments and are recombined and joined together [110]. If the developing cell makes a productive $\gamma\delta$ gene rearrangement, it expresses a TCR$\gamma\delta$ receptor. Alternatively, if the cell completes a productive β gene rearrangement, then the TCRβ chain is expressed on the cell surface coupled with a surrogate chain

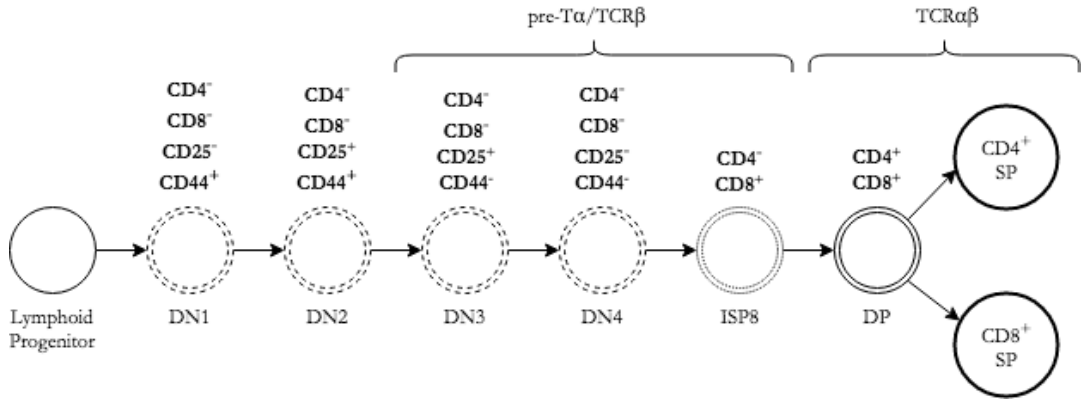


Fig. 4.1 T cell developmental stages in the thymus. Thymocytes differentiate from immature DN to immature single CD8 positive (ISP8), DP, and single-positive thymocytes. Each DN stage is characterized based on expression of cell surface markers CD25 and CD44: DN1 ($CD25^{-}CD44^{+}$), DN2 ($CD25^{+}CD44^{+}$), DN3 ($CD25^{+}CD44^{-}$), and DN4 ($CD25^{-}CD44^{-}$). My illustration here is based on details from [212]. Also note that the cell surface markers CD44 and CD25 are expressed in murine thymocytes and the corresponding human DN stages are characterized by the differential expression of CD34, CD38 and CD1A [110].

called pre-Tα, producing the pre-TCR complex. If neither of these gene rearrangements happen productively, the cell undergoes apoptosis [212].

Expression of pre-TCR stimulates recombination in the TCRα locus of these developing cells, induces their proliferation and the expression of CD4 and CD8 on their surfaces, allowing them to develop into $CD4^{+}CD8^{+}$ *double positive* (DP) thymocytes committed to the $\alpha\beta$ lineage. Next, the TCRα chain pairs with the TCRβ to form a functional signaling complex. At this point, DP thymocytes are $TCR\alpha\beta^{+}CD3^{+}CD4^{+}CD8^{+}$ [212]. CD3 serves as a T cell co-receptor that associates with the TCR and is a defining marker of the T cell lineage [117, Chapter9, p. 212].

As a next step, maturing T cells undergo two important processes: *positive* and *negative selection*. Positive selection only retains thymocytes which bear receptors that can recognize self-MHC molecules, while negative selection discards any thymocyte with a receptor that recognizes self-MHC molecules or self-antigens presented by self-MHC with high affinity. Cells that do not pass positive selection undergo apoptosis. The remaining cells go through negative selection ensuring that only the self-tolerant thymocytes survive. However, this poses the problem that negative selection would also remove any thymocyte that recognizes self MHC molecules, and therefore deplete the T cell population. Different hypotheses have been proposed to explain this contradiction, and are reviewed extensively in [258]. The exact process is currently a matter of debate, and it is possible that multiple mechanisms play a role in thymic selection [117, Chapter 10, p. 223-229].

Thymic selection further ensures that $CD4^+CD8^+$ DP thymocytes differentiate into $CD4^+CD8^-$ or $CD4^-CD8^+$ (SP, *single positive*) cells. However, the exact process is not yet established and several models have been proposed. For example, while one model suggests that the ability of a DP thymocyte to recognize MHC class I or class II molecule dictates whether it will commit to the CD8 or CD4 lineages respectively, some others suggest that this progression is stochastic and independent of T cell receptor MHC specificity. A review of the major models of CD4/CD8 lineage choice, including the aforementioned can be found in [220]. Additionally, it should be noted that $CD4^+CD8^+$ DP mature T cells have been reported in various disease settings [171], including melanoma [55].

Having gone through thymic selection, mature and functional T lymphocytes exit the thymus, traveling to peripheral tissues, or circulate in the blood or lymphatic system, ready to be activated upon encounter with the appropriate antigenic stimulus.

4.2.2 T cell activation

Mature lymphocytes are called *naive* (immunologically inexperienced) until they encounter the antigen for which they are specific. T cells do not directly interact with antigens but with *antigen presenting cells* (APCs). The interaction between antigen-specific T cell receptors and antigen-major histocompatibility complex (MHC) molecules found on the surfaces of APC initiates the activation of naive T cells. A small number of T cells that express the $\gamma\delta$ receptor do not require display by MHC proteins [126, Chapter 6, p. 188-198].

Naive T cells that exit the thymus travel to the periphery where they can encounter a multitude of APCs and get activated. If not, they keep circulating between the lymph and blood until a possible encounter occurs [34, Chapter 24, p. 1311-1313].

T cell activation leads to the proliferation and differentiation of naive T cells into *effector* cells. Upon activation, $CD8^+$ naive T lymphocytes may differentiate into *cytotoxic* T lymphocytes (CTLs or *Tc*) which have the ability to directly kill infected or neoplastic cells. Activation of $CD4^+$ T cells differentiates them into $CD4^+$ *helper* T cells (*Th*) and may initiate several downstream events, including synthesis of *cytokines*, signaling molecules, and the recruitment and activation of other immune cells. Another type of essential immune cells called *regulatory* T cells (*Treg*) have the phenotype $CD4^+CD25^+$ [196]. These cells are thought to monitor and control immune responses, thus preventing excessive immune activation and limiting potential collateral damage to healthy tissue. *FOXP3*, a transcription factor that confers immune suppressor function in this context, is commonly used as a marker to distinguish T regulatory cells from effector T cells [95].

Helper T cells can further differentiate into more specialized effector cells such as Th1, Th2, Th17, Th9, among others [28]. This subsetting is mainly based on the cytokine expression profile of the cell. Each helper T cell subtype performs a different function to regulate the immune response. Immune cell differentiation is not yet fully characterized; however, recent advances in single-cell genomics are helping distinguish further specialized subpopulations [209].

4.2.3 T cell diversity

T cells present a vast array of highly specific antigen receptors on their surface to detect pathogens. Each T cell expresses a specific T cell receptor as it develops into a functionally competent cell. A majority of the T cells (95%) express a combination of α and β chains [126, Chapter 6, p. 190]. The TCR$\alpha\beta$ recognizes antigens which are presented by MHC molecules on the surfaces of antigen-presenting cells [193]. A small population of TCRs expresses TCR$\gamma\delta$ which can recognize non-protein molecules such as lipids, without the need for display by MHC molecules [126, Chapter 6, p. 191].

Each TCR chain has a *variable* domain which recognizes and binds to the antigen, and a *constant* domain. TCR α and δ chains' variable domain is encoded by one of many V (variable) and J (joining) genes, and TCR β and γ chains have another D (diversity) gene between the two domains (Figure 4.2). During V(D)J recombination, one allele of V, D, and J genes recombine, forming the functional variable region. This variable region, combined with a constant gene segment, makes up the functional TCR chain [227]. Insertion and deletion of random nucleotides at the V-D, D-J, and V-J junction sites introduce junctional diversity. Later, the pairing of α and β or γ and δ chains provide an additional layer of diversity. This combinatorial, junctional, and coupling diversity creates an incredibly vast TCR repertoire guaranteeing the recognition of millions of antigens, including those of neoplastic cells. Analyzing this diversity can be crucial in assessing the immune response to cancer.

TCR chains contain three complementarity determining regions: *CDR1* and *CDR2* are encoded solely by V genes in the germ-line DNA segments, whereas *CDR3* is encoded by the terminal of V genes, the D genes in the case of TCRβ chains, the beginning of J genes, and the junctional sites between those genes, making it the most variable region in a TCR chain [117, Chapter 9, p. 207]. It is also the region which is primarily in contact with the peptide antigen presented by an APC [74]. Due to these characteristics, TCR clonal lineages can be identified by their CDR3 regions. When the receptor of a T cell detects a specific antigen and binds to it, that cell might receive a proliferation signal, leading to clonal expansion.

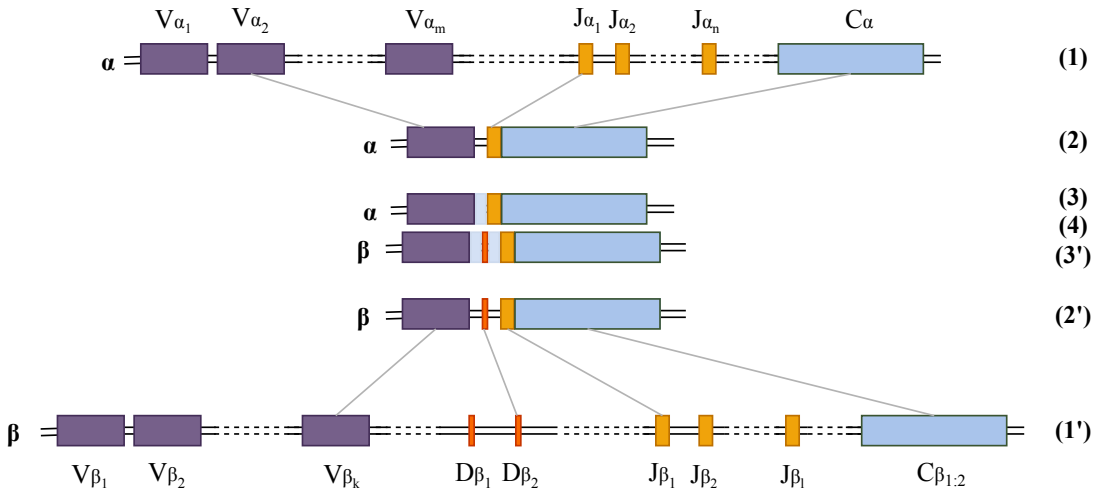


Fig. 4.2 TCR diversity. TCR α and β chains are generated by gene recombination. Multiple stages of diversity lead to a vast TCR repertoire. The germline V, D, J, and C alleles are randomly selected for recombination of the α and β chains (1,2 and 1′,2′). Random nucleotides are inserted and deleted at the junction sites, shown here in light gray (3,3′). α and β chains couple to create an additional layer of diversity (4). I illustrated this process based on the details provided in [117, Chapter 9, p. 204-212]. In this figure I have omitted the $\gamma\delta$ gene segments to simplify the process. See Figure 9-6 in [117, Chapter 9, p. 206] for a precise depiction.

T cells expressing identical receptors are said to be of the same *clone* and their TCRs of the same *clonotype* [137]. As mentioned before, since equality is determined via the CDR3 region, TCRs with the same productive CDR3 regions in their α and β chains are considered to be of the same clonotype. With some exceptions [239], it is highly unlikely that different cells with no common predecessor express the same CDR3, hence cells from the same clone are considered to have proliferated from the same T cell. All TCR clonotypes expressed by all the T cells of a host make up its TCR *repertoire*. The CDR3 amino-acid sequences can be considered as the unique index of the TCR repertoire.

In theory, there can exist between 10^{15} and 10^{20} unique TCR chains [131]. However, considering there are about 3.72 $\times 10^{13}$ non-bacterial cells in the human body [20], and that clonotypes tend to cluster in large groups due to thymic selection and antigen specificity, it is impossible that the full repertoire is realized in a single individual.

The TCR repertoire diversity varies throughout life, and in health and disease. For instance, by using high-throughput sequencing of TCRβ samples taken from patients with multiple sclerosis (MS) and control subjects, [147] demonstrated that the TCRβ sequences of Epstein-Barr virus (EBV) reactive $CD8^+$ T cells were enriched in MS patients only, while EBV-reactive $CD4^+$ T cells were enriched in both MS patients and control subjects. In cancer, $CD8^+$ cytotoxic T are believed to play a key role in killing cancer

cells [154] via perforins and granzymes, potent toxins secreted by cytotoxic lymphocytes. [87] assessed the phenotypic traits expressed by $CD8^+$ tumor-infiltrating lymphocytes (TILs) and the TCRβ chain clonotypic immunoprofiling in melanoma patients to identify the diverse repertoire of tumor-reactive cells. Studies have shown that profiling immune repertoires of TILs can provide a powerful platform to study anti-tumor T cell responses [118], [218], [14].

4.2.4 Tumor-infiltrating lymphocytes

The lymphatic system plays a vital role in immunity. A complex network of lymph vessels transports antigens and APCs to lymph nodes to generate adaptive immune responses [172]. In order to interact with an antigen and achieve activation, naive T cells circulate through secondary lymphoid organs which harbor a concentrated population of antigens. Activated effector T cells then travel to any site necessary in order to eliminate the specific antigen for which they were activated [126, Chapter 6, p. 194].

T cells are thought to recognize and eliminate cancer cells through the recognition of cancer-associated antigens, namely *tumor antigens*. Tumor antigens are expressed on the surface of tumor cells. In the broad sense, they can be *tumor-specific antigens* (TSA) - restricted only to tumor cells - or *tumor-associated antigens* (TAA) - present on both tumor and normal cells, but expressed at abnormal concentrations in a tumor cell [152, Chapter 16, p. 423-455]. The T cell response can differ based on the type of the tumor antigen: TAAs may trigger a weak immune response as T cells are programmed to be tolerant to self-antigens. Conversely, TSAs deemed foreign by the immune system may trigger a stronger immune response [263]. Cells involved in anti-tumor immunosurveillance such as T cells may recognize the tumor antigens resulting in a possible intratumoral immune response [117, Chapter 22, p. 507-512].

Tumor infiltration of lymphocytes is achieved via *selective trafficking* [223]. During activation in the secondary lymphoid organs, T cells improve their adhesive ability via the expression of specific *homing receptors*[1] which target ligands in the tumor tissue, allowing them to travel to and infiltrate specific tumors. Also, in recent studies, it has been shown that naive T lymphocytes can enter tumor tissue containing *tertiary lymphoid structures* (TLS) through lymph node (LN)-like blood vessels in tumors [63]. This infiltration is dependent on expression of L-selectin in lymphocytes and *CCR7* expression in lymphoid tissue and mediated by the LN-like blood vessels that express *PNAd* and *CCL21*. The TME might be an essential site of tumor-specific T cell clonal expansion [234]. The

[1]Lymphocyte homing receptors are molecules that enable cell adhesion

abundance of *tumor-infiltrating lymphocytes* (TILs), especially $CD8^+$ T cells, is associated with better prognosis in most solid tumors [206], [163], [82], [264]. This association has been extensively discussed in [71]. Along with activated T cells, the presence of naive T cells which activate within the tumor also results in increased tumor control [63]. It is therefore vital to analyze the TCR repertoire of TILs and identify efficient cancer cell fighting T cell subclones.

Although the immune system can deploy a large array of TILs to protect the host from tumors, the tumors themselves have elegant methods to create a microenvironment that obstructs immune responses. For instance, tumor cells can lose TAA/TSA expression, secrete immunosuppressive cytokines, generate a vasculature that represses T cell influx, or they can exhaust and inactivate TILs [152, Chapter 16, p. 423-455]. Hence, besides identifying efficacious tumor-reactive TILs, it is just as crucial to study how the tumor may impair them, and how they can be reinvigorated.

4.2.5 T cell impairment

When the metabolism or function of antigen-specific CD8 T cells is diminished, they can differentiate into a *hyporesponsive* state. Hyporesponsive T cells are usually referred to as *exhausted* in chronic infections [67], [249], and *dysfunctional* or again exhausted in the context of cancer [204], [235].

[262] first described exhaustion with virus-specific $CD8^+$ T cell responses during chronic lymphocytic choriomeningitis virus (LCMV) infection of mice. Responding $CD8^+$ T cells lost their effector functions during chronic infection in a hierarchical manner becoming unresponsive to the virus. Since then, various studies have investigated the phenotypic and functional traits of exhausted T cells in chronic viral infections showing they overexpressed various inhibitory receptors, including PD-1, TIM-3, LAG-3, CTLA-4, and 2B4 [124], [249], failed to secrete effector cytokines such as Interleukin-2 (IL-2), a cytokine with immune-modulating and anti-neoplastic effects, and expressed certain transcription factors such as Blimp-1 [217] and Eomesodermin (Eomes) [173].

In the context of cancer, although the accumulation of antigen-specific $CD8^+$ T cells in tumors is a positive prognostic marker, these immune cells deteriorate and become unresponsive to cancer within the TME due to a wide range of tumor-regulated immunosuppressive mechanisms [6], [15]. $CD8^+$ TILs that were once killing cancer cells differentiate into a dysfunctional state and, losing their effector functions, fail to perform as cytotoxic killer cells. Such dysfunctional T cells are a distinct lineage of differentiated T cells [235]. Similar to exhausted $CD8^+$ T cells in chronic infections, dysfunctional $CD8^+$ TILs show upregulation of inhibitory receptors such as PD-1 [6], [15] which has become

a major focus in cancer immunity along with CTLA-4 [210], have their effector cytokine productions corrupted [6], and display particular gene expression profiles. [181] provides an up-to-date review of exhaustive and dysfunctional T cell state studies in chronic viral infections and cancer and lists the distinguishing molecular signatures determined so far.

4.2.6 Immune checkpoint inhibitors in cancer therapy

PD-1 (also known as PDCD1), the programmed cell death protein 1, shows a sustained expression in exhausted T cells in chronic viral infections [240], [53], and has become a significant marker of T cell dysfunction in cancer [6]. PD-L1 is the host cell surface receptor which interacts with the inhibitory receptor PD-1 on T cells. Cancer cells express this ligand in order to overwhelmingly suppress the T cells through the PD-1/PD-L1 axis [228].

In checkpoint blockade therapy, a monoclonal antibody, eg. anti-PD-1 or anti-PD-L1, binds to the relevant membrane protein, PD-1 or PD-L1, to block the natural interaction of PD-L1 with the PD-1 receptor. This blockade removes cancer imposed suppression on the T cells resulting in the enhanced killing of the tumor cells. Anti-PD-1 checkpoint blockade therapy has been especially effective in malignant melanomas [191]. CTLA-4, cytotoxic T-lymphocyte antigen 4, is another "brake" on T cells, and anti-CTLA-4 checkpoint blockade therapy, though less effective than anti-PD-1 [252], has been shown to extend the life expectancy of advanced melanoma patients. Some studies suggest that PD-1 and CTLA-4 play complementary roles, and a combination of both therapies yield better outcome [253].

While checkpoint blockade therapy provides significant improvement in patient outcome, it only helps a minority of patients and is not effective long-term [140]. It is thus vital to associate the mechanisms that regulate or hinder permanent positive response with checkpoint therapy, in order to transform immunotherapy into a broadly applicable cancer treatment.

Furthermore, studies analyzing the TME have observed that response to immune checkpoint blockade favors pre-existing T cell infiltration in tumors, which they call "hot tumors", as opposed to "cold tumors" that *lack* T cell infiltration [26]. One common interpretation of this phenomenon is that hot tumors have already been infiltrated by substantial quantities of T cells, resulting in an elevated level of ongoing intratumoral immune response. Although the existing T cells have not completely eradicated the cancer cells, their level of activity can be increased by ICIs. Indeed, ICI therapy is thought to lift the restraints burdening dysfunctional T cells, thus allowing them to resume their effector functions.

Recent studies, while characterizing T cell states, have looked at biological markers that can predict clinical response to checkpoint therapy. For instance, [235] analyzed a total of 4645 malignant and tumor cells from 19 human melanoma samples. The authors examined $CD8^+$ T cells to identify exhaustion programs. The number of $CD8^+$ T cells extracted from the five selected tumors (where the selection was based on high $CD8^+$ T cell count) changed between 68 and 214. Their core exhaustion signature yielded 28 constantly up-regulated genes, including CXCL13, TNFRSF9, TIGIT, and CD27. Another study determined that TCF7 (TCF1 in mice) expression can predict positive clinical response [197]. TCF7, IL7R, and FOXP1 were shown to be associated with responder $CD8^+$ T cells, while previously discovered exhaustion markers such as TIM3 [199] and ENTPD1 were associated with non-responder $CD8^+$ T cells and dysfunction state. In addition, via TCR analysis, they showed that T cells can transition from one state to the other. This was later confirmed by [140] where again combining scRNA-seq and TCR-seq they showed that the dysfunctional $CD8^+$ T cell state rather than being a discrete pool constituted a gradient spectrum going from transitional state to early dysfunction, and highly dysfunctional. This seems to be a unique feature of melanoma infiltrating $CD8^+$ T cells. [140] too characterize dysfunction by LAG3 and PDCD1. However, they determine ETV1 and AKAPF as additional markers of dysfunctional $CD8^+$ T cells. Additionally, based on the coupled analysis of scRNA-seq and TCR-seq, they observed that dysfunctional T cells displayed the highest level of clonality.

As confirmed by [140], cells defined as dysfunctional could be in the initial state of dysfunctionality, which is thought to be reversible. For example, several recent studies have looked into the role that a specific nuclear factor, thymocyte selection-associated HMG box protein, TOX, plays in the differentiation of exhausted $CD8^+$ T cells in both chronic infection and tumors [257], [208], [211], [114]. TOX is highly expressed in melanoma, as well as other solid tumors. Their findings suggest that regulation of TOX could potentially reverse T cell exhaustion in human cancers.

4.3 Integrated approach to expose reactive-TIL dysfunction at single cell level

It is clear that there is a need to study the TME, focusing on the interactions of intratumoral malignant and non-malignant cells. If we can fully resolve the mechanisms behind tumor-infiltrating T cell dysfunction, we can improve the benefits gained from checkpoint blockade therapies. The TME is highly heterogeneous, as it comprises different populations of cancer cells, and a variety of resident and infiltrating host cells. This heterogeneity also characterises tumor-infiltrating T cells: the (i) type of T cells (e.g., $CD8^+$, $CD4^+$, Tregs) which can be (ii) at various states of (dys-)function, and (iii) target different antigens based on its antigen-specific TCR, lead to many T cell subpopulations. Resolving this heterogeneity requires an unbiased, single-cell level analysis of the TME, a cellular analysis which does not rely on pre-determined markers to allow for the discovery of new components. Additionally, the combined analysis of T cell transcriptional profiles together with their TCR clonotypes at single-cell level can uncover the links and transitions between transcriptional states. As previously shown [197], [140], analyzing shared TCRs between distinct cell states can shed light on the differentiation processes of TILs by unveiling which states they can occupy.

However, combining single-cell gene expression profiles with TCR information at the cellular level alone will not suffice. We need novel approaches which allow the coupling of cell state with the inherent potential of a T cell's ability to recognize specific tumor antigens [233]. Even though by linking single-cell transcriptional profiles and TCR sequences we can potentially determine tumor reactivity via clonality measures, the result may be confounded by different factors. For instance, the proliferation of the tumor-reactive T cells could be hindered by exhaustion or other immuno-suppressive mechanisms. More importantly, recent work provides evidence that intratumoral T cells may be bystanders with no tumor reactivity [219]. In order to provide opportunities to reinstate effective and long-lasting intratumoral T cell response, we need to determine, unambiguously, which specific T cells are responsible for tumor antigen recognition.

4.3.1 Antigen-Receptor Signalling Reporter (AgRSR) mouse

For this work, I collaborated with Dr. James Thaventhiran who has developed a novel transgenic mouse model, namely the Antigen-Receptor Signalling Reporter (*AgRSR*), which allows in vivo clonal tracking of activated B and T cells [93]. In this model, with the administration of a specific chemical trigger, *tamoxifen*, activated cells can be

fluorescently marked upon AgR signalling. AgRSR mice have been genetically modified by injecting a bacterial artificial chromosome (BAC) transgene which has been modified by inserting Katushka, a red fluorescent protein derived from the sea anemone Entacmaea quadricolor [215], and an E2A linked tamoxifen-regulated Cre-ERT2 recombinase, into the translation initiation codon of the Nur77 gene. The Nur77 gene is expressed upon AgR signaling in both B and T cells [12]. In this model, the activation signal can be detected with red fluorescence signals, via the expression of Katushka. This red fluorescence signal decays as the AgR trigger is removed, i.e., Katushka is not permanent. Therefore, in order to permanently mark activated B and T cells this mouse was crossed to the ROSA-Lox-Stop-Lox-EYFP reporter strain establishing a new strain in which, upon AgR activation, the administration of tamoxifen induces Cre-ERT2 recombinase mediated excision of the Stop cassette, which prevents the expression of the yellow fluorescent protein (EYFP). As a result, the administration of tamoxifen leads to the expression of EYFP. Since EYFP expression is determined by DNA-recombination, it is inheritable, so all clonal progeny of the marked cells remain EYFP$^+$ [James Thaventhiran, personal communication, September 10, 2018].

Model validation

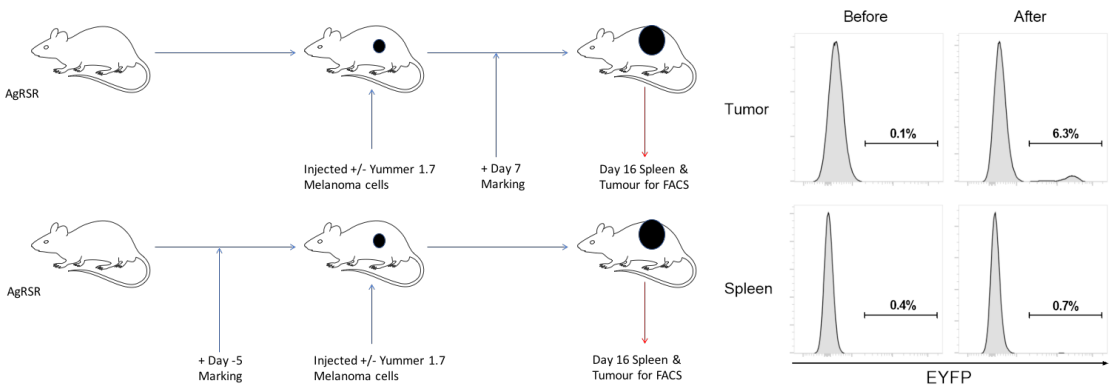


Fig. 4.3 Assessment of antigen specificity of marking. EYFP expression in spleen and tumor CD8$^+$ T cells, marked before and after melanoma tumor injection. Plots show that the marking is specific to cells recognizing tumor antigens. Cartoon and plots courtesy of Mr. Ty So.

Experiments carried out by Dr. James Thaventhiran and Mr. Ty So demonstrated that EYFP expression in T (and B) cell clones is dependent on tamoxifen exposure. Furthermore, in T cells receiving inflammatory or cytokine signaling, expression of EYFP is dependent on MHC molecules, confirming the TCR-specificity of the marking (Fig-

ure 4.3). This model can help track the clonal $CD8^+$ T cell response to tumor antigens. Furthermore, as this model kinetically timestamps TCR activation via tamoxifen administration, it can fate-map TCR activated $CD8^+$ T cell clones within an experimentally defined time frame [93].

4.3.2 Immune profiling of $EYFP^+$ cells

Current single-cell transcriptomic techniques provide limited insights on the causes of clonal T cell dysfunction as they do not allow us to determine how particular dysfunctional states and inherent tumor reactivity are linked. By immune profiling the $EYFP^+$ cells of the AgRSR mouse model, we aimed to overcome this limitation and provide a fully unambiguous analysis of T cell dysfunction within the TME.

4.3.3 Cell harvest and sequencing

Mr. Ty So challenged three mice with subcutaneous melanoma tumors at day 0 (Figure 4.4). The mice were later administered different dosages of tamoxifen at day 7. The first two mice were given half the amount the third mouse received. At day 15, the first mouse was sacrificed by cervical dislocation to harvest and digest the tumor and the spleen. Following FACS surface staining, tumor and spleen cells were sorted for $CD45^+CD11B^-EYFP^+$ [James Thaventhiran and Ty So, personal communication, September 10, 2018]. During FACS, in order to obtain $EYFP^+$ lymphocytes from mouse tumor and spleens, non-target cells, along with doublets, dead cells, non-lymphocytes, as well as $EYFP^-$ cells were removed (Figure 4.5).

As described in detail in Chapter 2, using 10x Genomics' Chromium Single Cell Immune Profiling Solution, libraries were generated for both tumor and spleen cells to obtain gene expression profiles, and full-length, paired, TCR repertoires from the same

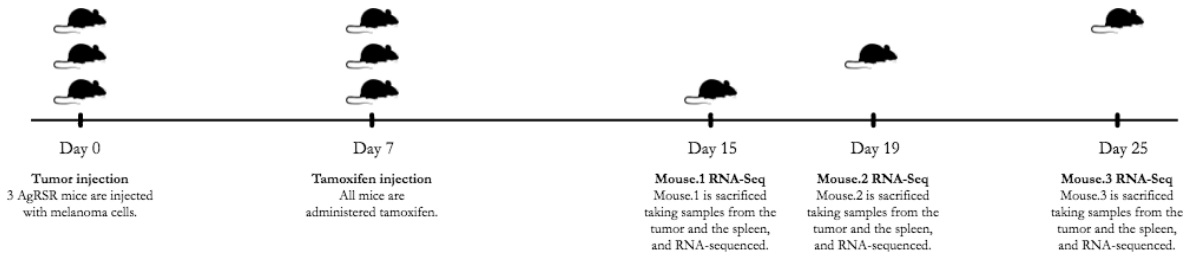


Fig. 4.4 Experimental timeline. All three mice were challenged with subcutaneous melanoma-tumors at day 0, later administered tamoxifen at day 7. The mice were sacrificed and their tumor and spleen cells were harvested and processed with the Single Cell Immune Profiling protocol at days 15, 19, and 25.

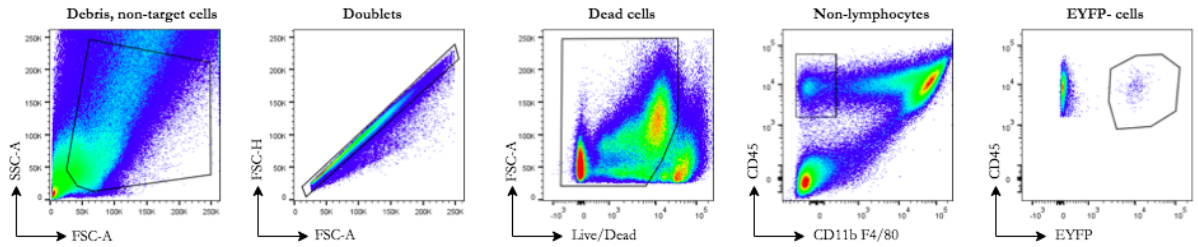


Fig. 4.5 Cell filtering. Regions with black borders are the selected cells. Forward Scatter Area (FSC-A) by Side Scatter Area (SSC-A) gate is used to remove debris. FSC-Height (FSC-H) by FSC-Area (FSC-A) gate is used to detect single cells which fall on the diagonal. Lymphocytes are $CD45^{+}CD11B^{-}$. Plots courtesy of Mr. Ty So.

input sample. Sequencer-ready scRNA-seq libraries with unique sample indices were sequenced on the Illumina HiSeq 4000 platform. The next two mice went through the same process later at day 19 and day 25. As mentioned previously, mice sacrificed on days 15 and 19 were given half the tamoxifen dosage given to the mouse sacrificed on day 25.

4.3.4 Chromium scRNA-seq output processing pipeline

Raw Illumina reads were converted into FASTQ files and processed using CellRanger v3.0.2. I used the default `cellranger count` arguments to align, filter, and count unique molecular identifiers (UMIs) to generate a (gene x cell) count matrix. I mapped the sequences to the mm10 genome and counted with the GRCm38 annotation. When reconstructing TCRs via `cellranger vdj`, as described in Section 2.4.2, I generated a V(D)J reference from the IMGT database. We obtained a total of six GEX profiles, and six TCR repertoires. Throughout this chapter, each mouse is labelled based on day of sacrifice; *mouse_d15*, *mouse_d19*, and *mouse_d25*.

cellranger count summary metrics

Table 4.1 and 4.2 list the summary metrics of the GEX barcoding and sequencing process on the cells obtained from the tumors and the spleens of each mouse, respectively. In both tumor and spleen samples, the total number of read pairs that were assigned to *mouse_d25* is significantly less compared to the other two samples. Nevertheless, they are still above the acceptable threshold of minimum 20,000 read pairs/cell [1]. It should be noted that *mouse_d25* samples were the first to be sequenced. The two other mouse samples were sequenced deeper to increase the read count per cell. However note that the sequencing saturation, i.e., the measure of the fraction of the total number of different

	mouse_d15	mouse_d19	mouse_d25
Estimated Number of Cells	7,633	5,343	8,351
Mean Reads per Cell	39,707	61,371	24,978
Median Genes per Cell	2,571	2,415	1,342
Number of Reads	303,085,056	327,908,254	208,593,470
Valid Barcodes	94.1%	94.9%	95.2%
Sequencing Saturation	61.8%	77.6%	72.6%
Reads Mapped Confidently to Genome	94.3%	95.2%	95.8%
Reads Mapped Confidently to Transcriptome	81.5%	83.1%	82.8%
Fraction Reads in Cells	95.6%	95.7%	95.0%
Total Genes Detected	17,069	16,825	18,149
Median UMI Counts per Cell	9,115	8,202	2,886

Table 4.1 GEX analysis metrics of the tumor samples

	mouse_d15	mouse_d19	mouse_d25
Estimated Number of Cells	3,992	3,082	4,918
Mean Reads per Cell	75,212	104,996	42,152
Median Genes per Cell	2,083	1,799	660
Number of Reads	300,249,980	323,599,839	207,304,501
Valid Barcodes	93.5%	94.8%	93.5%
Sequencing Saturation	87.4%	92.9%	83.8%
Reads Mapped Confidently to Genome	94.0%	95.3%	95.7%
Reads Mapped Confidently to Transcriptome	79.3%	81.6%	79.6%
Fraction Reads in Cells	97.1%	95.8%	36.9%
Total Genes Detected	15,621	15,046	14,896
Median UMI Counts per Cell	5,915	4,689	1,242

Table 4.2 GEX analysis metrics of the spleen samples

transcripts in the final library that was sequenced, of *mouse_d25* is similar to that of the other mice, again suggesting that the summary metrics are not a source of concern. As the sequencing saturation increases, additional sequencing does not result in many new unique UMIs for the library. In terms of the number of cells detected for each sample; as of version 3.0, Cell Ranger has implemented a cell-calling algorithm, based on [150], which can identify low RNA content cells even when mixed with high RNA content cells. The implemented algorithm is two-step:

- **Step 1 - Cutoff based on total UMI counts:** The Cell Ranger cell calling algorithm prior to version 3 is used to identify high RNA content cells using the cutoff method described in Section 2.4.2.
- **Step 2 - Captures low RNA content cells whose UMI counts may resemble empty GEMs:** Based on [150], the RNA profile of each barcode not called as a cell in step 1 due to its relatively low RNA content is compared to a background model which is created by modeling the RNA profile of a set of selected barcodes with low UMI counts that most likely represent empty GEMs. Barcodes whose RNA profile deviate from the background model are kept as cells.

In the TME, large tumor cells might be mixed with relatively smaller TILs. As the cells in our samples are tumor-reactive TILs, we expected to see cells with relatively lower UMI counts. This is most obvious in *mouse_d25*. The low median UMI counts per cell along with lower than expected median genes per cell could be indicative of a large population of TILs. This is later explored and clarified in Section 4.4.1. The *mouse_d25* spleen sample looks to be troublesome in terms of the fraction of reads in cells. This is explained by high ambient RNA in a sample. Ambient RNA comes from lysed/dead cells. Although the reads are aligned confidently, they are not associated with a valid cell-containing GEM. I have run `cellranger count` with the `--force-cells` option to increase the expected number of cells, however, did not achieve a significant increase in the cell count, suggesting that these cells may have been degraded. *mouse_d25* spleen sample was the first to be processed in the lab, and a low cell count was observed, which was speculated to be due to the delay in processing the sample. For the other spleen samples, they have reduced the time it takes to harvest the cells and RNA-sequence them which has increased the recovery of splenic cells, as observed both in the lab and post immune profiling analysis (Table 4.2). For all samples, the percentage of cells that mapped back to the mm10 genome is high, showing that there was no external contamination and that the detected cells were efficiently lysed.

cellranger vdj summary metrics

Tables 4.3 and 4.4 contain metrics about the barcoding and sequencing of the VDJ enrichment process. Looking at the metrics, other than the low fraction of reads in cells in the *mouse_d25* spleen sample, there is no noticeable issue.

Merging barcode information from cellranger count and cellranger vdj

Within the Chromium Single Cell Immune Profiling Solution workflow, GEMs are generated after cell suspension is loaded on to the Chromium controller, followed by reverse transcribing full-length cDNA in single-cell GEMs. The cDNA is then amplified and divided into two or three aliquots depending on whether VDJ is enriched for both BCR and TCR, or only for either type. Lastly, the preparation of libraries for 5′ gene expression and enrichment of VDJs followed by library preparation takes place.

Cell Ranger does not provide a consistent cell calling algorithm between GEX profiling and TCR reconstruction. There are barcodes accepted as cells via `cellranger count` but not `cellranger vdj` and vice versa. There were productive, full-length chains called with confidence by `cellranger vdj` that were not considered to be valid as the associated barcode was not detected as a cell with `cellranger vdj`, but was kept by

	mouse_d15	mouse_d19	mouse_d25
Estimated Number of Cells	6,881	4,734	5,999
Mean Read Pairs per Cell	2,462	7,946	4,763
Number of Cells With Productive V-J Spanning Pair	6,076	4,274	5,046
Number of Read Pairs	16,942,983	37,620,817	28,577,686
Valid Barcodes	96.2%	96.4%	95.8%
Reads Mapped to Any V(D)J Gene	88.9%	88.1%	86.8%
Reads Mapped to TRA	35.9%	38.3%	33.2%
Reads Mapped to TRB	36.2%	26.4%	27.1%
Cell Count Confidence	97.6%	97.1%	95.1%
Mean Used Read Pairs per Cell	2,142	6,936	4,042
Fraction Reads in Cells	96.3%	96.4%	93.9%
Median TRA UMIs per Cell	14	12	5
Median TRB UMIs per Cell	30	24	6
Cells With Productive V-J Spanning Pair	88.3%	90.3%	84.1%
Cells With Productive V-J Spanning (TRA, TRB) Pair	88.3%	90.3%	84.1%
Cells With TRA Contig	98.4%	98.1%	97.1%
Cells With TRB Contig	98.9%	99.2%	98.0%
Cells With CDR3-annotated TRA Contig	95.8%	96.3%	94.2%
Cells With CDR3-annotated TRB Contig	92.8%	92.8%	89.2%
Cells With V-J Spanning TRA Contig	96.8%	97.1%	95.8%
Cells With V-J Spanning TRB Contig	93.4%	93.2%	89.9%
Cells With Productive TRA Contig	95.1%	95.9%	93.6%
Cells With Productive TRB Contig	91.9%	92.3%	88.9%

Table 4.3 VDJ analysis metrics of all mouse samples taken from the tumor

	mouse_d15	mouse_d19	mouse_d25
Estimated Number of Cells	3,748	2,901	1,851
Mean Read Pairs per Cell	4,283	15,079	15,443
Number of Cells With Productive V-J Spanning Pair	3,102	2,407	1,304
Number of Read Pairs	16,055,573	43,746,295	28,586,455
Valid Barcodes	96.9%	97.3%	94.5%
Reads Mapped to Any V(D)J Gene	88.7%	89.0%	84.8%
Reads Mapped to TRA	27.6%	27.6%	25.4%
Reads Mapped to TRB	40.1%	40.1%	35.2%
Cell Count Confidence	98.9%	100.0%	48.3%
Mean Used Read Pairs per Cell	3,892	13,635	4,068
Fraction Reads in Cells	97.9%	96.5%	29.6%
Median TRA UMIs per Cell	6	5	2
Median TRB UMIs per Cell	17	13	5
Cells With Productive V-J Spanning Pair	82.8%	83.0%	70.4%
Cells With Productive V-J Spanning (TRA, TRB) Pair	82.8%	83.0%	70.4%
Cells With TRA Contig	93.6%	93.5%	83.6%
Cells With TRB Contig	99.8%	99.8%	99.6%
Cells With CDR3-annotated TRA Contig	88.8%	89.1%	77.9%
Cells With CDR3-annotated TRB Contig	91.5%	91.9%	91.3%
Cells With V-J Spanning TRA Contig	91.1%	91.0%	81.0%
Cells With V-J Spanning TRB Contig	92.2%	92.6%	91.7%
Cells With Productive TRA Contig	87.9%	88.1%	76.4%
Cells With Productive TRB Contig	90.4%	90.7%	90.7%

Table 4.4 VDJ analysis metrics of all mouse samples taken from the spleen

`cellranger count`. I updated the associated barcode of these VDJs as a true cell if it was detected with `cellranger count`. I discarded a portion of these cells, along with others, during QC, which I explain in Section 4.4.1.

4.4 Analysis

I present the downstream analysis of all the tumor samples. In the remainder of this chapter, "tumor samples" refer to cells harvested from the tumor of an AgRSR mouse, sorted with FACS for $CD45^{+}CD11B^{-}EYFP^{+}$, and processed with the Chromium Single Cell Immune Profiling Solution.

4.4.1 Gene expression analysis of $EYFP^{+}$ cells

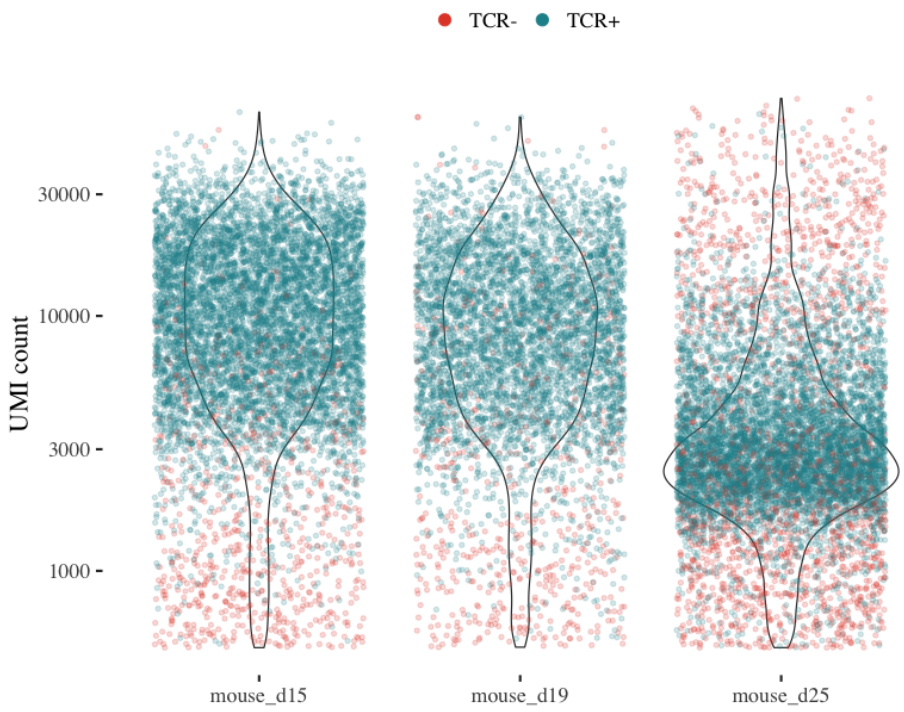
Unlike in Chapter 2, I did not call cells from empty droplets as Cell Ranger version 3.0 has implemented its own cell-calling algorithm based on [150] to detect cells more accurately than the prior versions. As mentioned before, the output of 5′ GEX analysis is a (gene x cell) UMI count matrix, henceforth referred to as the count matrix. The values in the count matrix represent the number of molecules for each gene, detected in each cell.

QC: Library sizes

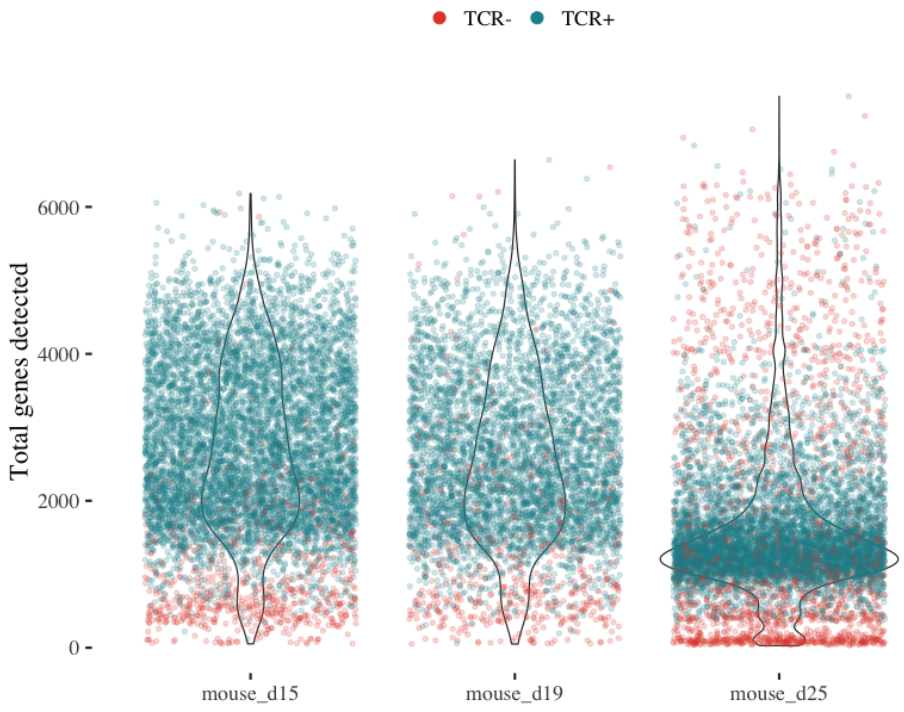
The sum of UMI counts, detected within a cell is referred to as the library size of the cell. Figure 4.6a shows the library sizes of each sample distinguished by whether it bears a TCR or not. Here, TCR^{-} cells may not be T cells, or a productive TCR may not have been detected. Furthermore, cells with small library sizes tend to be TCR^{-}, which may suggest that these are artifacts, non-T cells, or damaged or dead cells. There is also a noticeable difference in the UMI count per cell in the *mouse_d25* sample, which was discussed in Section 4.3.4. Additionally *mouse_d25* has high UMI count cells which are TCR^{-}.

QC: Unique genes detected

The number of uniquely detected genes within each cell, i.e., the count of all genes with UMI>0 within a cell, is a QC metric which can determine whether a cell is of low quality or an empty droplet (very few detected genes). High gene count may signify doublets/multiplets. Figure 4.6b shows unique gene count and TCR^{-} seem to correspond in low gene counts. These may be low-quality cells.



(a) Library size distributions.



(b) Uniquely detected gene count distributions.

Fig. 4.6 Library size (shown on log10 scale) and uniquely detected gene count distributions in the tumor samples. Each point represents a cell. Color denotes whether the cell bears a TCR or not.

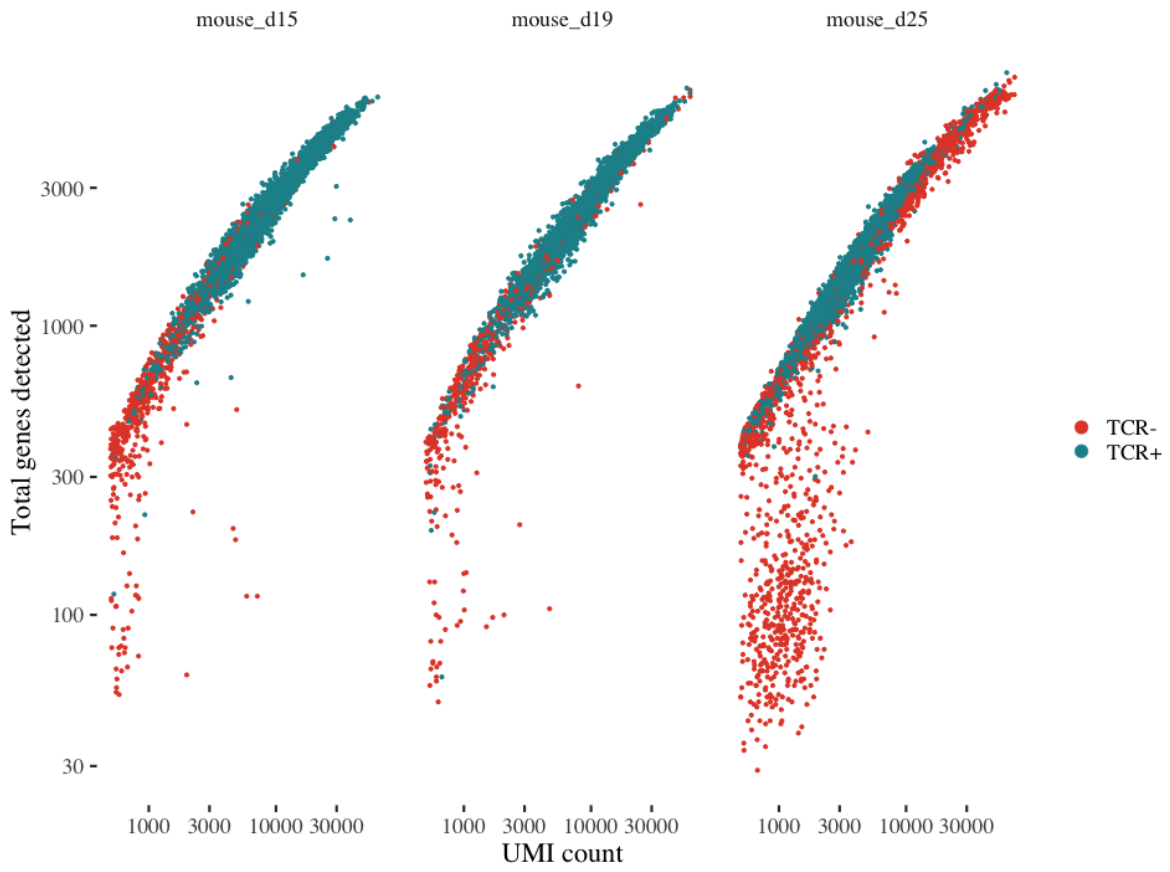


Fig. 4.7 Cell complexity of tumor samples. Each point represents a cell. Total genes detected and UMI count are shown on log10 scale.

QC: Cell complexity

Directly removing cells which have too few or too many detected genes, and large or small library sizes can remove specific cell populations. In a heterogeneous population like the TME, there may be different cell types belonging to different states, and hence expressing different numbers of genes at different rates.

Combining library sizes and uniquely detected gene counts QC metrics gives us *cell complexity*. Looking at Figure 4.7 we can see that there are TCR^- cells with small library sizes that also, in their majority, express less than 300 genes, which is especially evident in sample *mouse_d25*. These are most likely to be artifacts. However, they can also be a particular type of cell.

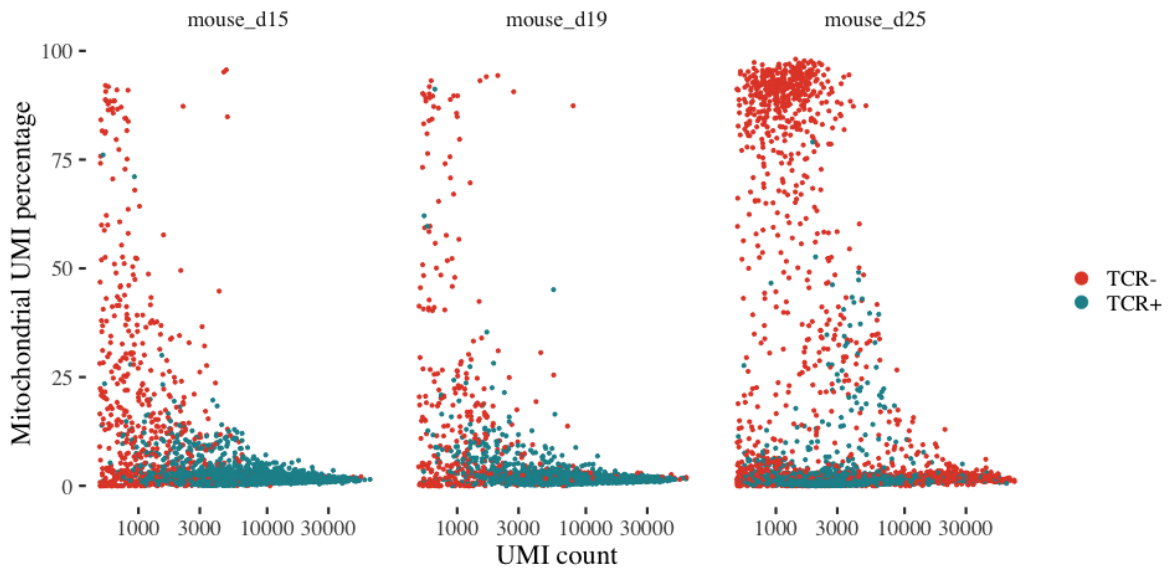


Fig. 4.8 Percentage of UMIs mapped to mitochondrial genes plotted against the UMI count (shown on log10 scale). Each point represents a cell.

QC: Mitochondrial UMI

Cells with high counts of UMIs that map to the mitochondrial genome may indicate stressed cells, or incomplete lysis [105]. Looking at Figure 4.8, we see that specific cells have very high mitochondrial content, and these cells do not bear a TCR.

QC: Mitochondrial UMI and uniquely detected genes

In Figure 4.9 we can see that genes with high mitochondrial expression have less uniquely detected genes, most of which are TCR$^-$. Based on Figures 4.7 and 4.9 I remove cells with uniquely detected gene count less than 300. However, I make sure that these cells are not enriched for a certain cell type (see Section 4.4.1, QC: Discarded cell inspection).

Next, I determine a threshold on mitochondrial gene expression to discard cells with high mitochondrial content, which may indicate dead/stressed cells or incomplete lysis.

Assuming that the mitochondrial UMI proportions are normally distributed, I use the robust Z-score method to identify outliers as described in Section 2.5.1. I consider cells to be outliers if their mitochondrial content is more than 3 MADs away from the median. Figure B.12 shows which cells are dropped/kept after this filtering.

Also observe that I increased the threshold of total genes detected to 600, 600, and 500, for samples *mouse_d15*, *mouse_d19*, and *mouse_d25* respectively. I determined this threshold in an iterative manner: visually inspecting Figure B.12 we can see that

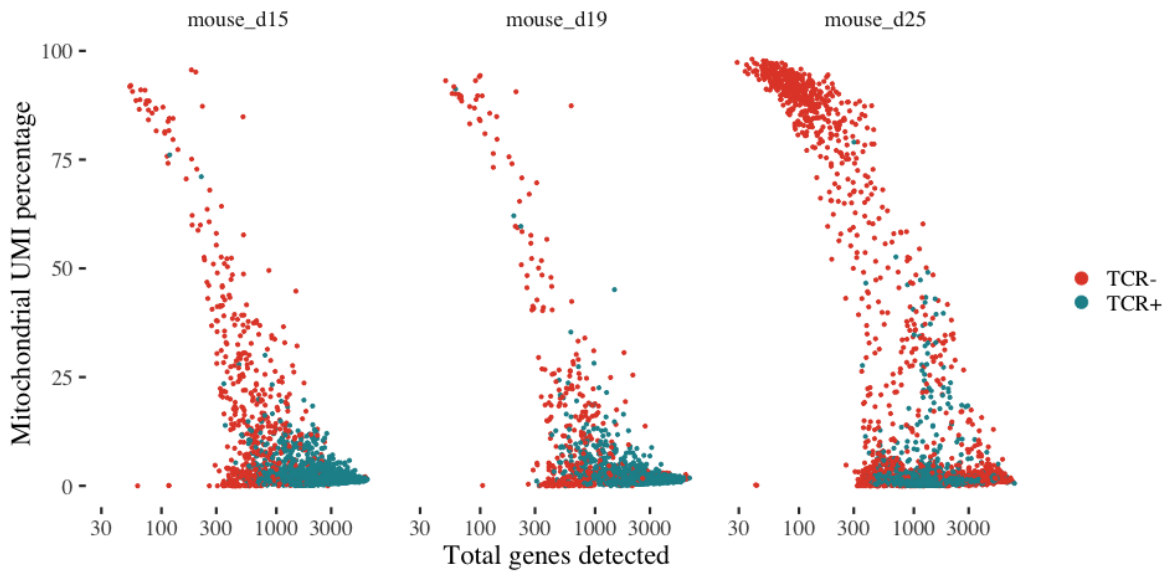


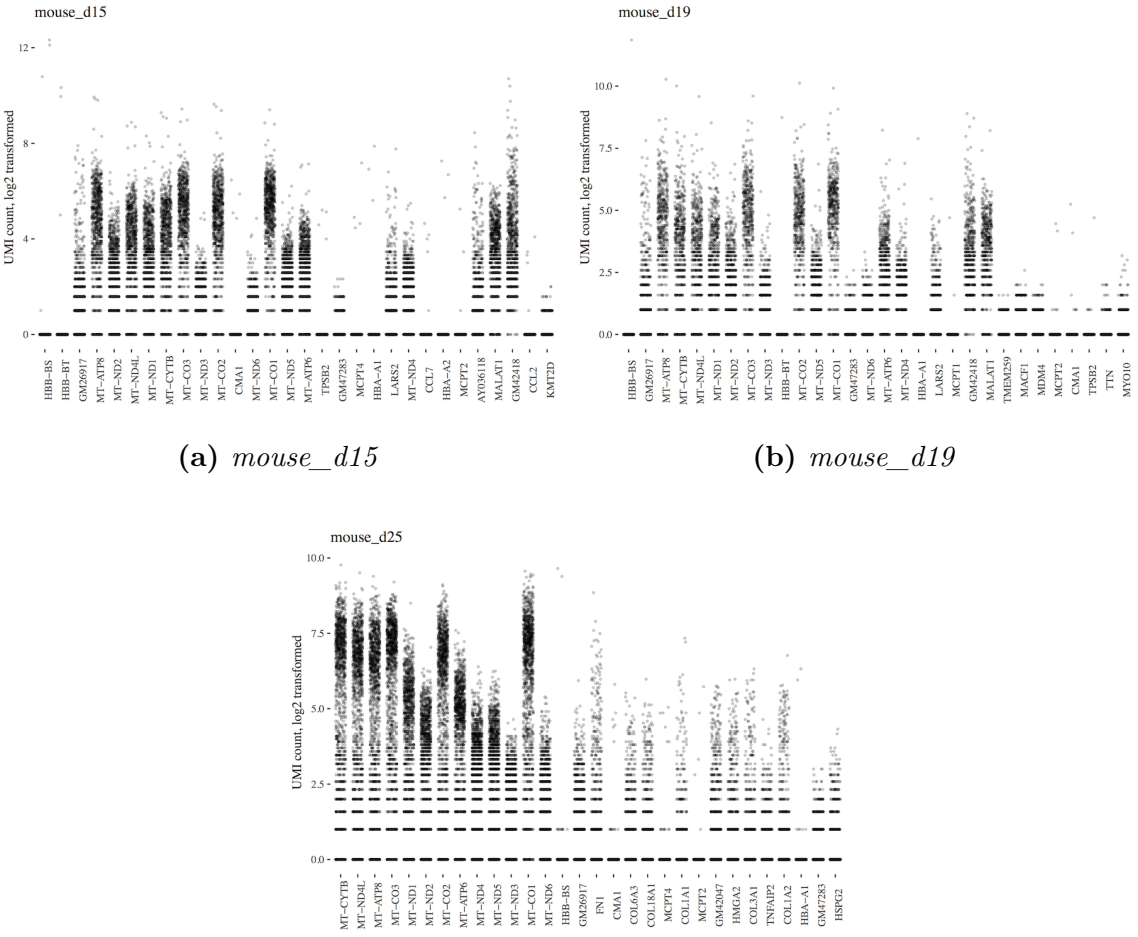
Fig. 4.9 Percentage of UMIs mapped to mitochondrial genes plotted against the number of detected genes (shown on log10 scale). Each point represents a cell.

there are dispersed TCR$^-$ cells with less than 1000 uniquely detected gene counts in samples *mouse_d15* and *mouse_d19*, and cells with less than 800 in sample *mouse_d25*. Starting at these levels, I decreased the threshold to the determined levels by examining the differences in gene expression between the dropped and kept cells, as described next.

QC: Discarded cell inspection

Figure B.13 shows the correlation between the average UMI counts of dropped and kept cells. For all samples, the populations are mostly in concordance. However, there are some distinct set of genes that are expressed higher in the dropped cell pool. For example, the genes HBB-BS, HBB-BT, HBA-A1 are hemoglobin markers. They are expressed higher in the dropped cell pool, indicating the removal of red blood cells. I examined the upregulation of genes in the discarded pool by computing log-fold changes between the two pools. The genes upregulated amongst the dropped cells are: MT-*: mitochondrial genes, HBB-BS, HBB-BT, HBA-A1 expressing hemoglobin genes. LARS2 is a mitochondrial, tRNA synthesis gene. MCPT1, MCPT2 expression indicate MAST cells, but only a few cells are expressing them (see Figure 4.10). FN-1 is a fibronectin gene, suggesting endothelial cells. AY036118 is a mouse transcription factor, and COL6A3, COL18A1, COL3A1 are collagen α chain genes which have been reported as biomarkers for various cancers.

These genes together are either not specific to a certain cell type, or they mark blood components. Hence it is reasonably safe to remove them. Removing cells with higher UMI counts (>600) caused certain naive lymphocytes to be discarded, hence I selected the determined thresholds for cell removal iteratively. After QC, each sample was reduced by approximately 10% in size.



(a) *mouse_d15* **(b)** *mouse_d19*

(c) *mouse_d25*

Fig. 4.10 Discarded cells' distinguishing gene expression.

Normalization

Our samples do not have cells with equivalent RNA abundances due to the heterogeneity of the TME. As described in Section 2.5.2, I used the deconvolution method presented in [148] to calculate size factors, after a pre-clustering step with the `scran::quickCluster` [149] function, specifying `igraph` as the clustering method to use. Figure 4.11 shows

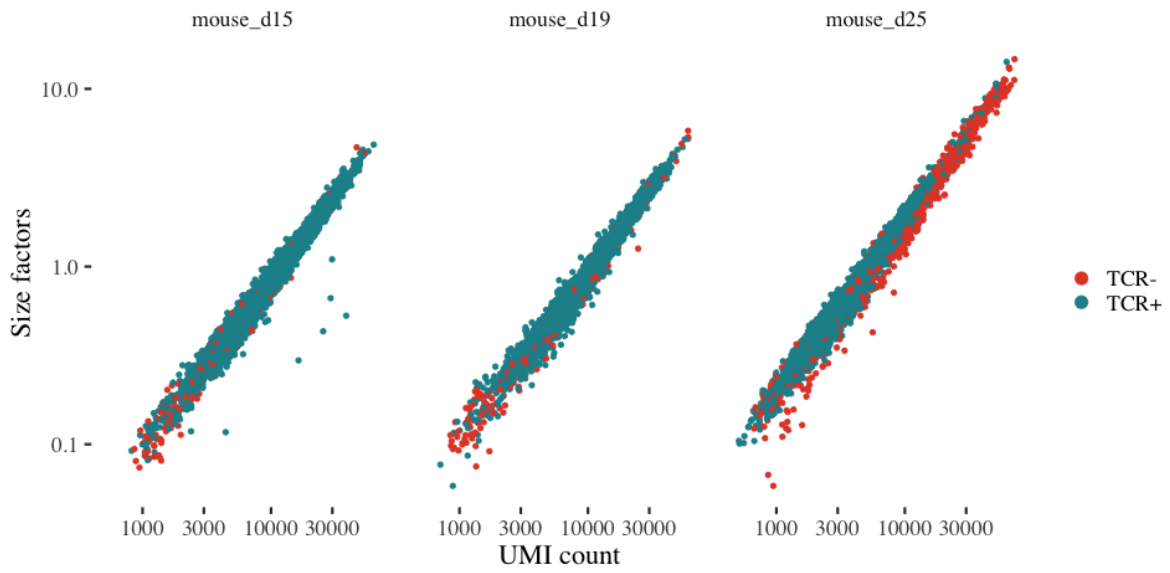


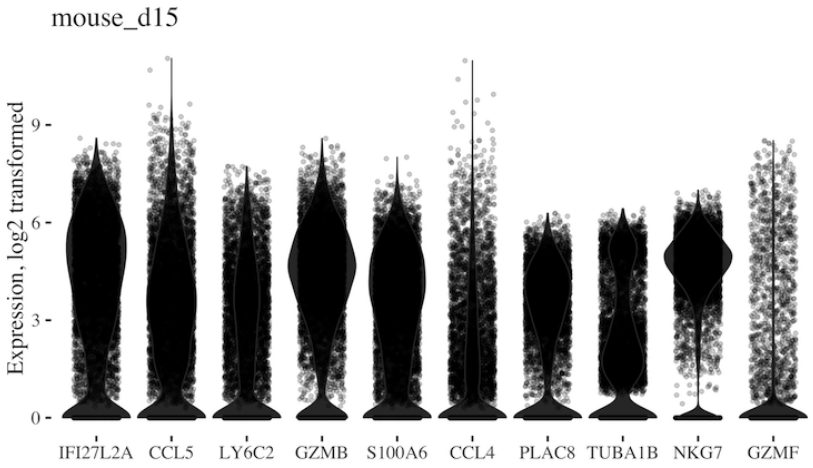
Fig. 4.11 Computed size factors plotted against the UMI count (shown on log10 scale)

the correlation of size factors against the library sizes. I normalized the samples using `scater::normalize` and log2-transformed the normalized data.

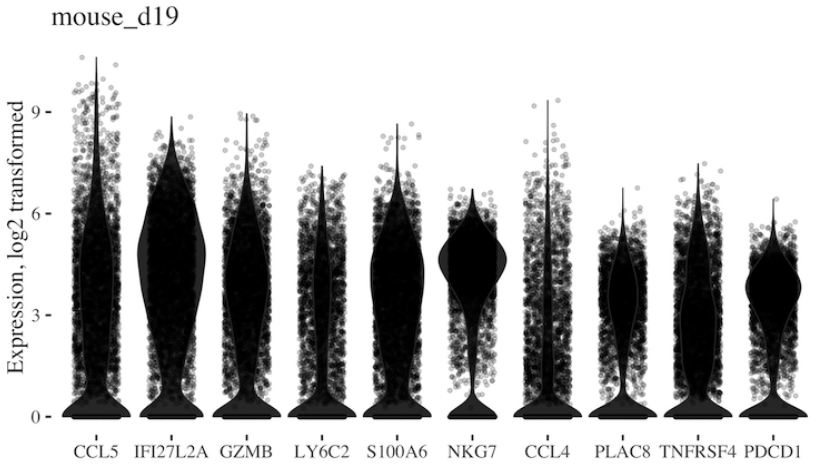
Highly variable genes

In the remainder of the analysis, I focus only on genes which show high variation in expression between cells. These highly variable genes (HVG) are highly expressed in some cells while lowly expressed in others. To determine these genes, I used `scran::decomposeVar` which decomposes the variance of each gene into biological and technical components while assuming the technical noise has a Poisson-based trend. I took the subset of genes which had a positive biological component. I also removed all TCR expressing genes, i.e., TRA* and TRB* genes as I did not want them to affect the downstream analysis directly.

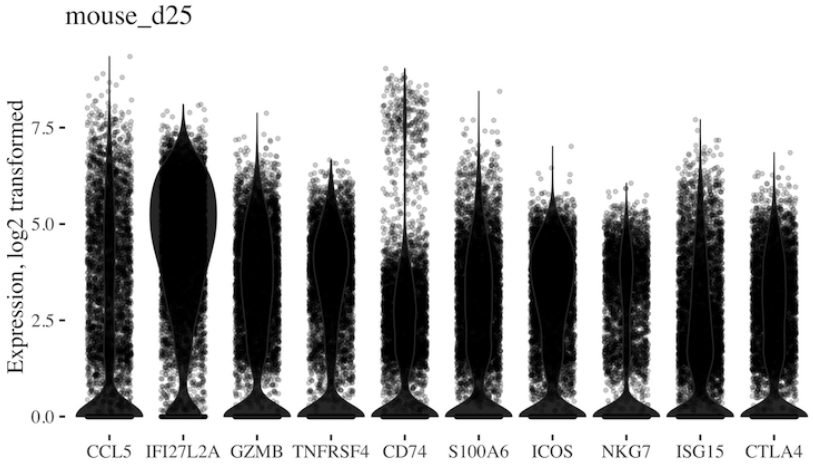
Figure 4.12 shows the distributions of normalized log2-transformed expression values of the top ten most highly variable genes of each sample. We see some promising T cell genes which mark different states and different functions.



(a) *mouse_d15* tumor HVGs



(b) *mouse_d19* tumor HVGs



(c) *mouse_d25* tumor HVGs

Fig. 4.12 Highly variable genes (HVGs) in the tumor samples.

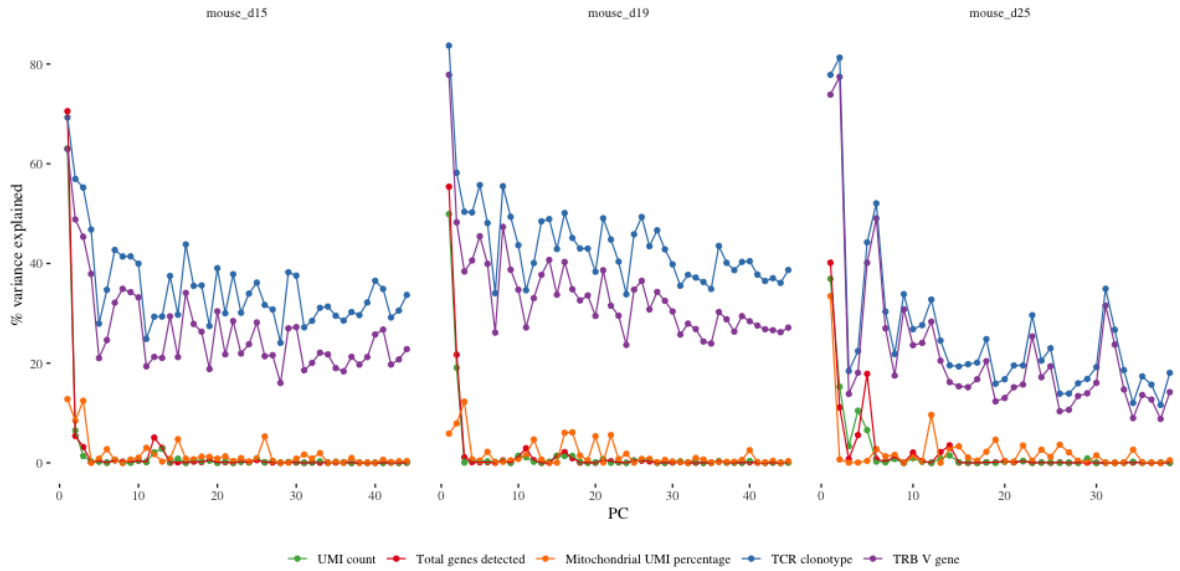


Fig. 4.13 The plots show the percentage of variance explained by the selected variables for each successive PC.

Dimension reduction using PCA

Using `denoisePCA` [149] with the assumed Poisson technical noise, I identified PCs most likely associated with actual biological phenomena, rather than technical artifacts. Here it is assumed that the underlying biology drives most of the variance and should be captured within the initial PCs, but technical noise is uncorrelated across genes and should be visible in later PCs. `denoisePCA` [149] removes these later PCs. I used only the HVGs when running `denoisePCA` but note that when selecting HVGs we did not lose any genes that are upregulated in rare populations as our filtering is very relaxed and is carried out primarily to remove TRA*/TRB* genes so that they do not drive any of the PCs directly.

In all samples, PC1 and PC2 correlate with the number of detected genes. This correlation is often observed. Also in *moused_25* PC1 correlates with the mitochondrial fraction. A correlation with library size can also be seen in the first PC. Immune cell populations are not equivalent in terms of size. For instance, T cells make cytokines, and plasma B cells make antibodies, which increase library size immensely. Here library size could be reflecting biological variation. I did not regress out the variation caused by the mitochondrial fraction or the library sizes as this could result in losing genuine biology.

The correlations of clonotypes and V genes with the PCs are interesting. It is important to recall that I omitted the TCR genes, hence this is not driven directly by the TRA*/TRB* genes. This could suggest that TCR clones express specific set of genes.

Projection

Based on my justifications in Section 2.5.6, I used the Uniform Manifold Approximation and Projection (UMAP) non-linear dimensionality reduction technique [156] to visualize the cells in two dimensions. The fact that UMAP captures global structure better than t-SNE is especially useful when visualizing cells of the TME. It allows us to see distinct clusters of $CD4^+$ and $CD8^+$ further apart from each other while $CD8^+$ cells in different stages or (dys-)functionality separate less apart. Also, as UMAP preserves distances, it tends to preserve the continuity of cell states. I performed UMAP on the prior reduced and denoised PCs. This helps reduce the computational time necessary to embed the data with UMAP.

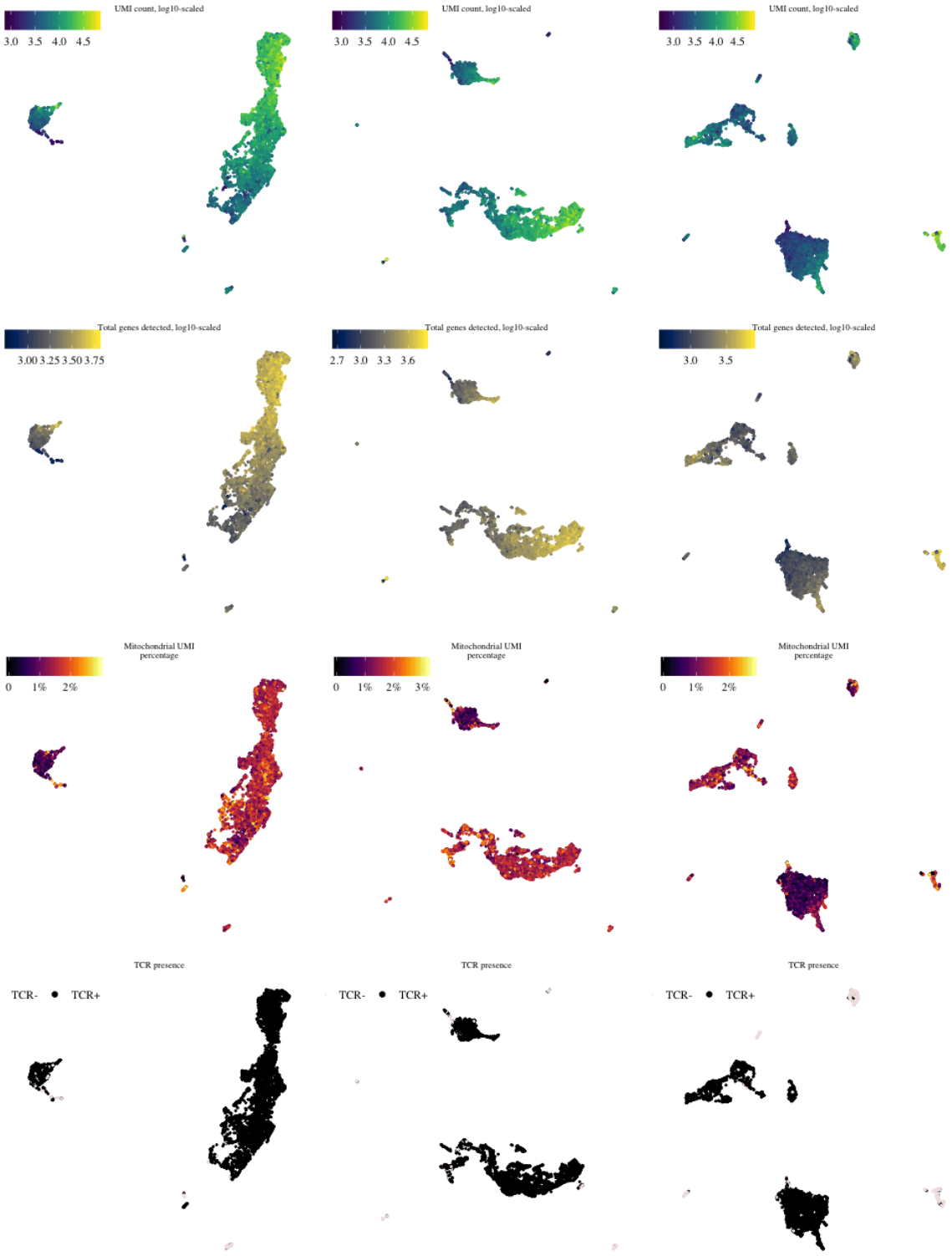
Figures 4.14 and 4.15 display the UMAPs overlaid with UMI count, uniquely detected gene counts, mitochondrial UMI percentage, TCR presence, and the normalized, log2-transformed gene expression of CD3, CD8, CD4, and FOXP3. We can see that subsets of T cells have separated further apart, as well as the cells with no TCR.

Clustering

As described in-depth in Section 2.5.5, I clustered the cells by first building a shared nearest neighbor (SNN) graph [255] using the pre-computed PCs. Then I compared different community detection algorithms (Fastgreedy [46], Walktrap [182], Louvain [78], and Infomap [194]) and Louvain showed an ideal[2] separation. Here the initial idea is to stratify the cells into high-level groups based on cell type rather than finding functional clusters. Using the Louvain algorithm, I found the community structure, i.e., clusters, which best captured the distinct Tc, Th, and Treg clusters. Figure 4.16 shows the clustering of each mouse tumor sample where the UMAP embedded data are colored by cluster membership.

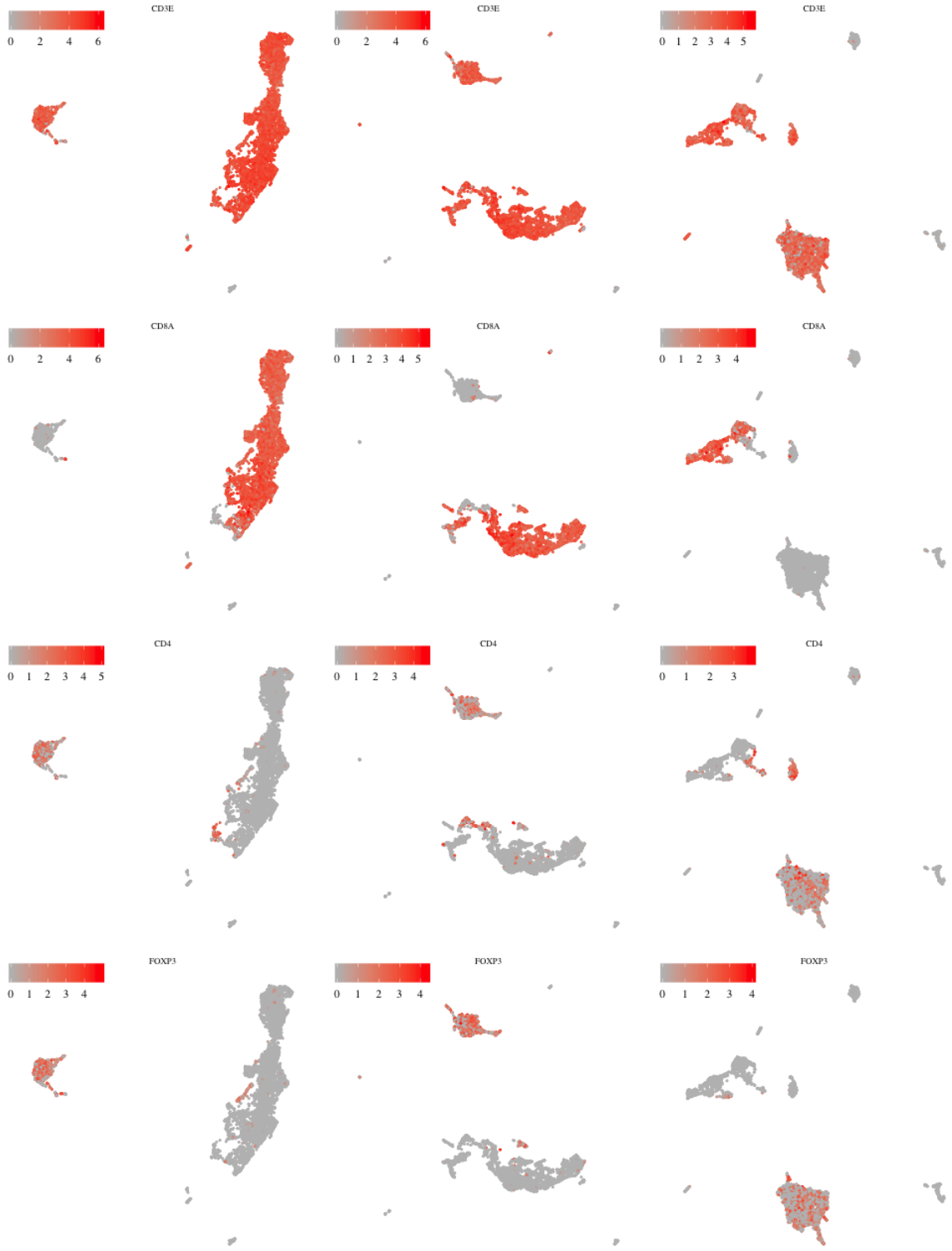
To identify clusters I examined the upregulated genes in each cluster which I found using `findMarkers` [149]. Figures B.14, B.15, and B.16 show the top marker gene expression via heatmaps. Gene expression levels are colored by the normalized, log2-transformed, and row-scaled expression of each gene in each cell. Also, in order to stratify cells into an initial high level grouping I inspected the expression of CD3, CD4, CD8, and FOXP3 genes (see Figure 4.15). Note that *mouse_d25* has a relatively small $CD8^+$ cell population.

[2] "Ideal" here means a desired separation in terms of biologically meaningful clusters. One may use the modularity score as a measure of good separation (which is of course only a technical measure and not a biologically informative one) but considering there is no right or wrong clustering, it is best to inspect the clusters in terms of gene expression.



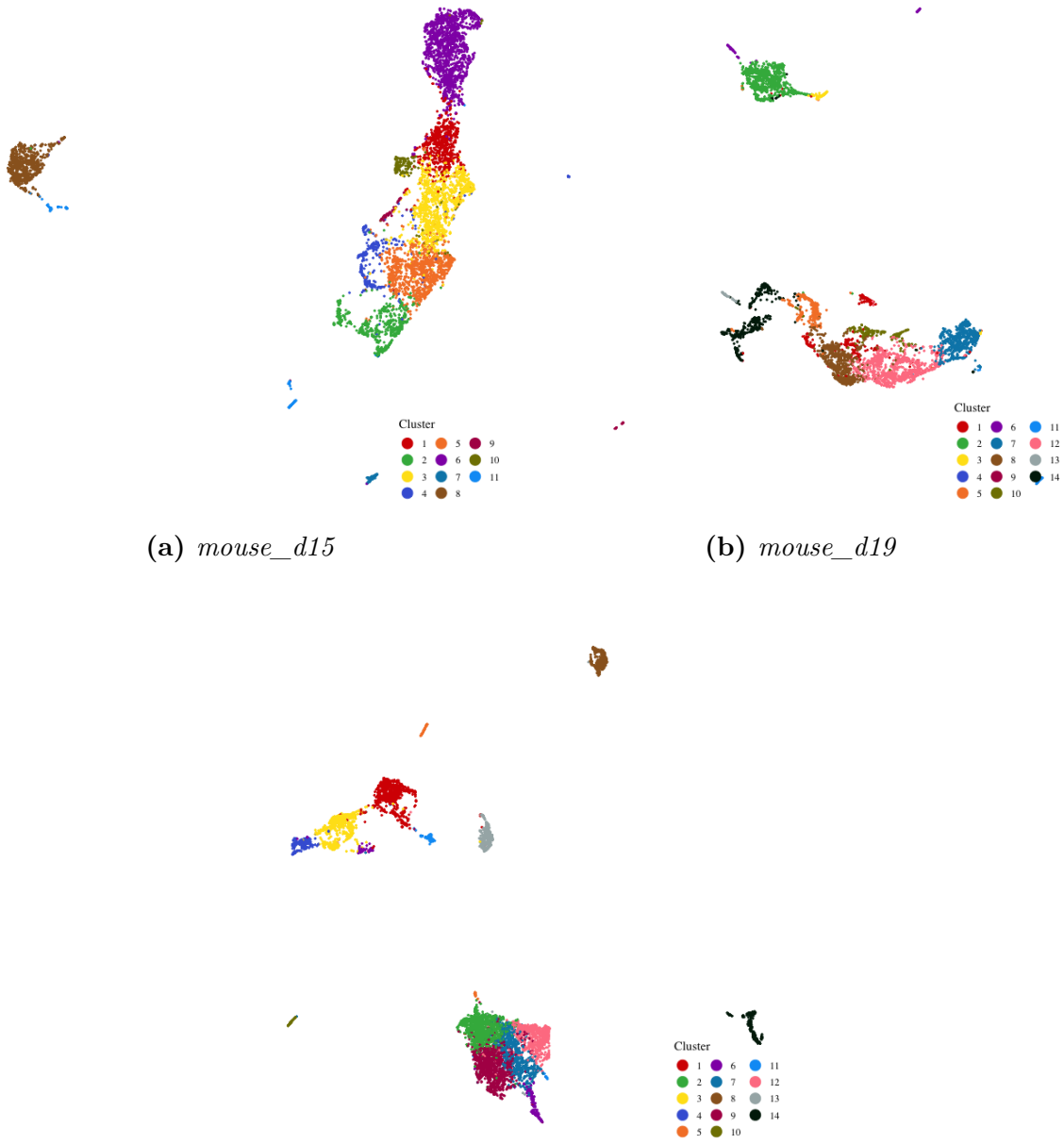
(a) *mouse_d15* **(b)** *mouse_d19* **(c)** *mouse_d25*

Fig. 4.14 UMAP of each sample overlaid with UMI count, uniquely detected gene counts, mitochondrial UMI percentage, and TCR presence. Each dot represents a cell.



(a) *mouse_d15* **(b)** *mouse_d19* **(c)** *mouse_d25*

Fig. 4.15 UMAPs of each sample overlaid with normalized, log2-transformed CD3, CD8, CD4, and FOXP3 gene expressions.



(a) *mouse_d15* **(b)** *mouse_d19*

(c) *mouse_d25*

Fig. 4.16 UMAP embedded data is colored by cluster membership. Each cluster is denoted with a different color. Points represent cells.

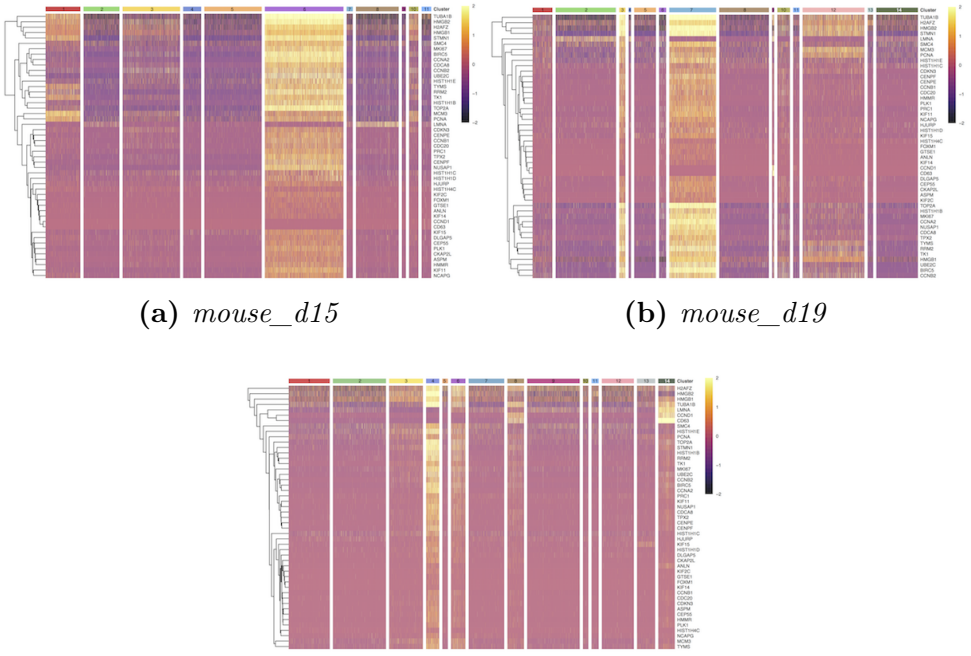
Based on their gene expression, in sample *mouse_d15*; cells in clusters 1, 2, 3, 4, 5, 6, 9, and 10 are $CD8^+$, cluster 8 are Tregs, and cluster 11 are naive T cells. In sample *mouse_d19*; cells in clusters 1, 5, 7, 8, 10, 12, 13, and 14 are $CD8^+$, clusters 2, 3, 4, and 6 are Tregs. And in sample *mouse_d25*; cells in clusters 1, 3, 4, and 6 (partially) are $CD8^+$, while cells in clusters 2, 6 (partially), 7, 9, and 12 are Tregs, and cells in clusters 11 and 13 are $CD4^+$.

At this point, there are a few things to note. Certain T cells cluster apart from their specific class of $CD8^+$ or $CD4^+$ based on the expression of ribosomal genes (RPL*/RPS*). For example, cluster 6 in *mouse_d19* can merge with cluster 2, but it has a very low RB-* expression. The low expression of ribosomal protein mRNA can signify cells that are in atrophy or senescence or the cell's cytoplasm could have been completely damaged. This can be used as another QC, but in this study, I let these cells cluster on their own instead, based on my justifications made in Section 2.5.1.

We can also observe cells displaying co-expression of CD8 and CD4, which is distorting the top-level clustering. One possible interpretation of this result is that we are detecting $CD8^+CD4^+$ DP mature T cells, which indeed have been reported in melanoma [55]. However, this observation could also be explained as an artifact introduced by the single-cell sequencing protocol. This second explanation is more likely to be true for some of the cells I found co-expressing IGHM and CD8A. These cells also bear TCRs and express genes which are normally involved in BCR recombination. While we could be observing a rare case of dual receptor-expressing lymphocytes as presented in [7], the artifact explanation remains more likely at this point. Since the V, D, and J gene expression levels are too low to be picked up directly from the GEX analysis, it is difficult to distinguish between real signal and noise in this case. It would be intriguing to speculate about IGHV genes expressed in these cells (Figure 4.52) and the correlation between CD8 and CD4, IGHM, and IGKC genes (Figure 4.51). However, at this point we cannot draw any conclusions without further, more complex validation experiments. It would be interesting to study the functionality and the TCR/BCR driven states of these cells and their receptor repertoires.

Lastly, looking at the heatmaps in Figures B.14, B.15, and B.16, we can see cell cycling clusters which are marked with genes such as TUBA1B, TOP2A, and STMN1. In order to distinguish the cell cycling clusters more surely, I took the cell cycling gene module of T cells from [140] and visualized their expression in a heatmap.

Based on the heatmaps shown in Figure 4.17, cluster 6 in *mouse_d15*, clusters 3, 7, and 12 in *mouse_d19*, and clusters 4 and 6 in *mouse_d25* are cell cycling. According to [140], the highest fraction of proliferating cells was among dysfunctional cells, and



(a) *mouse_d15* **(b)** *mouse_d19*

(c) *mouse_d25*

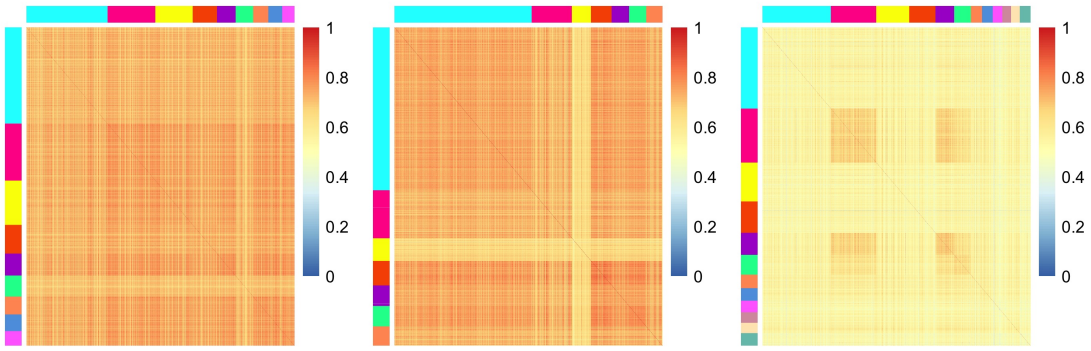
Fig. 4.17 Normalized, log2-transformed, and row-scaled expression of cell cycling genes determine the cell cycling clusters.

proliferation was most prominent in the earlier stages of dysfunctionality. However, dysfunctional cells occupied all stages of the cell cycle. Within the remainder of my analysis, I took cell cycle into consideration by assigning each cell to a cell cycle phase as previously described in Section 2.5.3.

Clonotype distribution within clusters

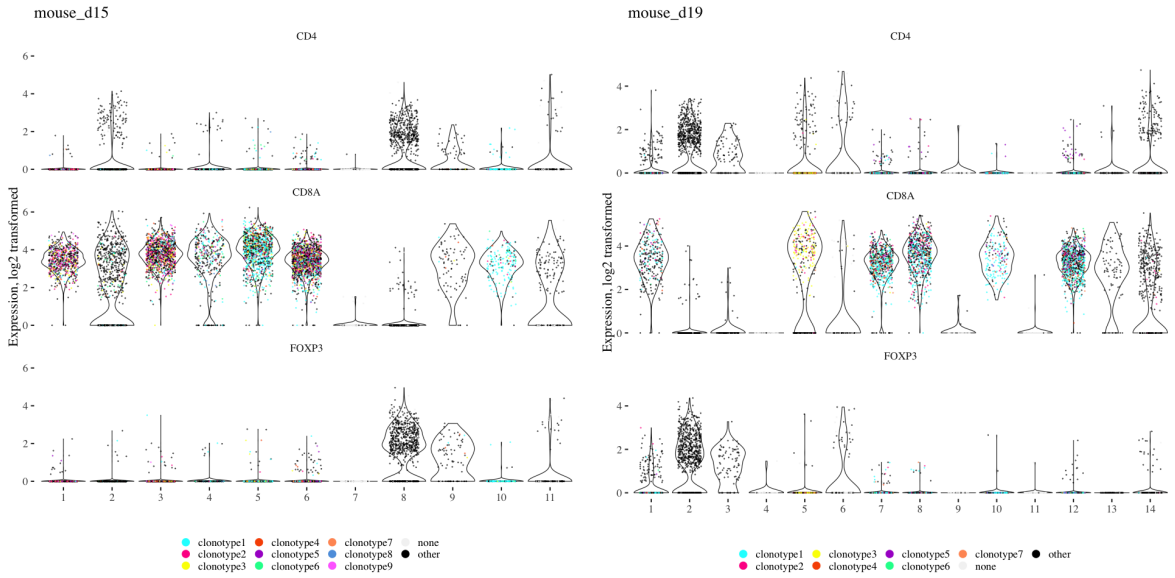
Having determined the clusters and their cell identities, I proceed to examine the clonotypes. First I investigated whether clonotypes were separated by top-level cell type; i.e., the presence of $CD8^+$, $CD4^+$ and $CD4^+$ $FOXP3^+$ clonotypes. In Figure 4.19 specific clonotypes are indeed of certain cell types. All of the clonotypes with at least 100 cells in *mouse_d15* and *mouse_d19* are $CD8^+$. However, this changes in *mouse_d25*; the most clonally expanded, i.e., the clone with the most number of cells, bears Treg clonotypes. Only clonotypes 2 and 5 are $CD8^+$. There is also a $CD4^+$ clonotype which we do not see in the other samples. As I mentioned before *mouse_d25* has a relatively small $CD8^+$ cell population. We can also see that some of the clonotypes are cluster-specific, suggesting that they are associated with the underlying gene expression profiles. Considering I did not use the TCR recombining genes when clustering the cells, any separation we

see is not driven directly by the TRA*/TRB* genes. As a next step, I overlaid the clonotypes on the UMAPs of each sample to visualize how the underlying GEX profile associates with the clonotypes in 2D space (see Figure 4.20). In samples *mouse_d15* and *mouse_d19* cells bearing the same clonotypes are not cluster-specific, however as expected they are cell type specific and they group close together in the 2D projection, whereas sample *mouse_d25* has clonotypes that cover specific clusters. At this stage it is of course not possible to draw any conclusions as the clustering can be made to be much more granular and for samples *mouse_d15* and *mouse_d19* we can start to see cluster-specific clonotypes as well. In order to investigate clonal specificity, I took the clonotypes as a unit of identity. For each sample, I calculated the cell-to-cell HVG expression correlation within clonotypes which have at least 100 cells; these are the clonotypes shown in Figure 4.20. Figure 4.18 shows heatmaps where grids are the calculated Pearson correlation coefficients for normalized log2-scale gene expression levels between all cells. In samples *mouse_d15* and *mouse_d19* we don't see an obvious difference within or between clonotypes (except maybe for clonotype3 in *mouse_d19*) but cells in clonotypes 2 and 5 in sample *mouse_d25* display greater correlation within clonotypes when compared to others. This may suggest that more granular clustering will not result in other cluster-specific clonotypes to emerge in sample *mouse_d25*. Note that clonotypes 2 and 5 of *mouse_d25* are $CD8^+$ whereas the others are Treg and $CD4^+$.



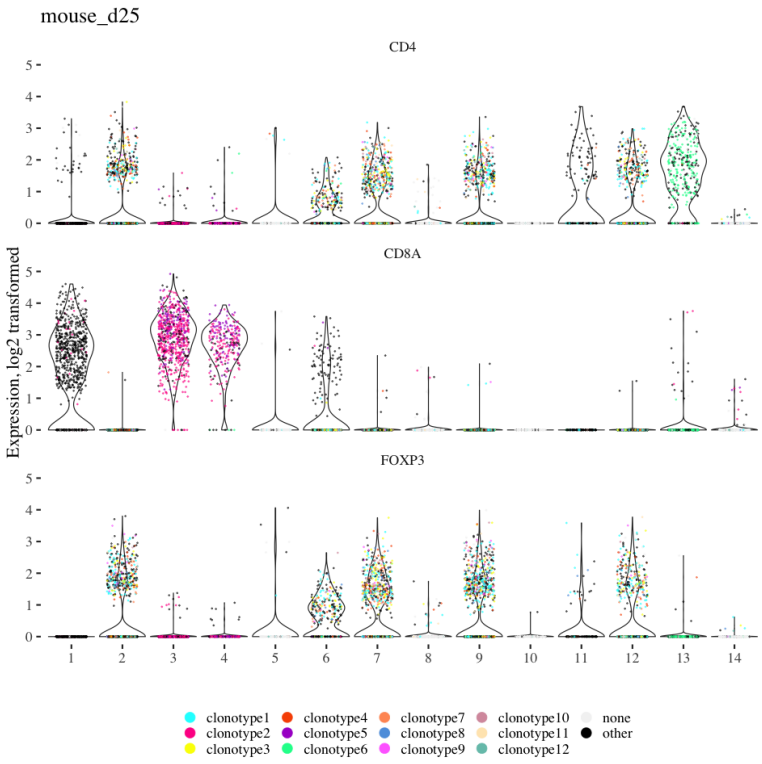
(a) *mouse_d15* **(b)** *mouse_d19* **(c)** *mouse_d25*

Fig. 4.18 Cell-to-cell correlation matrices for the normalized log2-scale gene expression levels between all cells within the top clonotypes. Each grid cell is colored based on the Pearson correlation coefficient.



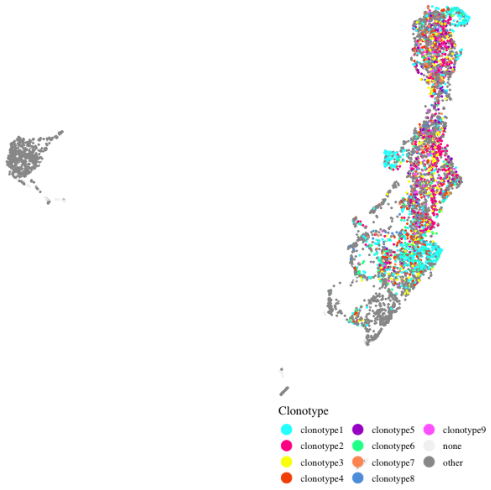
(a) *mouse_d15*

(b) *mouse_d19*

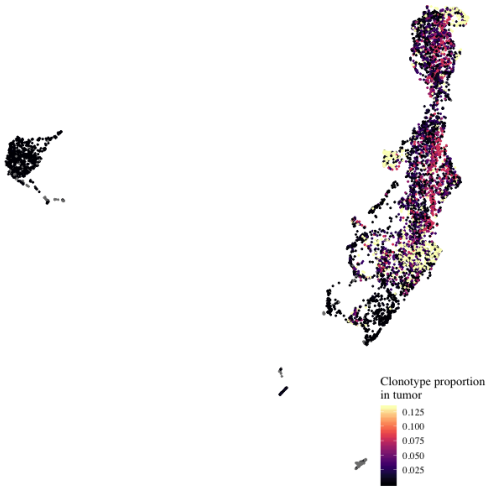


(c) *mouse_d25*

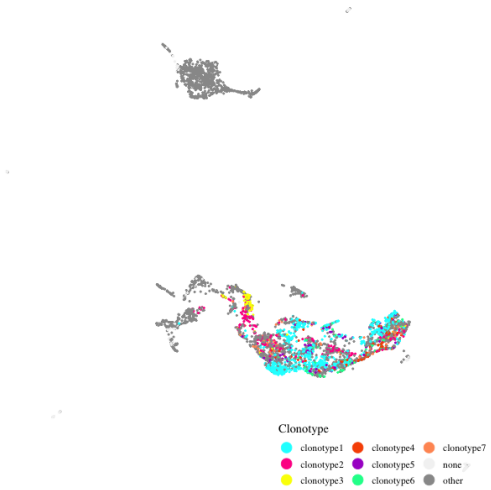
Fig. 4.19 Plots show the normalized, log2-transformed expression of CD4, CD8, and FOXP3 genes within each cluster. Each point represents a cell and is colored by its clone, i.e., its TCR's clonotype. Only the clonotypes represented with at least 100 cells are shown. None specifies cells with no TCR (TCR$^-$), and other specifies any other clonotype that did not make the cut as it had less than 100 cells. Each clonotype name is exclusive in its own sample; they are not shared between samples.



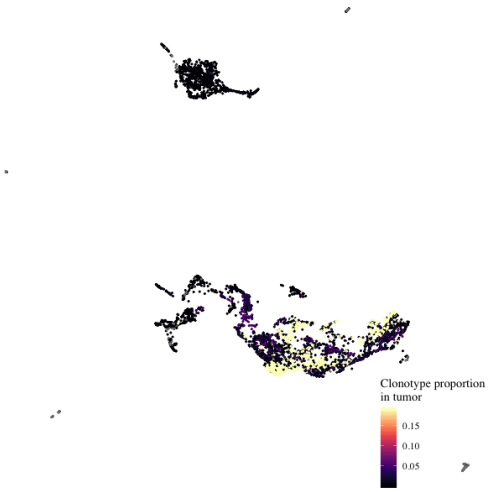
(a) *mouse_d15*



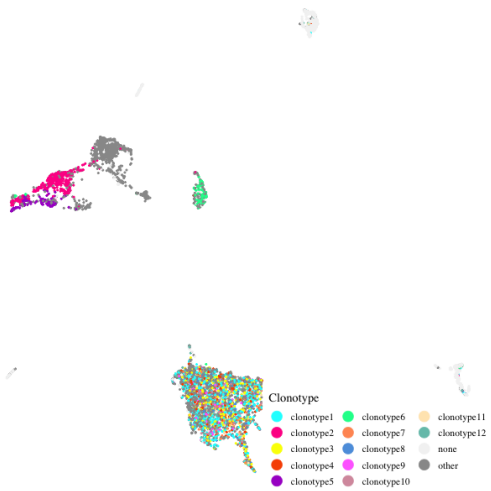
(b) *mouse_d15*



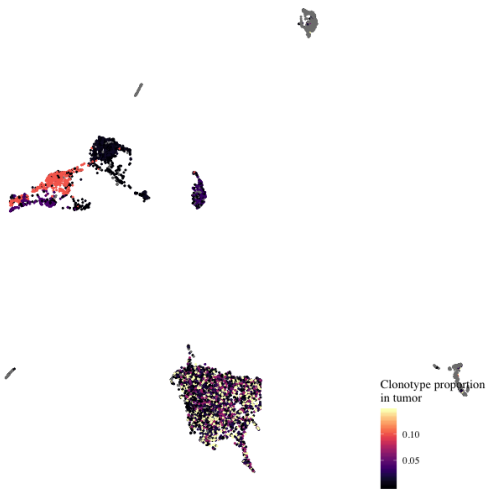
(c) *mouse_d19*



(d) *mouse_d19*



(e) *mouse_d25*



(f) *mouse_d19*

Fig. 4.20 Clonotype overlay on UMAPs.

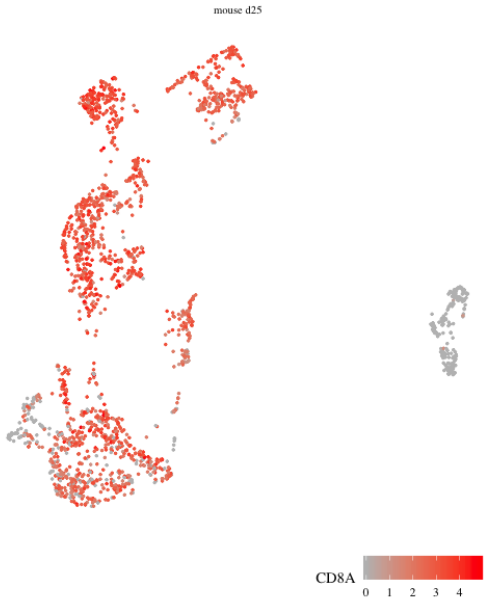


Fig. 4.21 UMAP of CD8 expression of the all cells in clusters 1, 3, 4, and 6 of *mouse_d25*, without any filtering on UMI count. Here all the cells with no CD8 UMIs cluster apart further to the right. They do not contribute to the analysis but can distort clustering due to purely technical reasons.

4.4.2 V(D)J and Gene Expression Analysis of CD8$^+$ TILs

Next, I subsetted the tumor samples for CD8$^+$ cells to further investigate their clone specific gene expression profiles. I took all the clusters defined as CD8$^+$, i.e., differentially expressing CD8 (Figure B.17), and filtered out cells in which no CD8 UMI was detected. As our initial clustering was not very granular, it is possible to see some cells which do not express CD8 but still cluster together with CD8$^+$ cells based on the closeness of other gene expression distances. I could have left these cells to cluster apart (see Figure 4.21 for a UMAP of the subsetted cells when keeping all the cells without filtering on positive CD8 UMI count) however I aimed to avoid their UMI count to distort normalization.

When compared to the other two samples *mouse_d25* has many fewer CD8$^+$ cells (see Figure 4.22). The TME of *mouse_d25* has a much more heterogeneous tumor-reactive cell population. I also removed cells that did not bear a TCR (less than 2%, 2%, and 4% in samples *mouse_d15*, *mouse_d19*, and *mouse_d25* respectively) as I wanted to focus on the clonotype specific response. Instead of performing further clusterings within clusters, I used the initial clustering as a single cell stratification. The immune cells vary in size, and when normalizing their UMI count, specific cell types may pull the size factors to the extreme and hence distort the gene expression levels. Here I pool together all the subsetted CD8$^+$ T cells and normalize them following the same pre-clustering and deconvolution method. See Figure 4.23 for the library size and uniquely detected gene count distributions, as well as cell complexity and the newly calculated size factors for the subsetted cells. It is worth noting that the range of the size factors is now narrower.

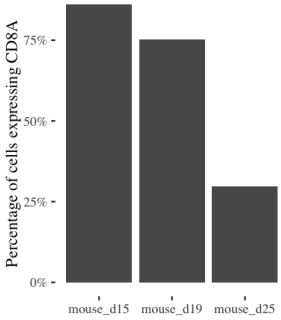
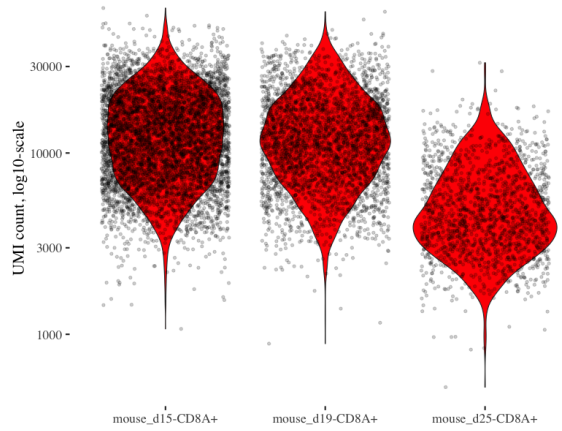
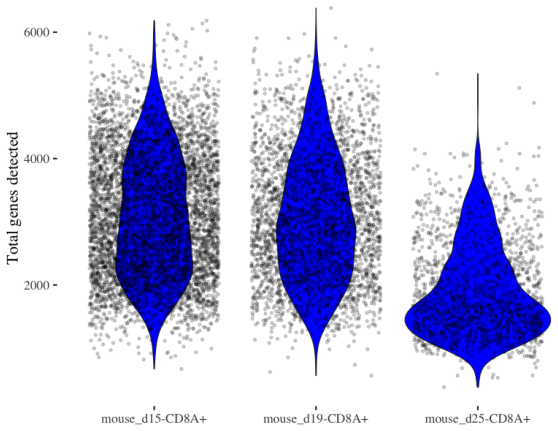


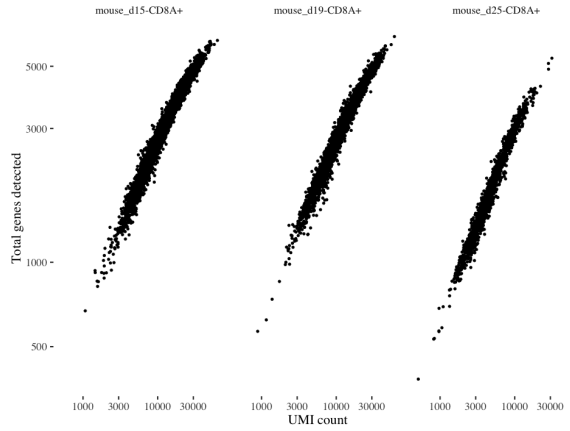
Fig. 4.22 CD8 expressing cell proportions. *mouse_d25* has much less $CD8^+$ cells when compared to the other samples. The TME of sample *mouse_d25* is much more heterogeneous which could be due to both the suppression of $CD8^+$ cells and the global response of the immune system mobilizing various immune cells to the tumor over time.



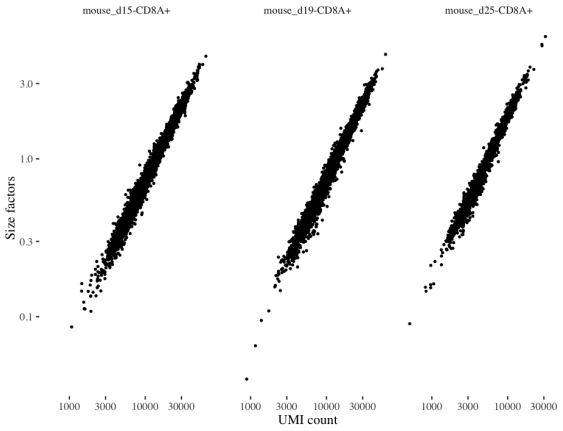
(a) Library size distributions



(b) Uniquely detected gene count distributions



(c) Cell complexities

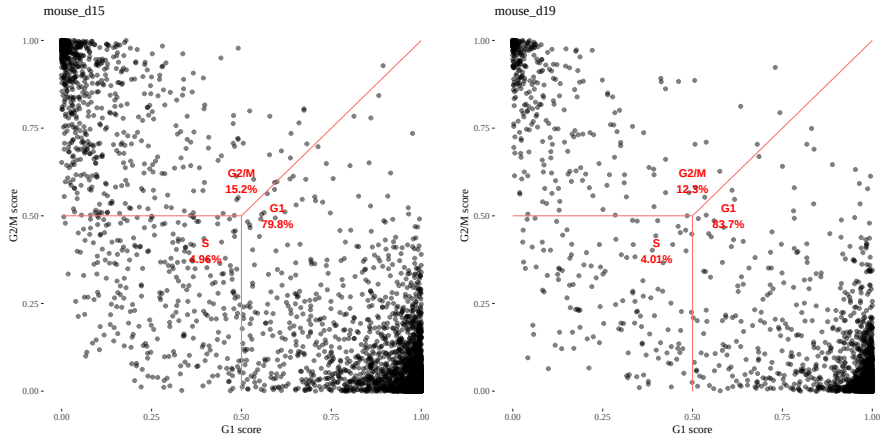


(d) Size factors

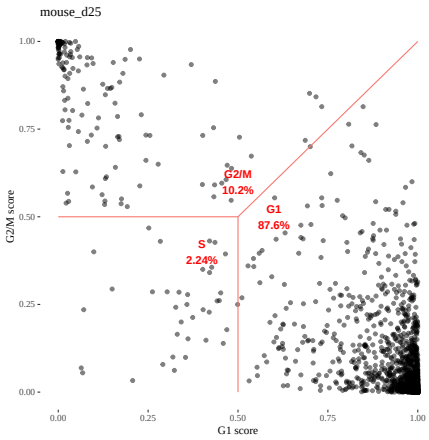
Fig. 4.23 Subsetted $CD8^+$ cells

Cell cycling

As I mentioned before in Section 4.4.1, our samples comprise cells from different cell cycle phases. I identified this by observing the expression of known cell-cycle marker genes. As such, [207] have proposed a method (*Pairs*) to infer the cell cycle phase of a cell directly from its gene expression profile. Their classification algorithm, using a training dataset, found pairs of marker genes that can assign a phase to a cell. This method is described in-depth in Section 2.5.3. A slightly altered version of this method is implemented in `scran::cyclone` with an existing pre-trained set of marker pairs for mouse data. Using this function I assigned cells into cell cycle phases based on their normalized gene expression data (see Figure 4.24).

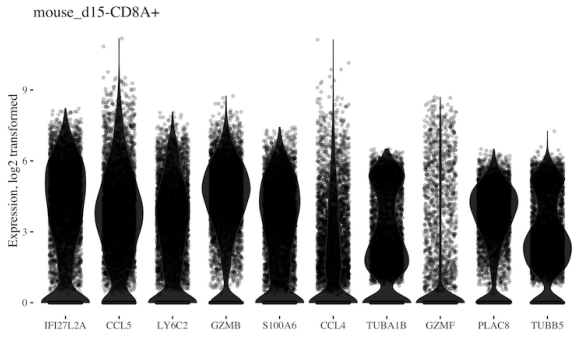


(a) *mouse_d15* **(b)** *mouse_d19*

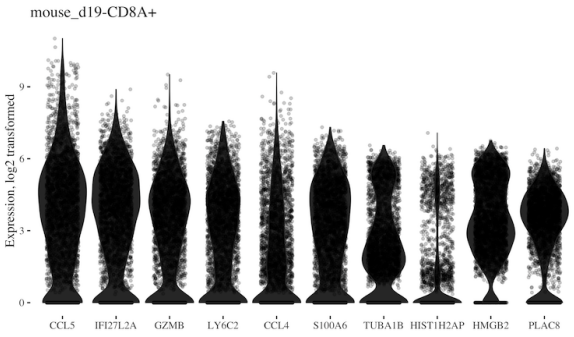


(c) *mouse_d25*

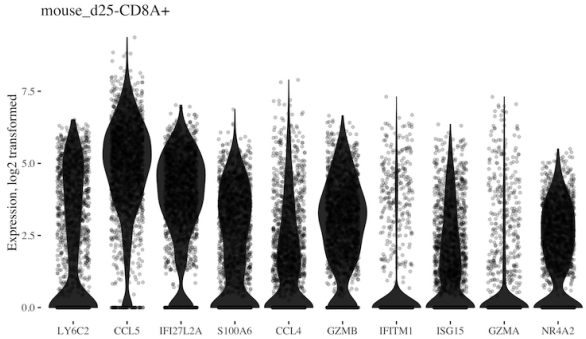
Fig. 4.24 Cell cycle phases of subsetted CD8$^+$ cells. Cell cycle phase scores of samples where each point represents a cell. Red lines separate cell phases and proportion of cells in each phase are written in red.



(a) *mouse_d15* CD8$^+$ HVGs



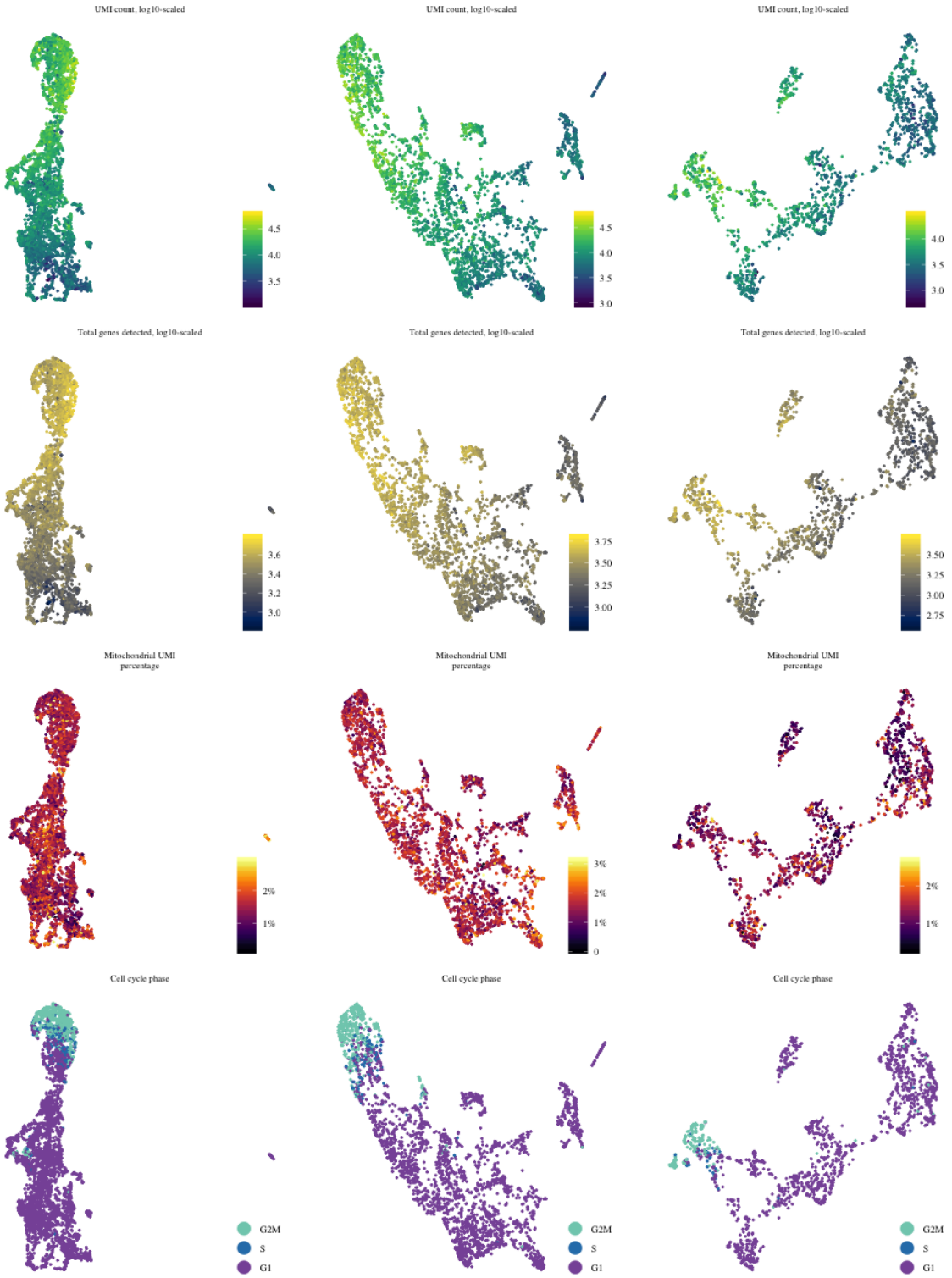
(b) *mouse_d19* CD8$^+$ HVGs



(c) *mouse_d25* CD8$^+$ HVGs

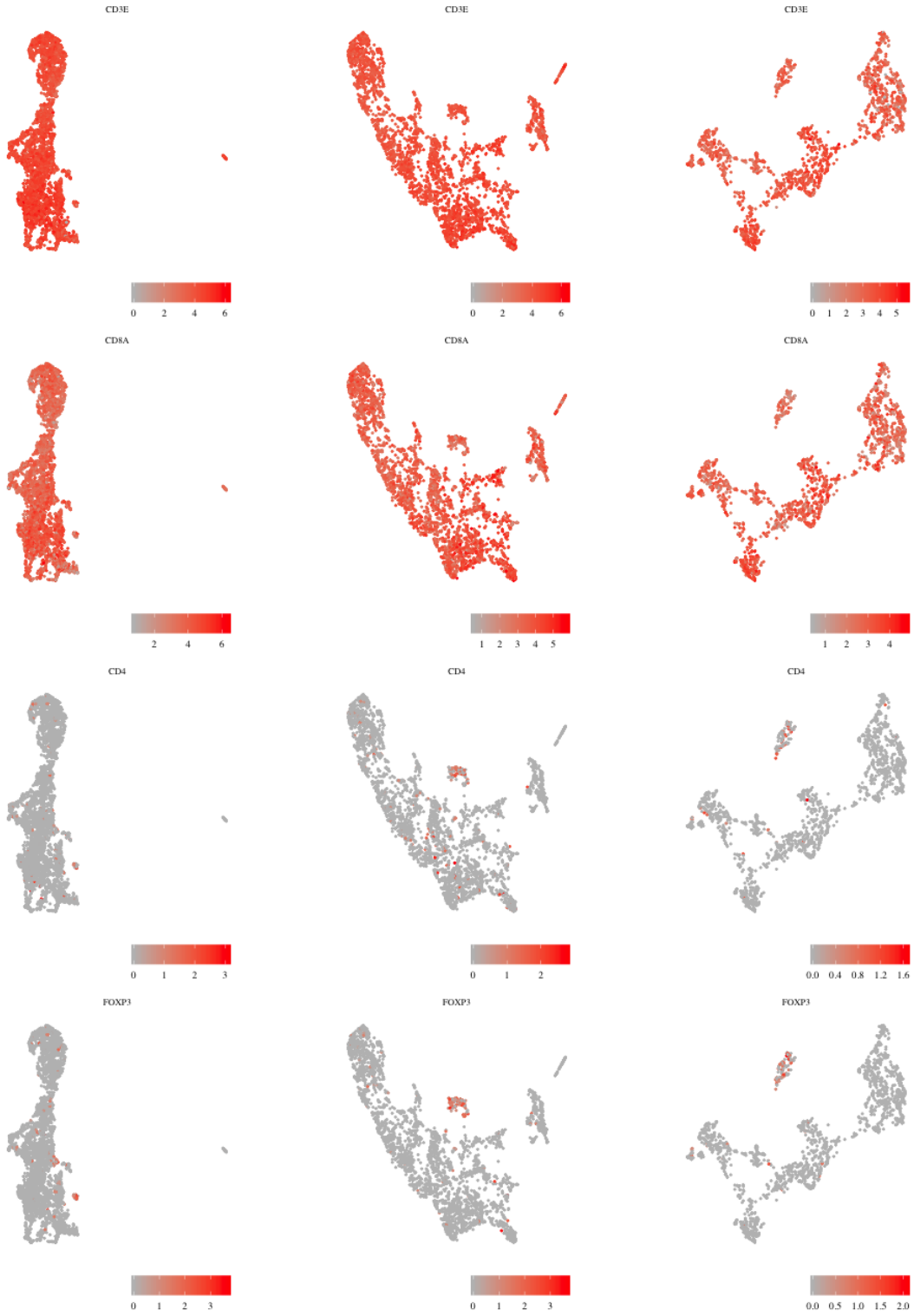
Fig. 4.25 HVGs of CD8$^+$ cells.

Next, similar to before, I found the HVGs (Figure 4.25), performed dimension reduction to determine PCs driven by biology, projected data onto two dimensions with UMAP (Figures 4.26, 4.27), and clustered (Figure 4.28) each sample.



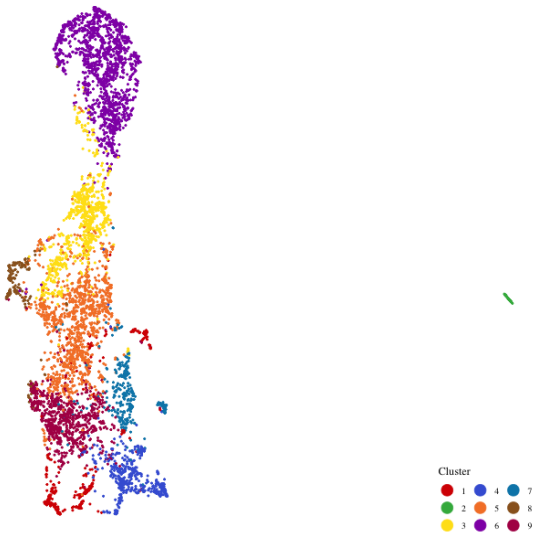
(a) *mouse_d15* **(b)** *mouse_d19* **(c)** *mouse_d25*

Fig. 4.26 UMAP of each CD8$^+$ subset overlaid with UMI count, uniquely detected gene counts, mitochondrial UMI percentage, and cell cycle phase. Each dot represents a cell.

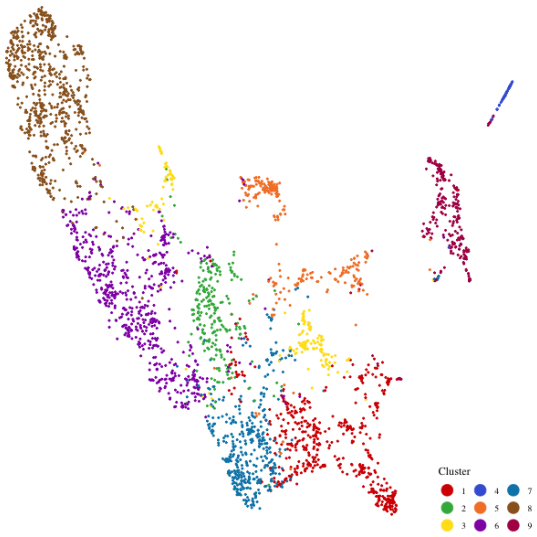


(a) *mouse_d15* **(b)** *mouse_d19* **(c)** *mouse_d25*

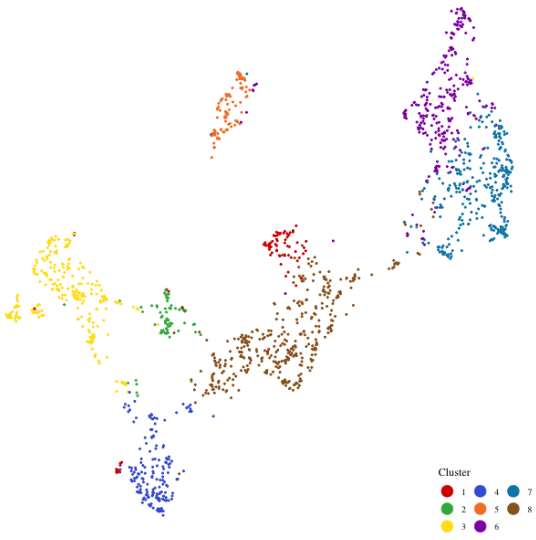
Fig. 4.27 UMAPs of each CD8$^+$ subsets overlaid with normalized, log2-transformed CD3, CD8, CD4, and FOXP3 genes.



(a) *mouse_d15*



(b) *mouse_d19*



(c) *mouse_d25*

Fig. 4.28 CD8$^+$ cell clusterings. UMAP embedded data are colored by cluster membership. Each cluster is denoted with a different color. Points represent cells.

Figure 4.29 shows the cell frequency in each cluster colored by the cell cycle phase. Cluster 6 in *mouse_d15*, 8 in *mouse_d19*, and 3 in *mouse_d25* are cell cycling. There are also cells in S and G2M phases in other clusters, but these are negligible. When finding the cluster biomarkers, I discarded these cells and the cell cycling clusters.

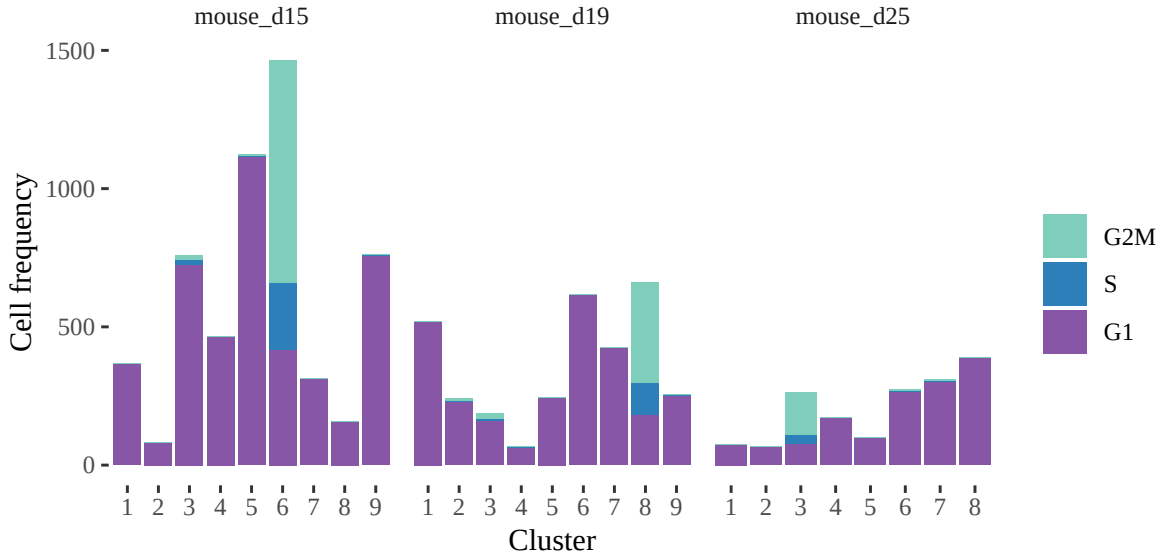


Fig. 4.29 For each sample, the bar plots show the cell count in clusters. Bars are colored by phase membership.

The rationale behind discarding the whole cell cycling cluster instead of only the G2M and S phase cells is that the phase allocations of these cells have a low degree of confidence when compared with the G1 phase cells in other clusters. Figure 4.30 shows this with *mouse_d19* as an example. This could suggest that these cells are cycling as well and close to duplication. Hence I removed the whole cell cycling cluster in order to identify subpopulations accurately.

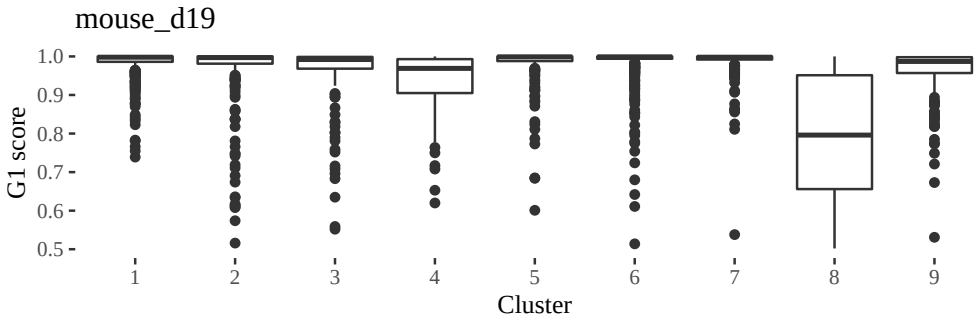
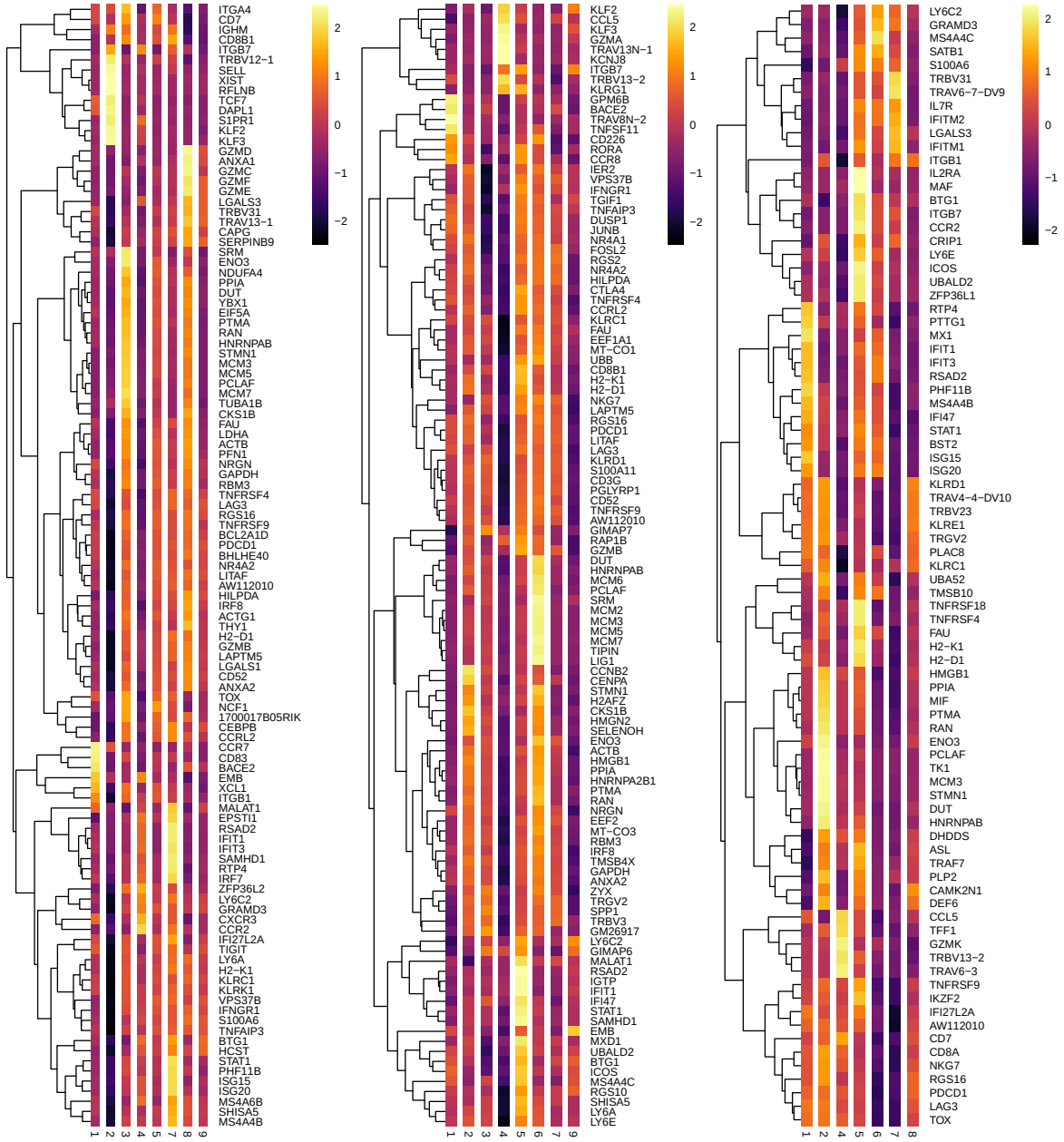


Fig. 4.30 G1 score distribution of cells in each cluster that are in the G1 phase.

Identifying clusters

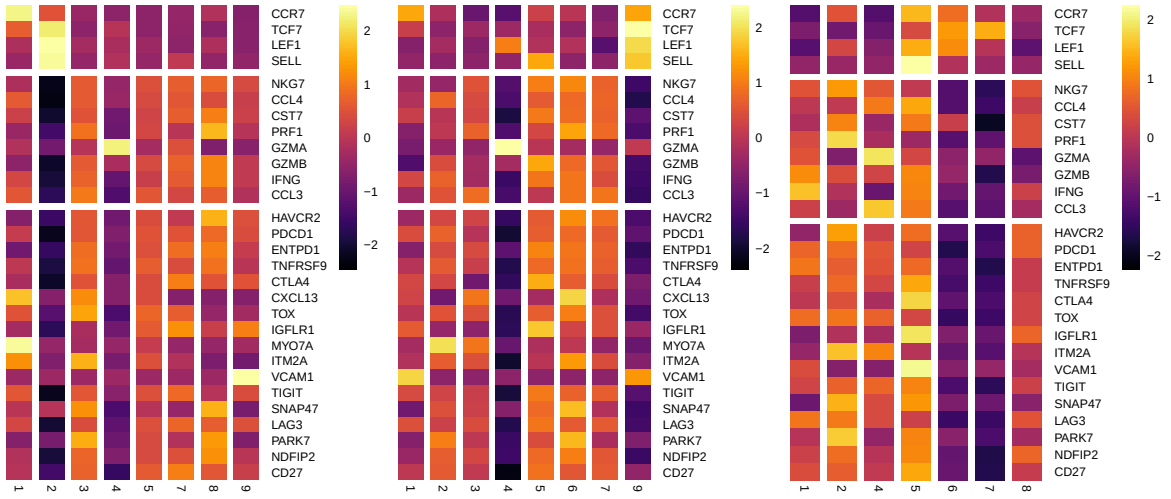


(a) *mouse_d15* **(b)** *mouse_d19* **(c)** *mouse_d25*

Fig. 4.31 Heatmaps show the log2-transformed average normalized expression of biomarkers.

I found the upregulated genes in each cluster to identify cluster biomarkers. Within this sub-clustering we can, for example, see that cluster 2 of *mouse_d15* are naive T cells (SELL, TCF7), cluster 4 of *mouse_d19* are cytotoxic T cells, and so are cluster 4 of *mouse_d25*. I next took CCR7, TCF7, LEF1, and SELL naive T cell markers, and

NKG7, CCL4, CST7, PRF1, GZMA, GZMB, IFNG, CCL3 cytotoxic T cell markers [235], and the intersection of tumor CD8$^+$ dysfunction-related genes identified in [235], [265], [268], and [89], and looked at their average normalized expression in each cluster in order to distinguish naive, cytotoxic, and dysfunctional clusters.



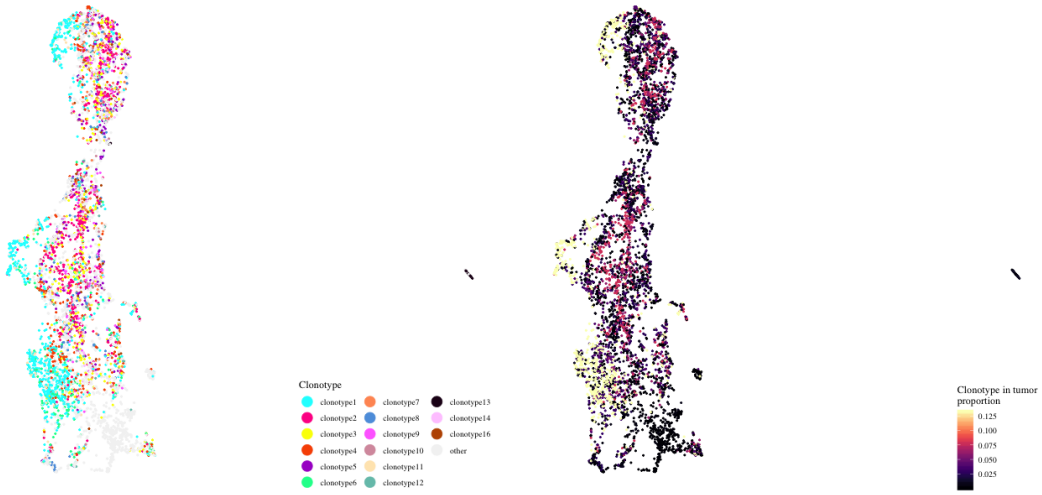
(a) *mouse_d15* **(b)** *mouse_d19* **(c)** *mouse_d25*

Fig. 4.32 Heatmaps show the log2-transformed average normalized expression of naive, cytotoxic, and dysfunctional biomarkers for CD8 cells.

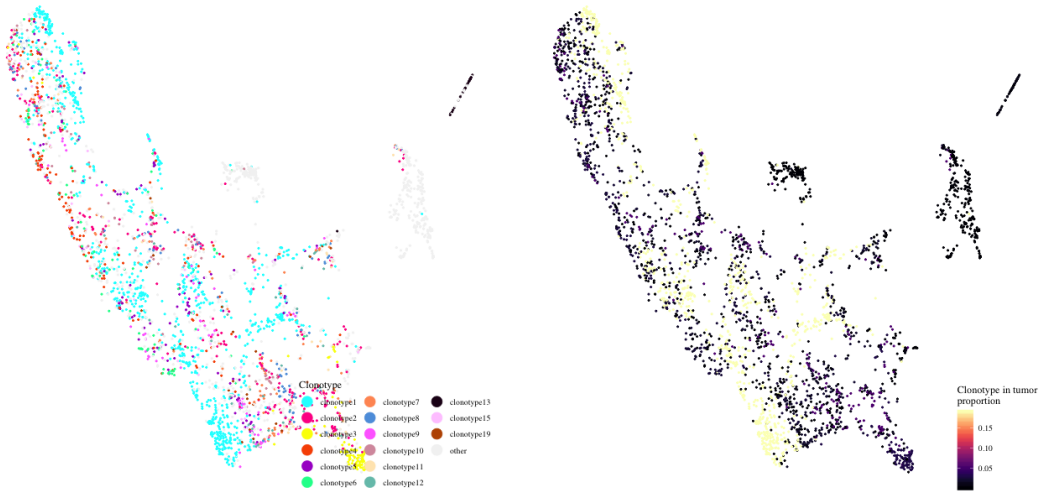
The separation is mostly clear for naive T cell clusters. However, it is possible to cluster the cytotoxic and dysfunctional clusters further. Instead of sub-clustering until "biologically desired" clusters are achieved, I took the clonotypes as the unit of grouping and found the clonotype biomarkers.

Identifying clonotype biomarkers

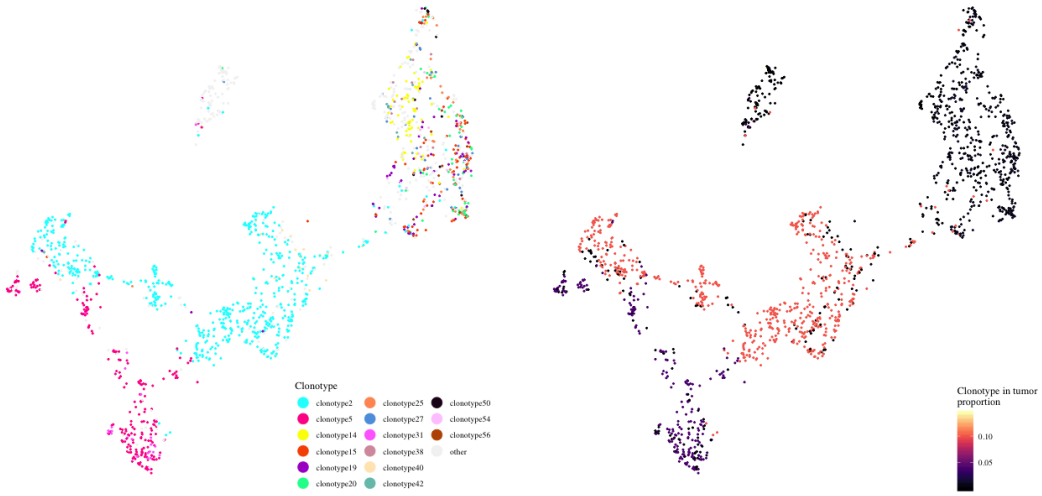
Figure 4.33 shows the overlay of clonotypes on the UMAPs of each sample. Observe that the gene expression clustering and clonotype bearing overlaps in sample *mouse_d25*. If we find the up/down-regulated genes of these clusters, we can identify the state of the clonotypes. Note that the overlays include the cell cycling clusters; however, when determining the clonotype markers, I did not include them in my analysis. I also do not consider the clonotypes that are represented by less than 50 cells. I ran both `scran::findMarkers` and `scran::overlapExprs` to identify the differentially expressed genes. `scran::overlapExprs` performs pairwise Wilcoxon rank sum tests between groups of cells, whereas `scran::findMarkers` uses the Welch t-test. Heatmaps showing the expression of clonotype biomarkers of each mouse can be found in Figures B.18 - B.23.



(a) *mouse_d15* **(b)** *mouse_d15*



(c) *mouse_d19* **(d)** *mouse_d19*



(e) *mouse_d25* **(f)** *mouse_d19*

Fig. 4.33 Clonotype overlay on UMAPs.

4.4.3 Integrated analysis of tumor and spleen cells

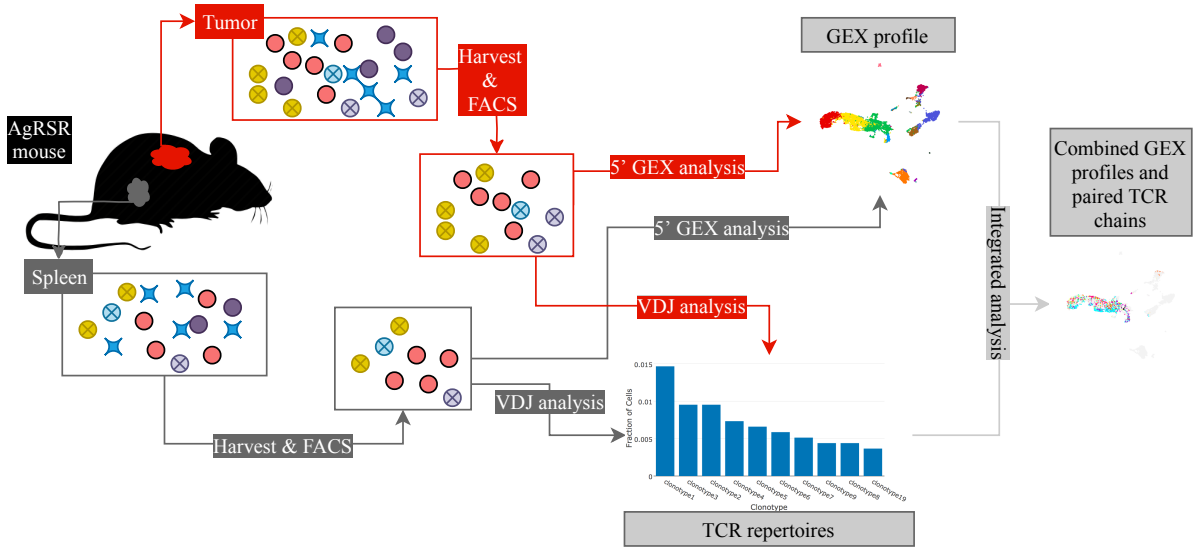


Fig. 4.34 Graphical overview of the integrated approach. Single cells were collected from the AgRSR mouse tumor and spleen, sorted with FACS, processed with Chromium Single Cell Immune Profiling Solution, and analyzed in combination.

I analyzed the cells harvested from the mouse tumor and spleen together to identify shared and differentiated cell states. I also wanted to see whether the immune response was localized, or if there were common clonotypes between the tumor and spleen. For this, I first compared each mouse's tumor and spleen TCR clonotypes. We might recall that cells with the same productive CDR3 regions in their TCRs' α and β chains are considered to be of the same clone, and their TCRs the same clonotype. Hence I compared the CDR3 regions of tumor and spleen cell TCRs with the requirement that they are an identical match in their amino acid sequence. For the tumor samples, I considered only the clonotypes with at least two cells, however, for the spleen samples I did not specify a cutoff as the clone frequencies are much lower in the spleen when compared to the tumor. Additionally, I discarded clonotypes with single chains. Table 4.5 shows the similarity between each mouse's tumor and spleen samples based on different metrics.

Compared to the other two samples, *mouse_d15* has the most overlapping clonotypes with the top clonotype (clonotype1) having 912 cells in the tumor and 136 in the spleen (see Figure 4.35). Also, we know that some of the cells within this clone express $CD8^+$ dysfunctional-state markers such as LAG3, and PDCD1. I say some as this clonotype spans multiple clusters: 5, 6, 8, and 9 (see Figure 4.33a) where cluster 6 is cell cycling. Furthermore, all of the top clonotypes, except for one (clonotype5), which are present in the tumor are also represented in the spleen.

	mouse_d15	mouse_d19	mouse_d25
Shared clonotypes	58	51	39
Shared CDR3s	119	109	55
Barcodes in shared clonotypes	3728	1160	3902
(tumor:spleen)	(3189:539)	(921:239)	(3738:164)

Table 4.5 The similarity between tumor and spleen samples.

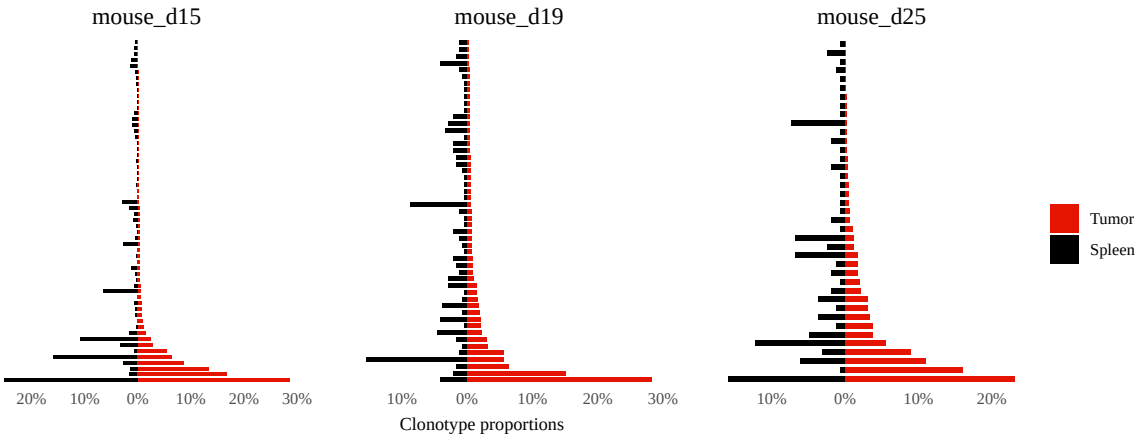


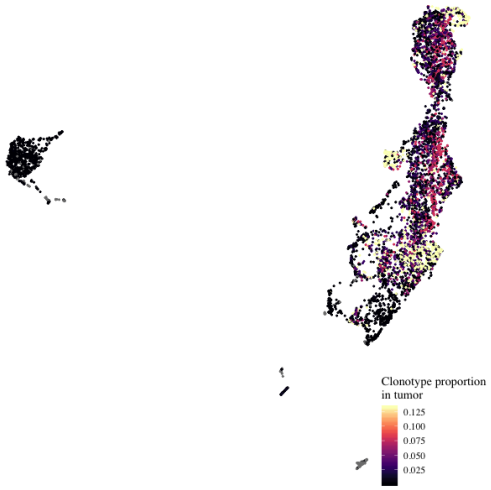
Fig. 4.35 Shared clonotype proportions in tumor and spleen samples.

Sample *mouse_d19* has its highest overlap in clonotype2 with 261 cells in the tumor and 10 cells in the spleen. The top clonotype of the tumor with 897 cells is not represented in the spleen. The next highest clonotype with 139 cells is represented with only 5 cells in the spleen. In fact, except for these two clonotypes, none of the clonal tumor sample TCR$\alpha\beta$ chain pairings are represented in the spleen.

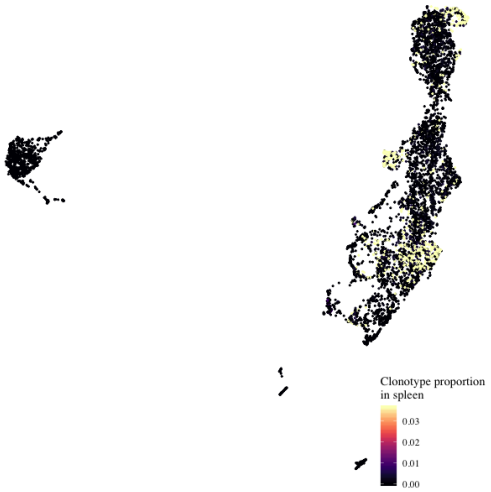
Although *mouse_d25* has overlapping clones these are not CD8$^+$ cells. The two cluster specific clonotypes 2 and 5 are not represented in the spleen. Figure 4.36 shows the clonotype proportions of all the clonotypes in the tumor, and the proportions of the overlapping clonotypes in the spleen.

It is also important to note that the cluster specific clonotypes 2 and 5 of *mouse_d25* express dysfunctional-state markers such as LAG3, TOX, TIGIT, and PDCD1. The top overlapping clones (clonotype1) are Tregs; this is clear in Figure 4.19c. This clonotype is represented with 863 cells in the tumor and with 26 cells in the spleen.

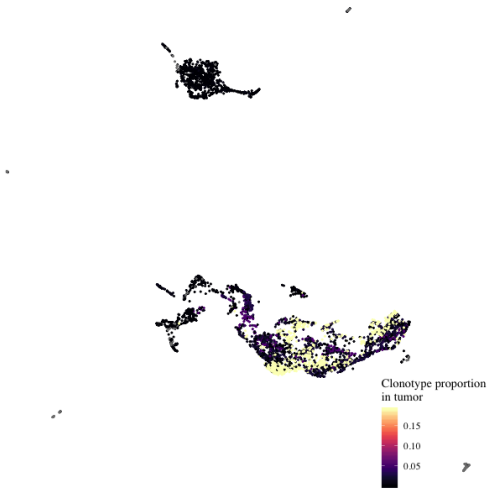
Because of our interest in the dysfunctional CD8$^+$ cells, and due to the significant number of cells in overlapping clonotypes, I decided to analyze the tumor and spleen cells of sample *mouse_d15*.



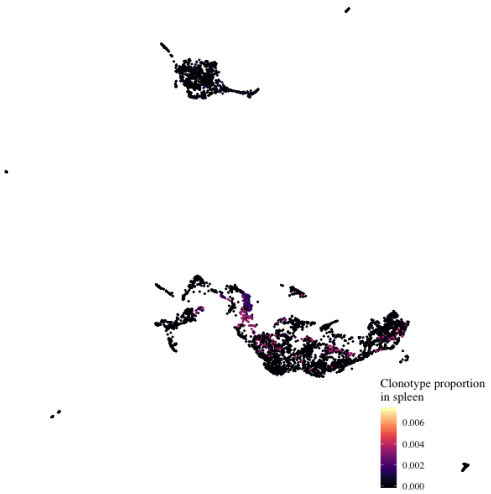
(a) *mouse_d15*



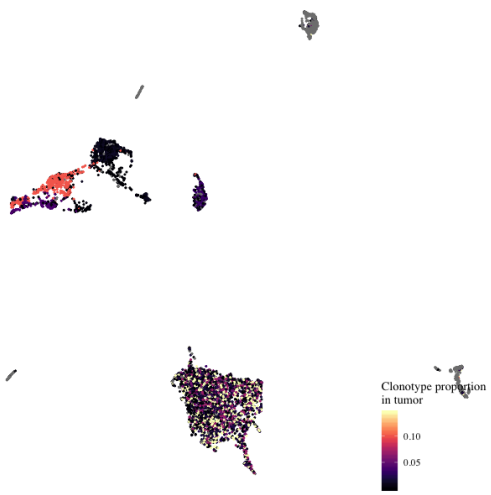
(b) *mouse_d15*



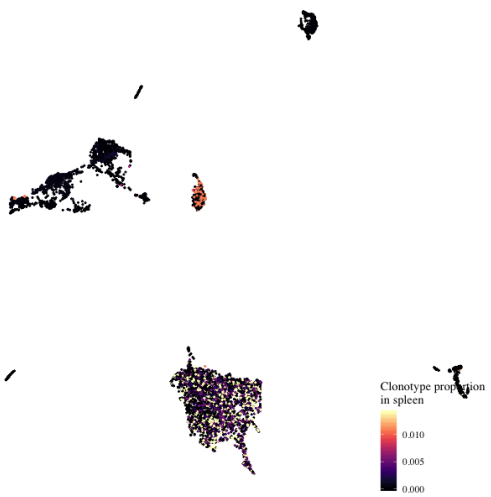
(c) *mouse_d19*



(d) *mouse_d19*



(e) *mouse_d25*



(f) *mouse_d19*

Fig. 4.36 UMAPs show the clonotype proportions, before $CD8^+$ stratification, in tumor and spleen samples.

UMI counts cannot be directly compared between samples: multiple reads with the same UMI and barcode which align to the same gene counts once towards the absolute number of observed transcripts per gene in a cell. However, these absolute counts are not normalized for sequencing depth or RNA content per cell. Furthermore, technical sources of variation may be added to the samples during handling. Called *batch effects*, these variations include external factors associated with labwork such as who prepared the libraries and when, what equipment they used, what was the reagent quality, what was the lab temperature that day, and so on. These circumstantial differences can cause systematic differences in gene expression in cells coming from different samples, making it difficult to identify the real biological variation by masking or intertwining the real signal with noise. Therefore, we need to correct for batch effects when analyzing the spleen and tumor samples together.

Preprocessing of the *mouse_d15* spleen sample

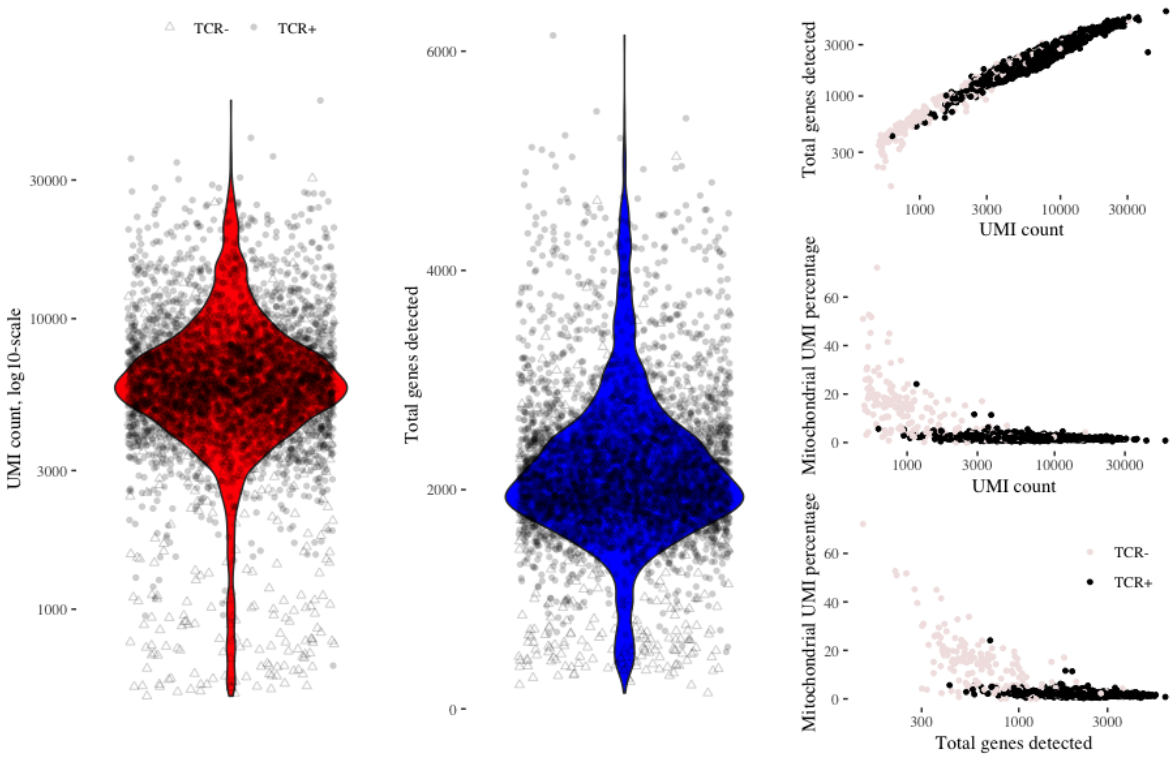


Fig. 4.37 *mouse_d15* spleen sample QC plots.

Just like the tumor samples, cells were harvested from the spleens of the AgRSR mice, sorted with FACS for CD45+CD11B-EYFP+, and processed with the Chromium Single Cell Immune Profiling Solution. And as before, I ran `cellranger count` and

`cellranger vdj` on the obtained spleen sample RNA-seq reads to obtain the gene expression profiles and the TCR repertoires, respectively.

`cellranger count` estimated a total of 3,992 cells in the *mouse_d15* spleen sample. Using the same QC approach as before (see Section 4.4.1) I removed 195 low quality cells which can be seen in Figure 4.38. Also see Figure 4.37 for the QC diagnostic plots. The cells with low numbers of uniquely detected genes and high mitochondrial read fractions are mostly TCR$^-$; hence, it is safe to discard them.

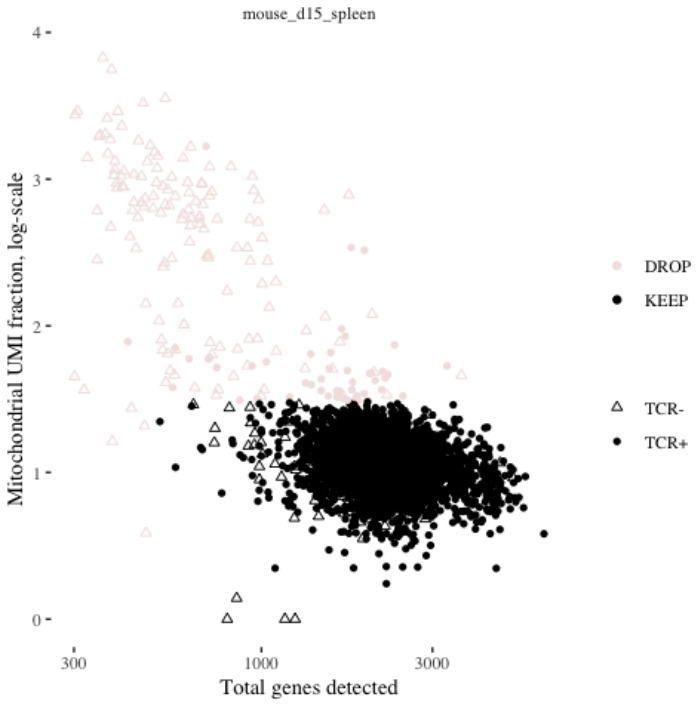


Fig. 4.38 *mouse_d15* spleen sample cells colored based on discard/keep, and differentiated by shape based on TCR detection.

I next merged the two tumor and spleen datasets subsetting on the shared HVGs (7,331 genes) between the samples. We can explore the possible batch effects with a PCA. Figure 4.39 plots the first two dimensions of the PCA; PC1 and PC2, which explain 22 and 7 percent of the total variation, respectively. Here each dot represents a cell colored by its originating sample. Having performed a PCA using the merged datasets, we can see a clear separation between the two samples: Cells from different samples mostly cluster together. As mentioned above, an apparent cause of this separation is the difference in the sequencing depth between samples. In order to remove this technical consequence, I used the aggregation method implemented as part of the Cell Ranger analysis pipelines.

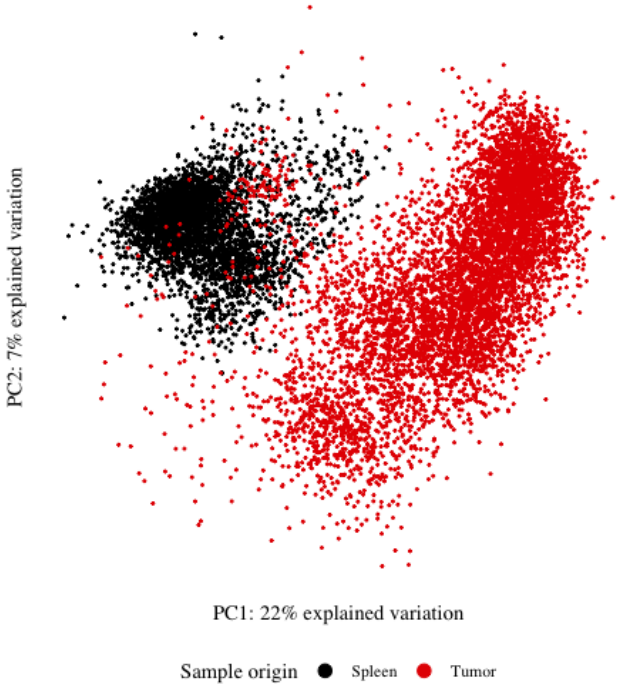


Fig. 4.39 PCA projection of *mouse_d15* merged tumor and spleen samples before batch correction where each dot represents a cell colored by its origin.

Normalizing for differences in sequencing depth between samples

The `cellranger aggr` pipeline aggregates the individual `cellranger count` outputs of each sample by normalizing them to the same sequencing depth, producing a single count matrix on the combined data. I first ran `cellranger count` on each sample separately. Then, to be able to compare them, I aggregated them with `cellranger aggr`. The `cellranger aggr` pipeline normalizes for sequencing depth by subsampling the reads. When subsampling, the pipeline only considers reads which aligned to valid barcodes and UMIs. No cell calling is performed as it directly uses the cells called via `cellranger count`. The pipeline finds the sample with the lowest number of reads and downsamples all the other samples to this threshold by computing a subsampling rate for each sample and keeping only the necessary fraction of reads. The subsampling rate is based on the mean number of filtered reads mapped to the transcriptome per cell.

Table 4.6 lists the summary metrics of the GEX barcoding and sequencing process on the cells obtained from the spleen and the tumor of *mouse_d15*. Here mean reads per cell is calculated by dividing the total number of sequenced reads by the estimated number of cells. This average is based on all the reads and not just the ones that mapped confidently

	Spleen	Tumor
Estimated Number of Cells	3,992	7,633
Mean Reads per Cell	75,212	39,707
Median Genes per Cell	2,083	2,571
Number of Reads	300,249,980	303,085,056
Valid Barcodes	93.5%	94.1%
Fraction Reads in Cells	97.1%	95.6%
Reads Mapped to Genome	95.9%	96.1%
Reads Mapped Confidently to Transcriptome	79.3%	81.5%

Table 4.6 GEX analysis metrics of *mouse_d15* samples

to the transcriptome. The mean number of confidently-mapped-to-transcriptome, valid-barcode reads per cell in the spleen sample, prior to depth normalization, is 55,938, and for the tumor, it is 29,949. So `cellranger aggr` took the lower count of 29,949 of the tumor sample as a threshold, and downsampled the spleen sample reads until they both had an equal number of confidently mapped reads per cell. The results can be seen in Table 4.7.

Estimated Number of Cells	11,625
Post-Normalization Mean Reads per Cell	39,899
Median Genes per Cell	2,267
Pre-Normalization Number of Reads	603,335,036
Post-Normalization Number of Reads	463,836,466
Fraction of Reads Kept in Tumor	100%
Fraction of Reads Kept in Spleen	53.5%
Pre-Normalization Confidently Mapped Barcoded Reads per Cell in Tumor	29,949
Pre-Normalization Confidently Mapped Barcoded Reads per Cell in Spleen	55,938

Table 4.7 Aggregate metrics of *mouse_d15* samples.

Here 53.5% of the spleen sample reads are kept so that the tumor and spleen samples both have the same sequencing depth, measured in terms of reads that are confidently mapped to the transcriptome per cell.

After normalizing the samples for sequencing depth, I processed the combined count matrix. I removed the barcodes that were discarded when processing the *mouse_d15* tumor and spleen samples separately. I then normalized the count matrix using the previously described method (see Section 4.4.1) and subsetted on the shared HVGs. The sequence-depth-corrected combined sample PCA projection can be seen in Figure 4.40.

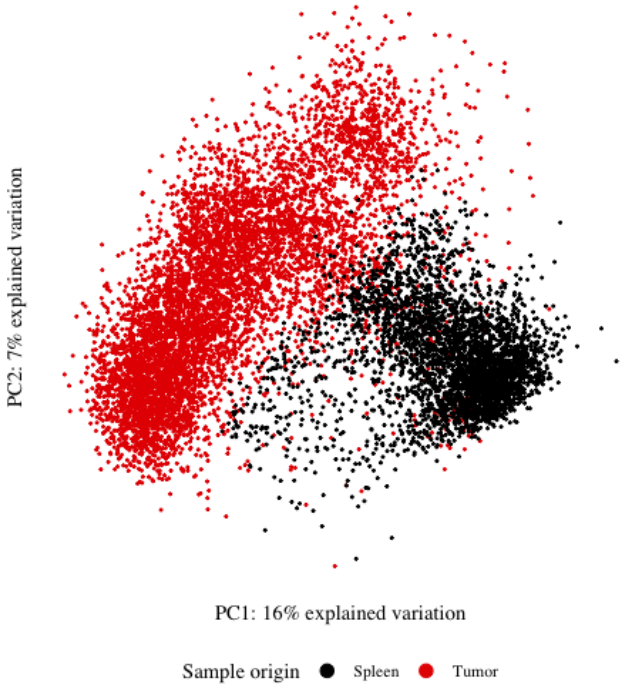


Fig. 4.40 PCA projection of *mouse_d15* after normalizing for differences in sequencing depth between the tumor and spleen samples where each dot represents a cell colored by its original sample.

After correction, samples associate more closely than before. However, separation is still apparent. At this point, we can assume that the biology of the tumor and spleen is so drastically different such that the separation between the cells is as clear-cut. This is not a realistic assumption as we expect to see some overlapping cell types, such as $CD4^+$ T helper and $CD8^+$ T suppressor/cytotoxic cells. We do, however, expect differing cell types as well; the T lymphocytes which differentiated into different states within the TME as the tumor took control of their transcriptional mechanism. Hence we have two samples with both distinct and common cell types and states. Furthermore, even when the cell types/states are the same, they can be present in differing abundance due to the proliferation/clonalization of immune cells which takes place in the tumor, resulting in differing cell population compositions.

Batch effect correction with mutual nearest neighbors

To further correct for possible batch effects, I used a method based on the detection of mutual nearest neighbor (MNN) pairs [94]. The MNN approach assumes that, across

the merged samples, there is a shared subset of the cell population, but does not require predefined or identical population compositions. This approach is ideal for our case as we cannot assume cell populations of identical or known compositions between the tumor and spleen samples; however we know they share common cell sub-populations.

Briefly, for each cell i_a in sample A, the MNN-based method finds k cells in sample B that are nearest neighbors to i_a in the high-dimensional gene expression space, thus obtaining its k nearest neighbors in sample B. Then for each cell in sample B, it finds their k nearest neighbors in sample A. Cells from different samples that are within each other's set of nearest neighbors are pairs of cells that are *mutually* nearest neighbors. These pairs of cells are thought to belong to the same biological cell type/state before any batch effect corrupted their signals. Hence, the difference in expression profiles between these paired cells should give an estimate of the batch effect which needs to be removed. This difference is then subtracted from all cells, including those that were not part of an MNN. Hence the batch effect correction is applied to both shared and distinct cell types.

The `batchelor::fastMNN()` function implements the MNN approach with a slight alteration: To decrease computational work, it performs a PCA to obtain a low-dimensional representation of the input data. All cells are projected into this low-dimensional space defined by the top d chosen principal components, and batch effect correction is performed in this space.

I applied `batchelor::fastMNN()` on `d=50` principal components computed from the shared HVGs. Here the shared HVGs are genes with a positive average biological component across the two samples. In total there were 11,164 HVGs. I excluded all the TCR generating V, D, J, and C genes. Figure 4.41 shows how the cells separate by the sample of origin after MNN correction.

We can now see a further association between the cells coming from the two different samples. To interpret the separation, we can also look at the corrected expression of T cell marker genes CD3, CD4, CD8, and FOXP3. In Figure 4.42 we see that T helpers, CD8 T cells, and Tregs have grouped tightly suggesting that we were able to separate the cells irrespective of their sample of origin, or less so.

The MNN-corrected values no longer make sense in terms of gene expression: As seen in Figure 4.42, they can take negative values. This is expected since we are subtracting vectors to correct for batch effects. However, we can still use these corrected values for cell-cell analysis where we only need the Euclidean distances between cells.

I used the MNN-corrected values to build a shared nearest-neighbors graph for all cells and applied the Walktrap community detection algorithm to identify the possible

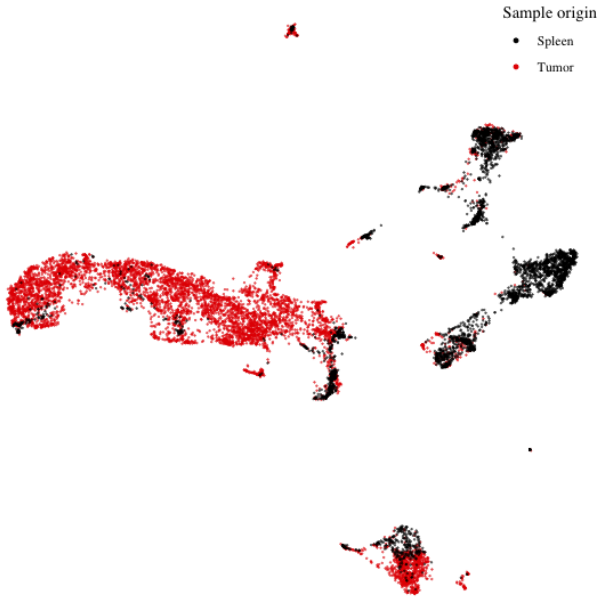


Fig. 4.41 UMAP of *mouse_d15* after MNN correction where each dot represents a cell colored by its original sample.

clusters. Figure 4.43a shows the detected clusters overlaid on the two-dimensional UMAP of the merged samples.

In Figure 4.43 we see that clusters 6, 12, 13, 14, and 15 have cells that do not bear TCRs. Additionally, cluster 4 in majority comprises spleen sample cells (1372 spleen and 38 tumor sample cells). To identify these clusters, I performed differential expression analysis between clusters, but within each sample using the uncorrected gene expression data, then combined the p-values across samples to find the top cluster markers. Figure B.25 is a heatmap showing the uncorrected scaled gene expression of the cluster biomarkers. Cluster 1 shows high expression of TUBA1B, TOP2A, and STMN1, indicating it is a cell cycling cluster. Cluster 4 comprises naive $CD4^+$ cells.

I next overlaid the shared clonotypes on the merged samples selecting only the top 15 most frequent ones (see Figure 4.44). There are 912 tumor cells and 136 spleen cells that bear the *common_clonotype_1* TCRs. This clonotype spreads over clusters 1, 2, 3, and 11. Considering cluster 1 is cell cycling, I focused on the remaining clusters.

I took the clonotype as the unit of identity and found the markers that distinguish between tumor and spleen to see how this clonotype differentiates between the two samples. There are a total of 776 cells in this subset consisting of 104 spleen and 672 tumor cells.

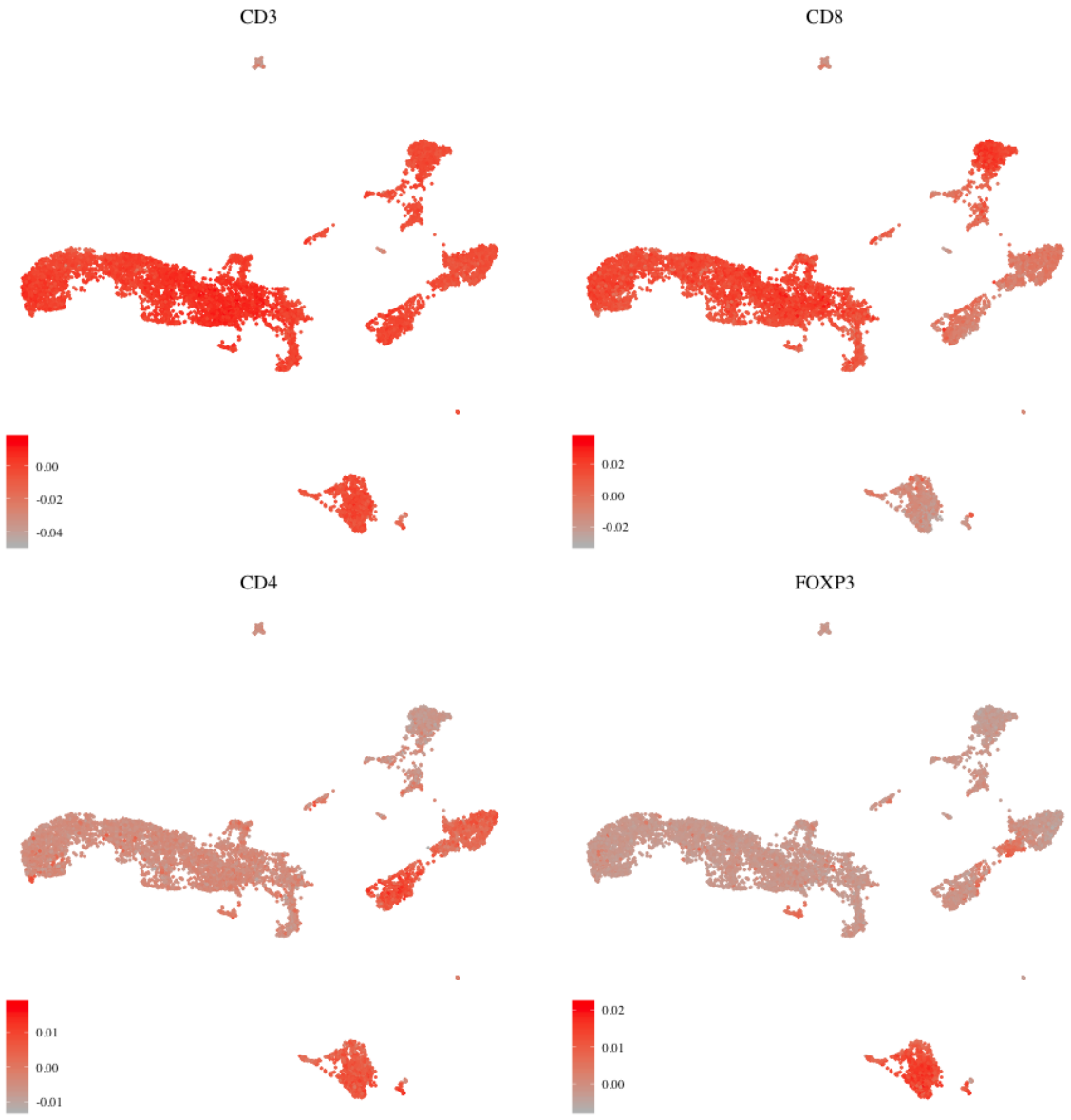
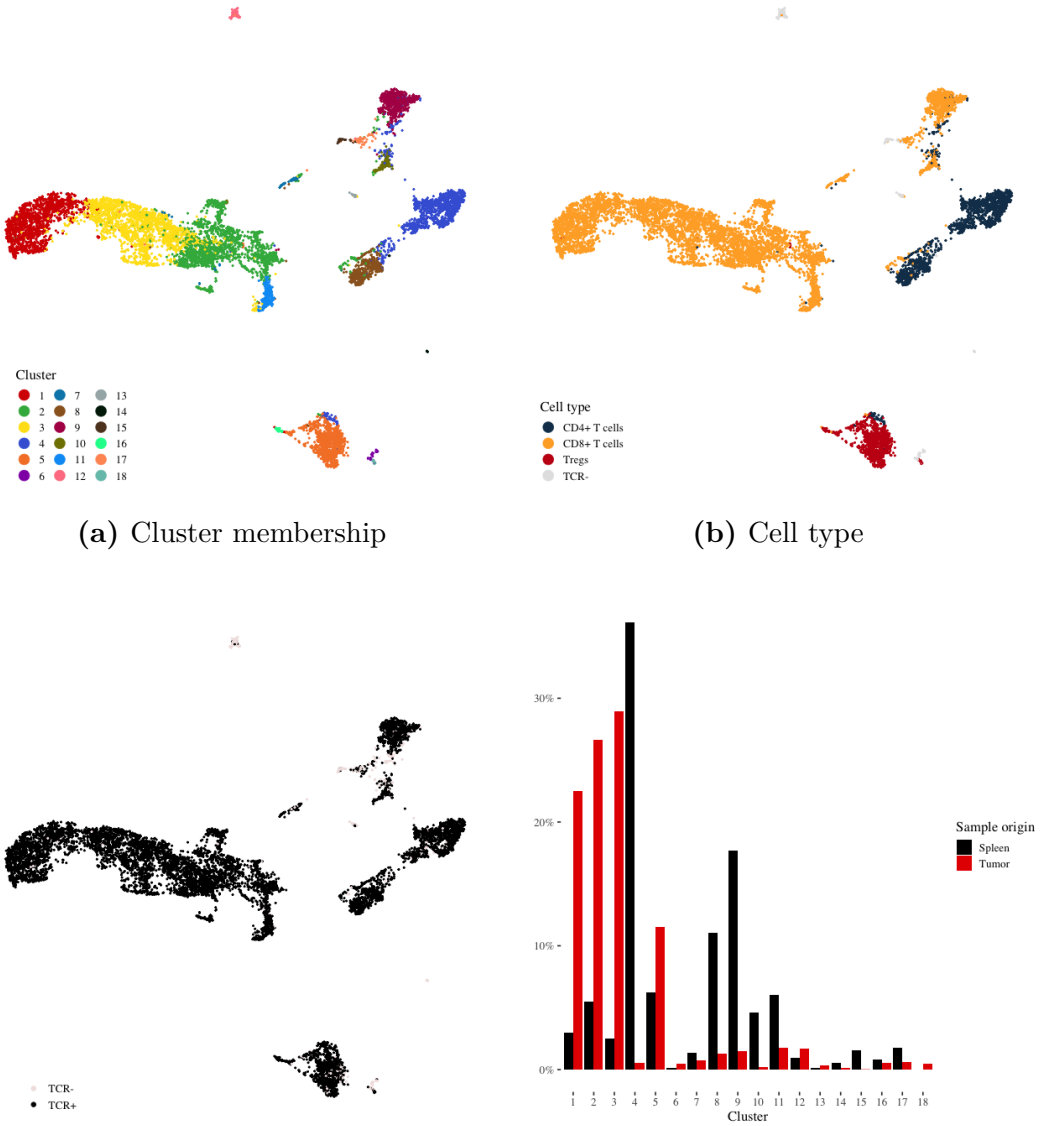


Fig. 4.42 UMAP of *mouse_d15* after MNN correction where each dot represents a cell colored by the corrected expression of CD3, CD4, CD8, and FOXP3 genes.



(a) Cluster membership **(b)** Cell type

(c) TCR presence **(d)** Cluster proportions by sample

Fig. 4.43 (a-c) UMAPs of the merged *mouse_d15* sample after MNN correction where each dot represents a cell and is colored by cluster membership, TCR presence, and cell type. (d) Proportions of cells, relative to within samples, in each cluster separated by sample of origin.

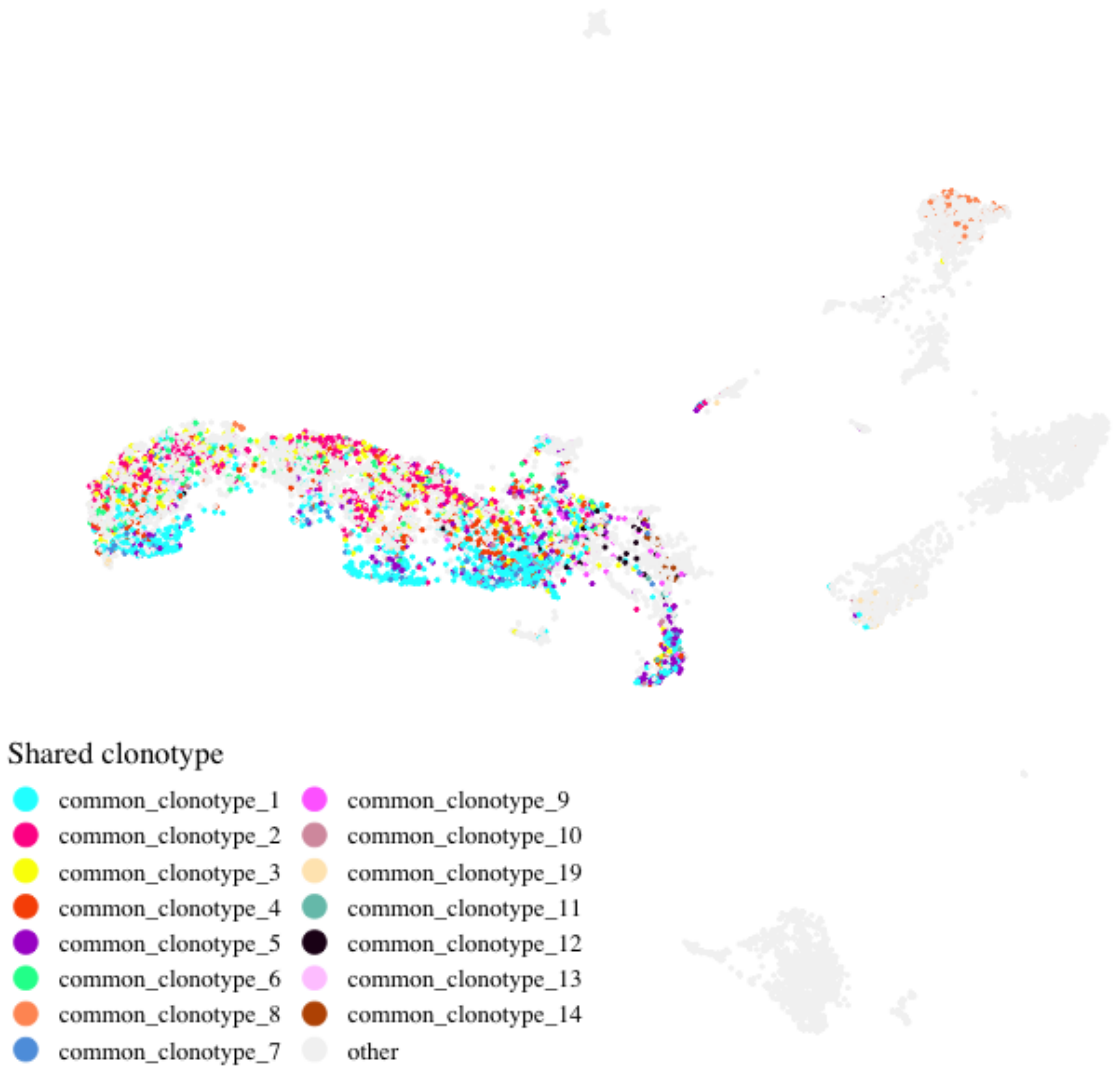


Fig. 4.44 UMAP of *mouse_d15* after MNN correction where each dot represents a cell colored by the shared clonotypes between the tumor and spleen samples.

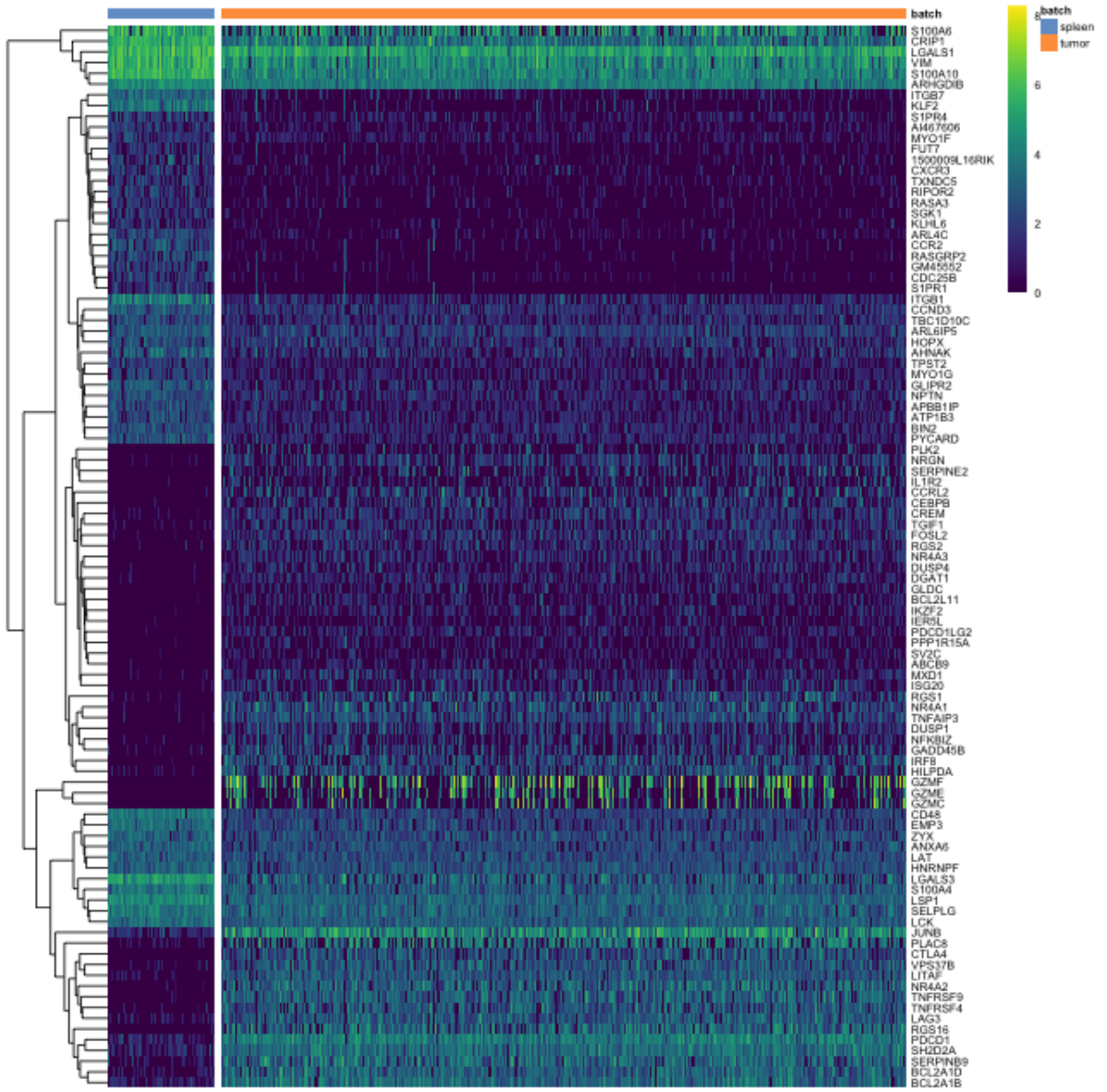
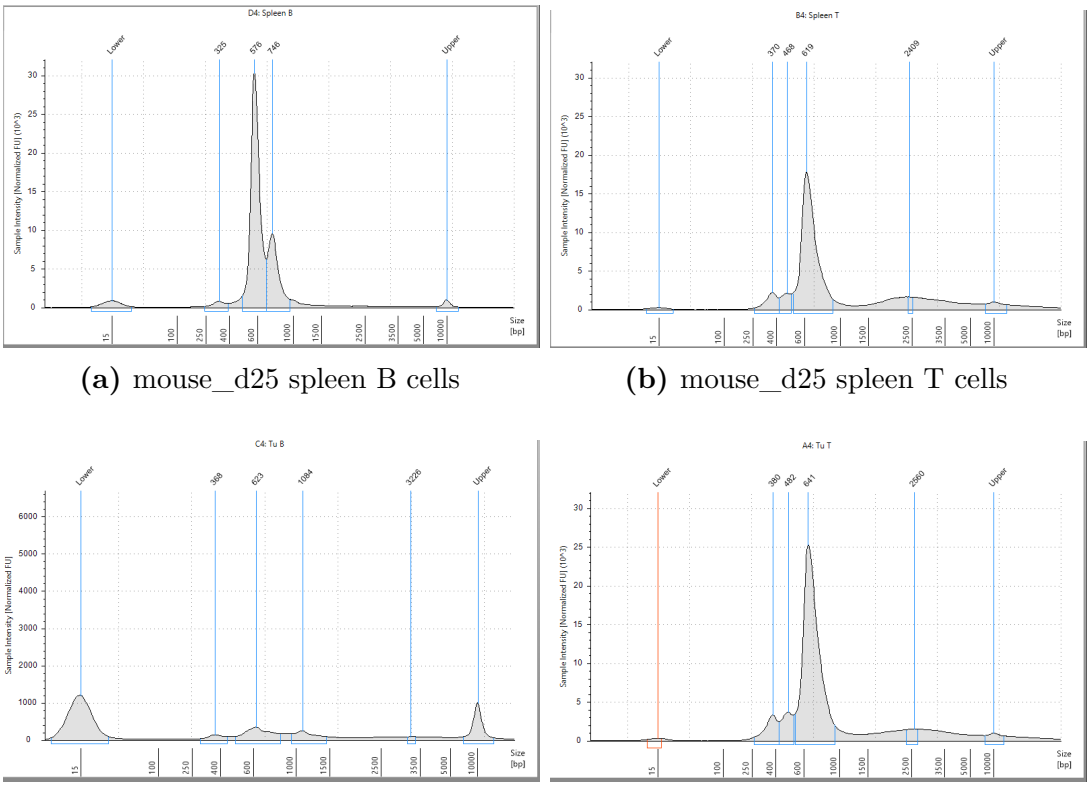


Fig. 4.45 Heatmap showing *mouse_d15 common_clonotype_1* tumor and spleen cells' uncorrected gene expression of biomarkers found to differentiate the tumor and spleen cells. The gene expression is not scaled.

Looking at Figure 4.45 we can see the TME-immune response interplay in action. The *common_clonotype_1* cells differentiate in the tumor expressing granzymes (GZMC, GZME, GZMF), meaning these cytotoxic T clones are inducing apoptosis to eliminate cancerous cells. We also see the differential expression of PDCD1, CTLA-4, and LAG3 in the tumor sample clones, which shows that the TME is trying to counteract the immune response by putting the clones in a dysfunctional state.



(a) mouse_d25 spleen B cells **(b)** mouse_d25 spleen T cells

(c) mouse_d25 tumor B cells **(d)** mouse_d25 tumor T cells

Fig. 4.46 cDNA QC plots show very low yield for BCR enrichment. Plots courtesy of Ms. Katarzyna Kania, CRUK CI Genomics Core

4.5 Results

4.5.1 Lack of intratumoral B cells

BCR enrichment was performed along with TCR enrichment only for *mouse_d25* reason being that no B cells were detected in the tumor sample. The QC results of the cDNA manufactured from the tumor sample showed a very low yield for BCR enrichment (Figure 4.46). This was consistent with a follow-up FACS analysis: Remaining tissue from FACS were stained with CD19 to check for B cells. However there were no $CD19^{+}EYFP^{+}$ expressing cells detected (Figure 4.47).

Looking at Table 4.8, we see that 1,060 B cells have been detected in the tumor where only 11 of them have a productive V(D)J pair. Only 54.9% of the reads have mapped to any VDJ gene, and very low portions have mapped to Ig genes. The fraction of reads is very low as well, indicating ambient RNA and hence dead cells or artifacts. The first obvious explanation of why we see 1,060 cells, but only 11 of them have productive Ig chains is that these barcodes have been falsely called as cells. Next, as very few reads

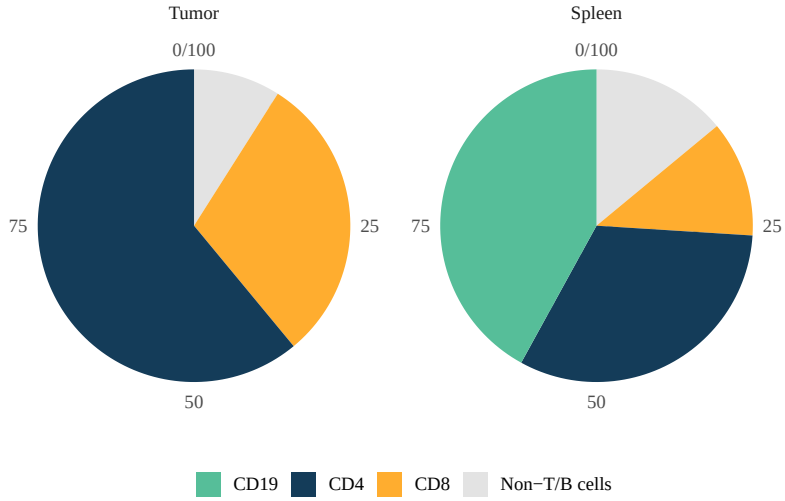


Fig. 4.47 Composition of EYFP^{+} sorted cells. Non-T/B cells are defined as CD45^{+}CD11B^{-}F4/80^{-}CD8^{-}CD4^{-}CD19^{-}

	mouse_d25 spleen	mouse_d25 tumor
Estimated Number of Cells	1,579	1,060
Mean Read Pairs per Cell	16,407	35,611
Number of Cells With Productive V-J Spanning Pair	1,505	11
Number of Read Pairs	25,908,141	37,748,147
Valid Barcodes	97.1%	87.5%
Reads Mapped to Any V(D)J Gene	97.3%	54.9%
Reads Mapped to IGH	17.1%	17.6%
Reads Mapped to IGK	65.5%	14.3%
Reads Mapped to IGL	11.3%	0.3%
Cell Count Confidence	77.7%	58.1%
Mean Used Read Pairs per Cell	9,922	3,302
Fraction Reads in Cells	73.4%	34.4%
Median IGH UMIs per Cell	18	0
Median IGK UMIs per Cell	67	0
Median IGL UMIs per Cell	240	0
Cells With Productive V-J Spanning Pair	95.3%	1.0%
Cells With Productive V-J Spanning (IGK, IGH) Pair	90.1%	0.8%
Cells With Productive V-J Spanning (IGL, IGH) Pair	6.6%	0.2%
Paired Clonotype Diversity	27.48	11.00
Cells With IGH Contig	97.3%	18.0%
Cells With IGK Contig	95.8%	29.6%
Cells With IGL Contig	42.1%	1.1%
Cells With CDR3-annotated IGH Contig	96.2%	1.0%
Cells With CDR3-annotated IGK Contig	94.6%	0.8%
Cells With CDR3-annotated IGL Contig	6.8%	0.2%
Cells With V-J Spanning IGH Contig	95.8%	1.0%
Cells With V-J Spanning IGK Contig	95.5%	0.8%
Cells With V-J Spanning IGL Contig	10.8%	0.2%
Cells With Productive IGH Contig	95.5%	1.0%
Cells With Productive IGK Contig	94.6%	0.8%
Cells With Productive IGL Contig	6.8%	0.2%
Contigs Unannotated	0.2%	0.3%

Table 4.8 BCR enrichment analysis metrics of *mouse_d25* samples

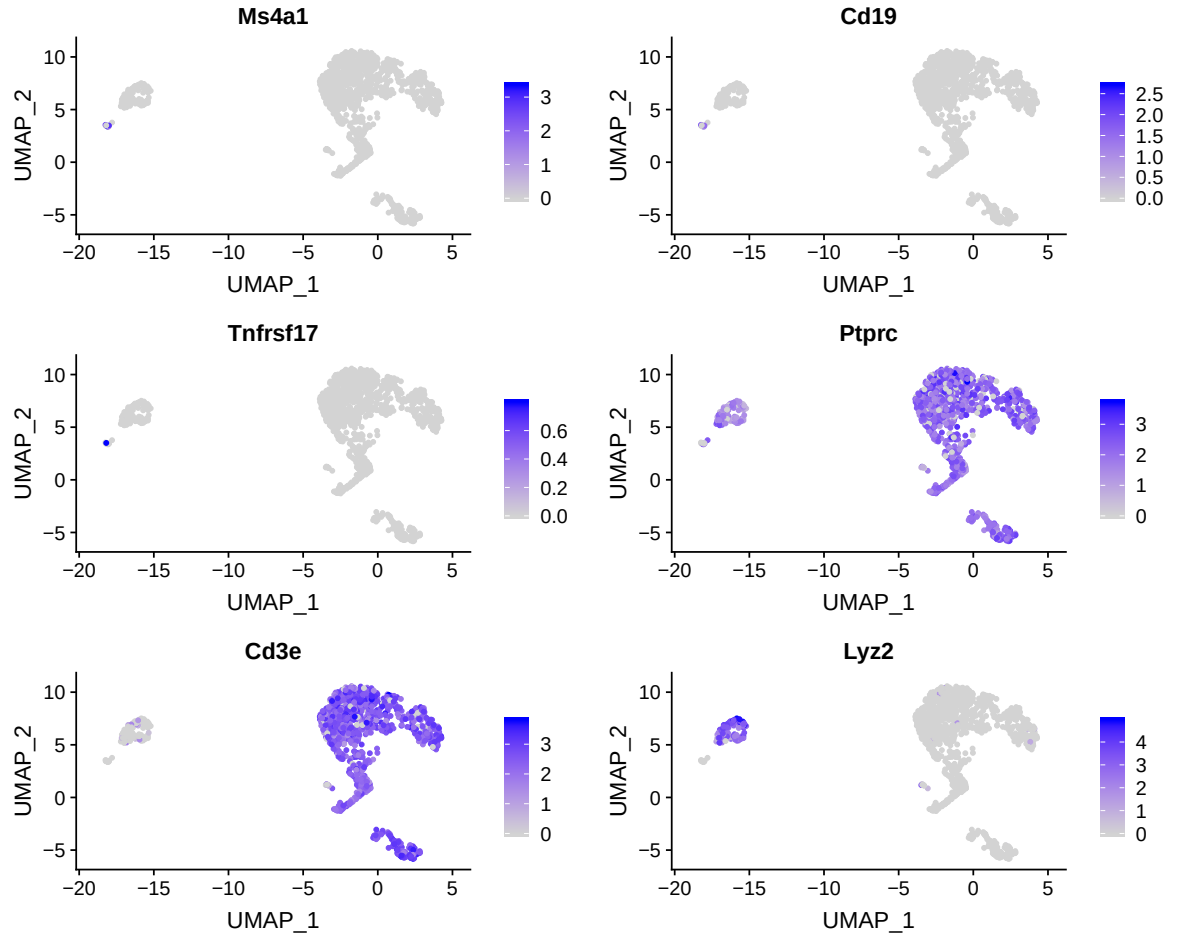


Fig. 4.48 UMAP of the 1,060 cells showing the gene expression for MS4A1, CD19, TNFRSF17, PTPRC, CD3, and LYZ2

map to Ig genes, it is possible that some of the cells are instead T cells. In order to further investigate, I analyzed these cells' gene expression profiles.

Figure 4.48 visualizes feature expression of selected genes on a UMAP of the 1,060 cells. Looking at the expression of MS4A1, CD19, and TNFRSF17, we can see that there are only a few B cells or plasma B cells. Majority of the cells express CD3 suggesting if these are real cells, they are in fact T cells. Figure 4.49 shows the expression heatmap of the top markers for each cluster. Only cluster 6 shows B cell markers. So if we are to accept these barcodes as cells, it is evident that they are in majority T cells.

Having investigated the nature of the aforementioned 1,060 cells, we need to address the lack of B cells in our samples. There are multiple possible explanations for this observation. One possible explanation is that during B cell differentiation CD19 might be downregulated. Our B cells might represent activated B cells of a different developmental

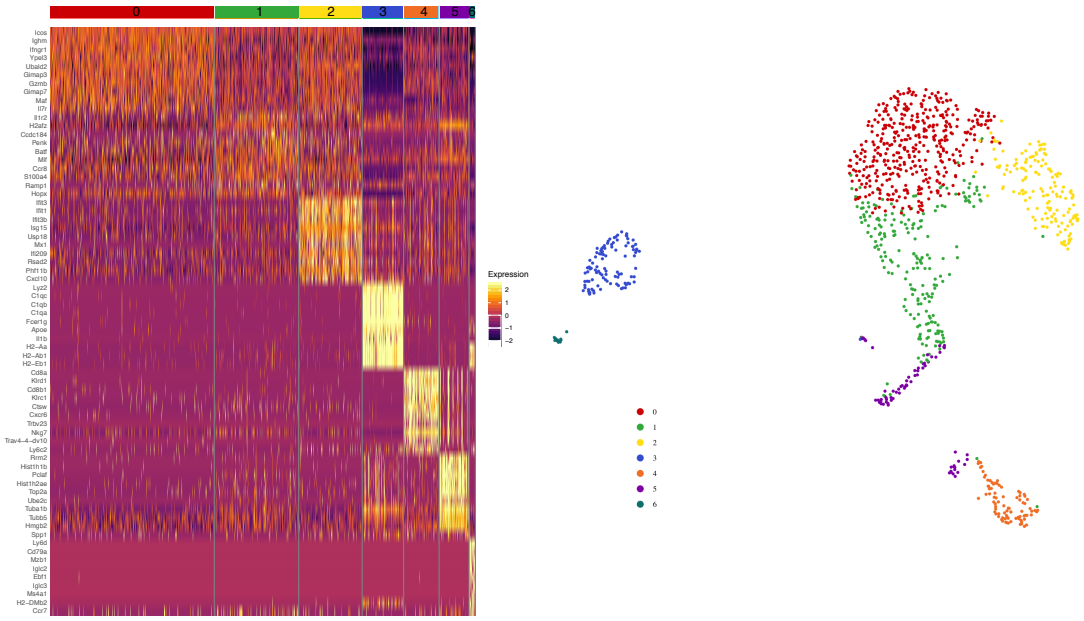


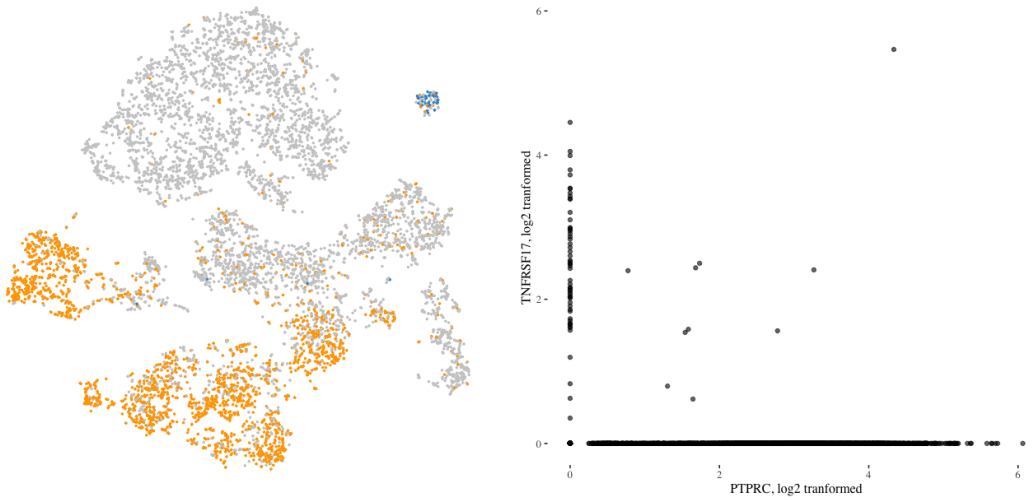
Fig. 4.49 Majority of the markers determine that the cells are T cells. Only cluster 6 expresses B cell biomarkers MZB1, IGLC3, and MS4A1.

stage, e.g., plasma cells, which were not recovered by the CD19 FACS analysis. Moreover, the AgRSR mouse cells are sorted for $CD45^{+}CD11B^{-}EYFP^{+}$. Indeed CD45 is expected to be low in plasma cells and plasmablasts [183], thus impairing the recovery of these cell populations in our experimental system. This explanation would be consistent with what we observed in the human melanoma samples presented in Chapter 2. In these unselected cell populations, the plasma cell cluster does not express PTPRC, the gene which encodes for the CD45 molecule, or expresses it at very low levels (see Figure 4.50). While we have not directly observed a presence of plasma cells in our mouse single cell sequencing data, we cannot exclude their presence as they might have been counterselected by our $CD45^{+}$ FACS analysis.

Additionally, to further account for the absence of B cells, I can hypothesize that considering this melanoma is an injected tumor, rather than one that has developed naturally over time, it is possible that the immune system of this mouse model does not portray one that developed naturally over time, either. Injected tumors form very quickly, showing accelerated disease progression, and as a result contain less stroma and are rich in cancer cells [225], which may not reproduce the histological nature of a clinical cancer. A recent study has compared B cell infiltration and activation in a genetic model of murine PDAC (KPC mouse) as well as an injectable orthotopic model, to find that significant B cell infiltration was only observed in KPC tumors and correlated with T

cell infiltration, while orthotopic tumors showed a low infiltration rate of B cells [225]. Furthermore, KPC-derived B cells expressed GC entry, B cell memory, and plasma cell differentiation markers accompanying significant intratumoral Ig deposition, a feature detected less in orthotopic tumors.

I can therefore speculate that the absence of B cells is due to technical choices, cell sorting and injection of syngeneic tumor cells, rather than a reflection of the underlying true immunological mechanism. Neither of the two other mouse samples was enriched for BCRs, hence, in this study, I did not further examine B cells.

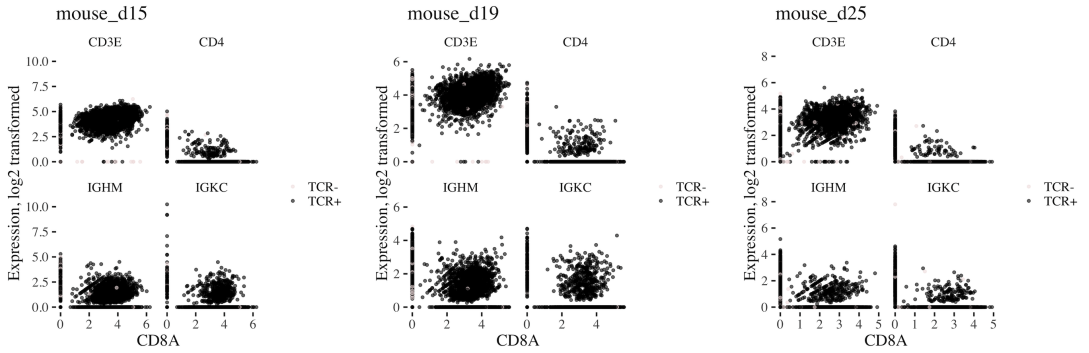


(a) t-SNE projection **(b)** Feature plot

Fig. 4.50 Expression of PTPRC and TNFRSF17 in the *patient*1 sample presented in the Chapter 2. Plots show that cells which express TNFRSF17 do not express PTPRC. (a) Blue dots represent cells expressing TNFRSF17, orange dots PTPRC. (b) Correlating TNFRSF17 and PTPRC expression. Gene expression values are normalized and log2-transformed.

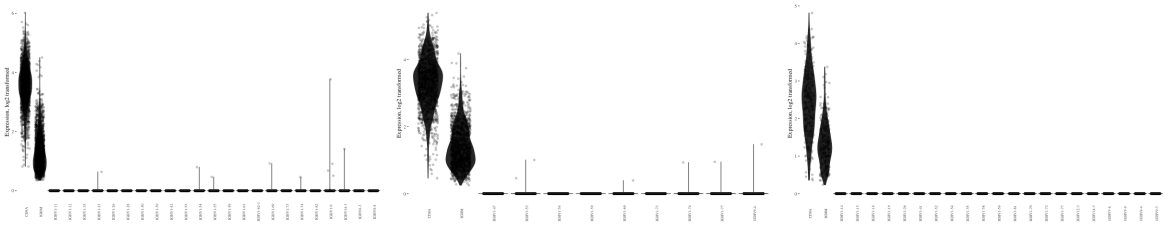
4.5.2 Co-expression of CD8 with CD4 and B cell genes IGHM and IGKC

In all of our samples taken from the tumor of the AgRSR mice, I found cells displaying co-expression of the surface markers CD8 and CD4. During positive selection which takes place in the thymus, double-positive thymocytes differentiate into $CD8^+$ or $CD4^+$ single positive cells depending on whether they recognize the MHC class I or II molecule. Therefore we should not be detecting DP $CD8^+CD4^+$ T cells in the tumor samples. However $CD8^+CD4^+$ DP mature T cells have been reported in melanoma [55]. Additionally, we see TCR bearing cells co-expressing IGHM/IGKC and CD8 even though these



(a) *mouse_d15* **(b)** *mouse_d19* **(c)** *mouse_d25*

Fig. 4.51 The correlation between CD8 and CD4, IGHM, and IGKC. Note that these are normalized log2-transformed gene expression levels.



(a) *mouse_d15* **(b)** *mouse_d19* **(c)** *mouse_d25*

Fig. 4.52 IGHV gene expressions of $CD8^+$ cells. Expression levels are normalized and log2-transformed.

genes are specific to B and T cells, respectively. See Figure 4.51 for the expression of these genes. We could be observing a case of dual receptor-expressing lymphocytes as presented in [7]. However, we did not perform BCR specific sequencing for these samples, so we cannot determine whether these cells also express BCRs. The BCR recombinant V, D, and J gene expression levels are too low to be picked up directly from the GEX analysis. Figure 4.52 shows the co-expression of IGHM and CD8 along with BCR V genes. I could not observe significant expression of Ig V segments. For this reason, it is impossible to draw meaningful conclusion from this data regarding the nature of this group of cells. Validation experiments would be required to investigate this population.

4.5.3 Dysfunctionality signatures

In all three tumor samples taken from the AgRSR mice, I detected different $CD8^+$ T cell clusters that correlate with different functional states. All tumor samples had cells in the dysfunctional state, which was determined by the expression of PDCD1, LAG3, TOX, HAVCR2, and CTLA-4. When I took the clonotype as the unit of identity, I was

able to identify clonal dysfunctionality. It was most evident in *mouse_d25* which was the mouse kept alive the longest. Here I found clonotypes which differentially expressed TOX, LAG3, and PDCD1, indicating dysfunctional clonotypes.

Furthermore, I detected the clonal expression of genes such as TCF7 and IL7R, which have been shown to predict clinical response to immune checkpoint immunotherapy in melanoma [197]. Another clonotype showed high expressions of GZMA and GZMK, as well as high expressions of TIGIT, TOX, and EOMES. Similar exhaustion populations have been described in chronic viral infections [173]. Considering we have different clonotypes exhibiting dysfunction/exhaustion, it would be interesting to study further what these different TCRs are binding.

4.5.4 The immune response is specific to each host

I compared all TCR chain clonotypes that were found in the tumor samples of each mouse to see if they had any common clonotypes.

The two clonotypes that were common in all three tumor samples were single TRA chains and present at rather small proportions: neither chain was observed in more than 0.13% of the cells in any sample (Table 4.9). Considering that these were not paired TRA-TRB chains and that they were represented with a low, and non-increasing frequency of cells, it is likely that these clonotypes were not particularly active in the ongoing intratumoral immune response.

Clonotype CDR3	mouse_d15	mouse_d19	mouse_d25
TRA:CAASASSGSWQLIF	0.00073	0.00042	0.00130
TRA:CAAEANYGNEKITF	0.00015	0.00021	0.00017

Table 4.9 Shared TCR clonotypes across all tumor samples.

I also compared the TCR chains found in the cells taken from the mouse spleens and found no shared chain.

Our results do not show any evidence of shared immune response to melanoma in the AgRSR mice. This may suggest that individuals subjected to similar antigenic challenges might develop completely different TCR repertoires. This is even more striking if we consider that the mice used in our experiments are almost genetically identical.

4.5.5 The immune response is not localized

In Section 4.4.3 I showed that the tumor and spleen samples share multiple clonotypes. Looking at their GEX profiles, we see that their clones are of different cell types. In *mouse_d15* and *mouse_d19* the common clonotypes with the most cells in either the spleen or the tumor samples are $CD8^+$. However, in *mouse_d25*, the common clones are Tregs expressing CD4 and FOXP3. With this dataset, we were able to show the clonal expansion of Tregs. Additionally, in *mouse_d25* the $CD8^+$ dysfunctional T cells were not represented in the spleen. However, we do see common clonotypes expressing dysfunctionality signatures within the two other mouse samples.

4.5.6 Possible evidence of allelic inclusion

In all of the tumor samples combined, I detected one α and one β chain pairing in 74% (9,171/12,454) of cells, two α and two β chain recombinants in 1.3% (156/12,454) of cells, and solely two α chain recombinants in 0.56% (70/12,454). 510 cells had only one α chain, and 246 had only one β. I detected two α and one β chain recombinants in 1666 cells out of 12,454 (13%) of cells, and one α and two β chain recombinants in 635 (5.1%) cells. I considered only the clonotypes represented by at least 10 cells. All chains are productive. Even though it is generally accepted that TCRs comprise one α and one β chain, these observations are in line with the phenomenon of allelic inclusion described in the literature [227], [30].

A more interesting case can be observed in the *mouse_d25* tumor sample dataset. Here we have three clonotypes, *clonotype5*, *clonotype31*, and *clonotype44* which clustered tightly based on their GEX profile (see Figure 4.53). Clonotype5 is represented by 246 cells which bear two α and one β chains: TRAV14-2, TRAV6-3, and TRBV13-2. Clonotype31 has 30 cells with the chains TRAV6-3 and TRBV13-2, and clonotype44 has 14 cells which bear the chains TRAV14-2 and TRBV13-2. It is possible to see cells with one α and one β chain, with additional captures of minor transcripts resulting in multiple chains. In such cases, we would see one α chain with a high UMI count, one β chain with a high UMI count, and all the other chains would have low UMI counts. However, looking at Figure 4.54, it is clear that both α chains of clonotype5 have high UMI counts. So here clonotype5 is expressing the same α and β chains as clonotype31 and clonotype44 and is expressing an α chain additional to the other two clonotypes. Clonotype5 is more clonally expanded than both clonotype31 and clonotype44. Also, observe that only clonotype5 has cell cycling clones (see the bottom panel of Figure 4.26c which plots the cell cycling cluster). Although the simplest explanation could be that

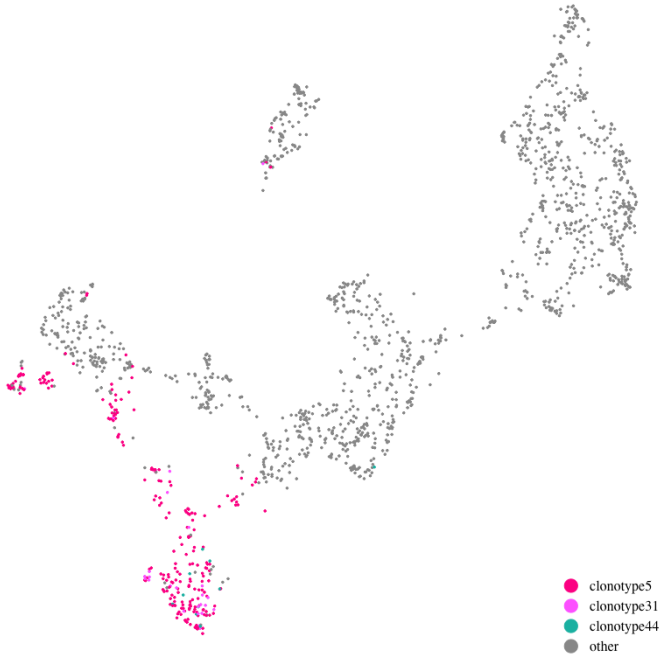


Fig. 4.53 UMAP of the *mouse_d25* tumor sample CD8$^+$ cells colored by clonotype membership.

all these clones, in fact, express the same double-α single-β chains on their surface but `cellranger vdj` could not detect the additional α chains, it is possible that there is a descendant relationship between clonotype5 and the other clonotypes. Additional experiments would be required to investigate this further.

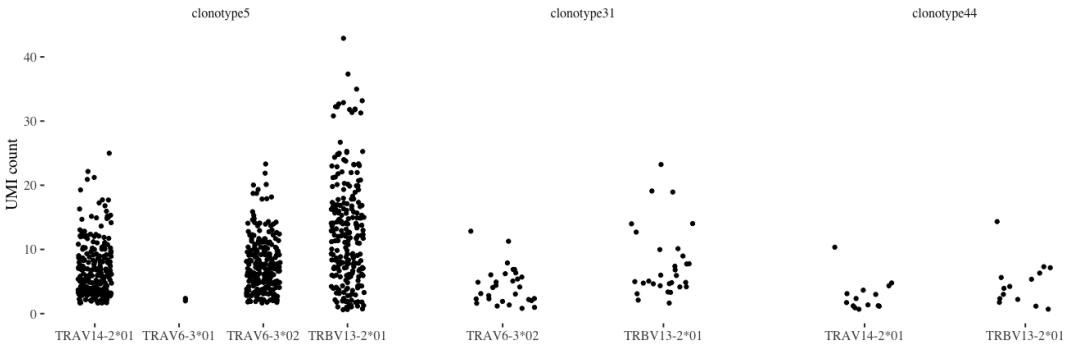


Fig. 4.54 The per cell UMI counts of the *mouse_d25* tumor sample clonotypes. These are the UMIs that were picked up with TCR enrichment.

4.6 Discussion

Tumor infiltration by T cells is considered a proxy to immune response and has been shown to reflect better prognosis in solid tumors [206], [163], [82], [264]. Currently, immunotherapy studies are investigating ways to maintain this immune response, which can, in time, be suppressed by the TME. Reactivating dysfunctional T cells can, in some cases, lead to a sustained immune response [189]. Several recent studies have coupled scRNA-seq and TCR repertoire information to characterize the tumor infiltration in melanoma [235], [197], [140]. However, they provide limited insights as the TCR repertoires they investigate are not particularly tumor-antigen specific. Recent work provides evidence that some intratumoral T cells may be bystanders with no tumor reactivity [219]. To detect T cell clones that are truly able to recognize cancer antigens, we need to be able to identify the repertoire of tumor-reactive TCRs. This is of crucial importance as developing new methods to overcome immune escape in cancer requires a deeper understanding of cancer antigens and TCRs that actually drive the intratumoral immune response.

In this study, I defined the clonal tumor infiltration of T cells in an unbiased manner by linking single-cell RNA sequencing and TCR repertoire datasets which were obtained by using a novel transgenic mouse model that allows for the in vivo identification of activated T cells. In the cells of this animal, the injection of tamoxifen initiates the expression of yellow fluorescent protein (EYFP) in TCR-activated clones. As a DNA mediated event, the expression is inheritable, and the clonally expanded cells remain $EYFP^+$.

In collaboration with Dr. James Thaventhiran and Mr. Ty So, by performing FACS and immune profiling of more than 30,000 $EYFP^+$ cells present in the spleen and tumor, we were able to identify tumor responding Treg, conventional $CD4^+$, and $CD8^+$ T cell clones. Our findings show that tumoral $CD8^+$ T cells comprise similar populations found in human melanoma [140], [197]. By analyzing the GEX profile of tumor-reactive $CD8^+$ T cells, I distinguished different clusters that correlate with different functional states. By coupling the GEX profile with cell level TCR repertoire information, and taking the clonotype as the unit of identity, I detected T cell clones which exhibit different functional states. I was able to detect clonotypes with markers of dysfunctionality such as PDCD1 (PD1 in humans), CTLA-4, TOX, EOMES, and LAG3.

With this mouse model, we fate mapped a clonally diverse population of T cells that are TCR-activated towards melanoma tumors within an experimentally defined time frame. The cells of the tumor sample taken from the mouse kept alive the longest showed a decrease in the $CD8^+$ T cell population. Furthermore, the $CD8^+$ T cell clonotypes

of this mouse showed expression of exhaustion and apparent clustering. Unless in a cell cycling state, dysfunctional $CD8^+$ T cells exclusively cluster separately indicating a convergent differentiation state which may suggest that although multiple clonotypes are proliferating within the tumor, their intratumoral differentiation state is dictated by their TCR. This separation was not as evident in the two other earlier sacrificed mice. For the immune response, a week might be a significant amount of time. By analyzing the GEX profiles of mice cells harvested at different time points, I was also able to suggest a preferable time of harvest in order to obtain and further study cells exhibiting clear dysfunctionality markers. This would enable to improve the design of further experiments.

With this study, we were able to match clonally expanded TCRs of the tumor with those of the spleen. A majority of expanded T cell clones were shown to have clonally-matched cells present in the secondary lymphoid tissues. Also, by analyzing the GEX profile of the matching clonotypes, I was able to begin elucidating the gene expression activity of different intratumoral T cell clones, and the interplay between local and systemic response to cancer.

I was not able to study the clonality of intratumoral antigen-reactive plasma cells. However, I could speculate that their absence in the tumor is due to cell sorting for $CD45^+$ which is expected to be low in plasma cells and plasmablasts [183]. Furthermore, the injection of syngeneic tumor cells could have caused accelerated disease progression which may have prevented the development of a natural immune response which would have been expected in a tumor that occurred spontaneously [225].

I am aware that my findings are not conclusive since we only had single samples with differing conditions taken at different time points. For this reason, it is necessary to conduct further experiments using multiple samples, i.e., technical and biological replicates while controlling for possible batch effects.

This study forms a first proof-of-principle effort for the comprehensive analysis of combined scRNA-seq and TCR profiling of tumor-responsive T cells in-vivo in melanomas. By performing larger scale experiments, it will now be possible to obtain better insights into tumor antigen binding T cells, their receptors, and their functional status in the TME.

Chapter 5

Conclusions and future study

William B. Coley, a 19th-century surgeon, may have performed the first application of immunotherapy in cancer when he administered patients with "Coley's mixed bacterial toxins" to treat their inoperable malignant tumors [49]. The field of anti-tumor immunity has since seen numerous milestones: Paul Ehrlich's cancer immunosurveillance hypothesis [61], Ruth and John Graham's cancer vaccine study of a 114 patient cohort [85], evidence of tumor-specific antigens [121], the identification of immune cells including that of B and T lymphocytes and their antigen specificity through receptor diversity [113], and Dunn and Schreiber's cancer immunoediting concept which replaced the immune surveillance hypothesis [58], are amongst the numerous discoveries in the field. Most recently, the independent discoveries of the checkpoint proteins CTLA-4 and PD-1 by James P. Allison and Tasuku Honjo brought them the 2018 Nobel Prize in Physiology or Medicine, and has led to tremendous scientific and clinical breakthroughs in anti-tumor immunity. While ICB therapy has shown great success, its effects are not shared universally, and the results are often not long-lasting [189], [52], [266]. It is necessary to associate the mechanisms that limit durable positive response with ICB therapy in order to transform immunotherapy into a more broadly applicable cancer treatment.

Through numerous genetic and epigenetic alterations, tumors might accumulate neoantigens that can be recognized by the immune system [229]. The anti-tumor immune response involves the innate and adaptive immune components which aim to control tumor growth [153]. While a baseline immune response is well observed, it is not always effective as tumors develop ways of resistance through immune suppression, tolerance, and escape [205], [159]. With the studies on TILs, as well as TLSs, and their associations with patient survival [26], [201], it has become clear that the TME needs to be analyzed in-depth to expose the mechanisms that tumors adopt in order to counter an immune response, and improve the benefits of ICB therapies.

In Chapter 2, I characterized the TME of melanoma metastases. As expected, I found significant infiltration of T lymphocytes, including Tc cells, Th cells, and Tregs. The role of the effector T cells in anti-tumor immunity is well known, and the pre-existing mass of antigen-experienced TILs in the TME has been associated with effective ICB therapies [242], [214], [6], [26]. The role of B cells, on the other hand, remain uncertain [241], [254]. By analyzing gene expression in 12,000+ intratumoral cells at single-cell level, I confirmed the presence of TIL-Bs in melanomas, including $CD19^+$, $CD20^+$, and plasma B cells, as well as proliferating B cells at various differentiation stages. With known cell type biomarkers, I identified the presence of LTi and Tfh cells, and FRCs, which indicated the formation of intratumoral TLS. The presence of TLS and the abundance of TIL-Bs and TILs suggest a combined effort of B and T lymphocytes in generating an anti-tumor response.

Furthermore, by coupling gene expression profiles with BCR immune repertoires, I was able to detect clonally expanded plasma B cells. The fact that these plasma cells are expressing class-switched antibodies suggests prolonged exposure to the target antigen. The likely presence of TLS and clonally expanded plasma B cells expressing class-switched antibodies at baseline may suggest an ongoing adaptive intratumoral immune response.

Having observed abundant plasma cells and highly clonal antibodies pre-immunotherapy in melanoma metastases, I wanted to study B cell infiltration of tumors pre- and post-therapy with AMD3100, and assess the difference in clonality in the receptor repertoires of intratumoral lymphocytes. AMD3100 has been shown to re-accumulate effector T cells amongst the cancer cells by blocking the CXCL12-CXCR4 axis [65]. I was interested in the effects of this treatment on intratumoral B cells, and whether or not the cooperation of B and T cells would be visible here as well. To this end, as presented in Chapter 3, I reconstructed the B and T cell receptor repertoires of 37 samples taken before and after treatment with AMD3100 from patients with CRC and PDAC. In all of the samples, the number of unique BCR clonotypes was significantly higher than TCR clonotypes, showing intratumoral antibody-secreting plasma cells. IgG and IgA were the most abundant antibody isotypes, further showing that the majority of the detected Ig clonotypes were expressed by B cells that have encountered an antigen and differentiated into plasma and memory B cells. These findings strongly suggest B cell infiltration, and the presence of plasma cells inside the CRC and PDAC tumors I studied.

Moreover, in some patients, the B cell infiltration increased after treatment, which may be a result of the AMD3100 treatment. In a recent study, patients who did not show a positive response to PD-1 inhibitor therapy had a greater increase in clonotype diversity among B and T cells [266]. Based on these findings, I decided to identify patients with

highly clonal BCRs. Although these patients were not administered with anti-PD-1 therapy, they may already have an ongoing anti-tumor immune response and may be good candidates for an ICB therapy.

In both studies, I was able to show an intratumoral immune response involving both B and T lymphocytes. However, we cannot conclude that these cells were necessarily recognizing tumor-specific antigens. While high clonality may suggest tumor-reactive antibodies and effector T cells, alternative explanations are also possible: the proliferation of the actual tumor-reactive T cells may have been hindered by exhaustion [235], and the antibodies may be in response to a non-tumor antigen. In addition, recent work provides evidence of bystander T cells with no tumor reactivity [219]. In Chapter 4, I analyzed 30,000+ $CD45^{+}CD11B^{-}EYFP^{+}$ cells harvested from the tumor and spleen of the AgRSR mouse in order to unambiguously determine which specific B and T cells contribute to anti-tumor immunity. In the cells of this mouse model, the injection of tamoxifen leads to the expression of yellow fluorescent protein (EYFP) in activated B and T cells. As a DNA mediated event, the clonally expanded daughter cells remain $EYFP^{+}$. With this mouse model, we were able to fate map a clonally diverse population of TCR-activated T cells responding to melanoma tumors within an experimentally defined time frame. I identified clonally expanded tumor-reactive Treg, $CD4^{+}$, and $CD8^{+}$ T cell clones and the $CD8^{+}$ T cells comprised similar populations found in human melanoma [140], [197]. By analyzing the GEX profile of tumor-reactive $CD8^{+}$ T cells, I distinguished clusters that correspond to different functional states. Furthermore, by linking the GEX profile with cell level TCR repertoire information, and taking the clonotype as the unit of identity, I detected T cell clones which exhibit different functional states. I was able to detect clones with markers of dysfunctionality.

Additionally, in this study, being able to profile gene expression simultaneously with the TCR repertoire at a single cell level, I was able to match clonally expanded TCRs of the tumor with those of the spleen and partially assess their activation status. This allowed me to investigate the interplay between local and systemic response to cancer. The immunosuppressive effects of the TME were evident in the dysfunctional intratumoral T cell clones. Unfortunately, I was not able to study the clonality of intratumoral antigen-reactive B cells. However, I can speculate that cell sorting for $CD45^{+}$ may have discarded plasma cells and that the injection of syngeneic tumor cells may have caused accelerated disease progression limiting the development of a natural immune response as suggested in [225].

scRNA-seq is an ideal method to study AgRs as it allows for the reconstruction and pairing of receptor chains within a single cell. In Chapter 2, I recovered the AgR

repertoires of melanoma metastases. The *patient2* sample demonstrated a highly clonal Ig repertoire in which the most dominant clonotype showed both class-switching and possible SHMs. This intratumoral antibody may be generating an immune response towards a possible melanoma tumor-antigen. With RNA-seq, it is not possible to pair Ig heavy and light chains with full certainty. However, detecting a highly dominant antibody in the melanoma tumor suggested the opportunity of finding such highly clonal antibody heavy and light chains in the CRC and PDAC samples that where sequenced using RNA-seq. Indeed, in patients 2, 11, and especially 25, I was able to detect heavy and light Ig chains that were clonal enough to pair. All the reconstructed antibody sequences are given in Appendix A. These sequences could allow one to have the specific antibody expressed in mammalian cells. These antibodies could, in turn, be used to find the tumor antigens they bind by using in-vitro screening methods.

In immunotherapy, it is crucial to determine which patients can gain the most benefit from treatments, and what the best biomarkers are to guide treatment decisions. While T cells have been studied in depth to reveal that response to ICB favors pre-existing T cell infiltration in tumors [26], TIL-Bs have only recently incited interest [241], [254]. In Chapter 2 of this thesis, I characterized the TME of melanoma tumors prior to immunotherapy. Using scRNA-seq coupled with targeted AgR enrichment allowed me to determine clonally expanded, antigen-challenged immune components. Immediate future work could be to study all the patients in the MelResist clinical trial [51] using the same technique, with samples taken both pre- and post-immunotherapy. This could lead to finding the effect of baseline TIL-Bs and TLSs, as well as detecting possible B cell related biomarkers towards a positive patient response. Furthermore, we can determine the fate of the clonally expanded antibodies post-immunotherapy and find candidate antibodies for tumor-antigen detection.

In parallel, while in Chapter 3 I had access to pre- and post-treatment data from a larger cohort, it would be beneficial to use a single-cell RNA-seq and targeted AgR enrichment protocol to retrieve the gene expression profiles and VDJ sequences simultaneously from single B cells before and after treatment. By using this approach, we could obtain the activation and differentiation status of each cell and thus potentially identify the Igs which recognize highly immunogenic antigens in the tumor. These antibodies would, in turn, allow for the identification of antigens that drive intratumoral adaptive immune response. Either future work would benefit the field of cancer immunology by further helping to interpret the mechanisms regulating the interactions between the immune system and the TME, and possibly expand the benefits of immunotherapy to more patients.

To conclude, this thesis aimed to investigate anti-tumor immunity by characterizing the TME and the antigen receptor repertoires associated with intratumoral immune responses. I hope that the findings presented in Chapters 2 and 3 will further ignite interest in the role of B cells in anti-tumor immunity, while Chapter 4 establishes a solid analytical groundwork for further experiments.

References

[1] 10x Genomics (2019a). Sequencing requirements for single cell V(D)J - specifications - sequencing - single cell immune profiling - official 10x Genomics support. support.10xgenomics.com/single-cell-vdj/sequencing/doc/specifications-sequencing-requirements-for-single-cell-vdj. [Accessed 13/07/2019].

[2] 10x Genomics (2019b). Single cell immune profiling. support.10xgenomics.com/single-cell-vdj/software/analysis-of-multiple-libraries/latest/overview. [Accessed 10/10/2019].

[3] Actor, J. K. (2014). *Introductory immunology: Basic concepts for interdisciplinary applications*. Elsevier.

[4] Afik, S., Raulet, G., and Yosef, N. (2019). Reconstructing B-cell receptor sequences from short-read single-cell RNA sequencing with BRAPeS. *Life Science Alliance*, 2(4):e201900371.

[5] Afik, S., Yates, K. B., Bi, K., Darko, S., Godec, J., Gerdemann, U., Swadling, L., Douek, D. C., Klenerman, P., Barnes, E. J., Sharpe, A. H., Haining, W. N., and Yosef, N. (2017). Targeted reconstruction of T cell receptor sequence from single cell RNA-seq links CDR3 length to T cell differentiation state. *Nucleic Acids Research*, 45(16).

[6] Ahmadzadeh, M., Johnson, L. A., Heemskerk, B., Wunderlich, J. R., Dudley, M. E., White, D. E., and Rosenberg, S. A. (2009). Tumor antigen-specific CD8 T cells infiltrating the tumor express high levels of PD-1 and are functionally impaired. *Blood*, 114(8):1537–1544.

[7] Ahmed, R., Omidian, Z., Giwa, A., Cornwell, B., Majety, N., Bell, D. R., Lee, S., Zhang, H., Michels, A., Desiderio, S., Sadegh-Nasseri, S., Rabb, H., Gritsch, S., Suva, M. L., Cahan, P., Zhou, R., Jie, C., Donner, T., and Hamad, A. R. A. (2019). A public BCR present in a unique dual-receptor-expressing lymphocyte from type 1 diabetes patients encodes a potent T cell autoantigen. *Cell*, 177(6):1583–1599.e16.

[8] American Cancer Society (2019a). Key statistics for melanoma skin cancer. www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html. [Accessed 10/10/2019].

[9] American Cancer Society (2019b). Survival rates for colorectal cancer. www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html. [Accessed 28/09/2019].

[10] Ammirante, M., Luo, J. L., Grivennikov, S., Nedospasov, S., and Karin, M. (2010). B-cell-derived lymphotoxin promotes castration-resistant prostate cancer. *Nature*, 464(7286):302–305.

[11] Arstila, T. P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., and Kourilsky, P. (1999). A direct estimate of the human $\alpha\beta$ T cell receptor diversity. *Science*, 286(5441):958–961.

[12] Ashouri, J. F. and Weiss, A. (2016). Endogenous nur77 is a specific indicator of antigen receptor signaling in human T and B cells. *The Journal of Immunology*, 198(2):657–668.

[13] Azizi, E., Carr, A. J., Plitas, G., Cornish, A. E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., Choi, K., Fromme, R. M., Dao, P., McKenney, P. T., Wasti, R. C., Kadaveru, K., Mazutis, L., Rudensky, A. Y., and Pe'er, D. (2018). Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell*, 174(5):1293–1308.e36.

[14] Bai, X., Zhang, Q., Wu, S., Zhang, X., Wang, M., He, F., Wei, T., Yang, J., Lou, Y., Cai, Z., and Liang, T. (2015). Characteristics of tumor infiltrating lymphocyte and circulating lymphocyte repertoires in pancreatic cancer by the sequencing of T cell receptors. *Scientific Reports*, 5(1):13664.

[15] Baitsch, L., Baumgaertner, P., Devêvre, E., Raghav, S. K., Legat, A., Barba, L., Wieckowski, S., Bouzourene, H., Deplancke, B., Romero, P., Rufer, N., and Speiser, D. E. (2011). Exhaustion of tumor-specific CD8+ T cells in metastases from melanoma patients. *The Journal of Clinical Investigation*, 121(6):2350–2360.

[16] Barone, F., Gardner, D. H., Nayar, S., Steinthal, N., Buckley, C. D., and Luther, S. A. (2016). Stromal fibroblasts in tertiary lymphoid structures: A novel target in chronic inflammation. *Frontiers in Immunology*, 7(NOV).

[17] Bashford-Rogers, R. J. M., Palser, A. L., Idris, S. F., Carter, L., Epstein, M., Callard, R. E., Douek, D. C., Vassiliou, G. S., Follows, G. A., Hubank, M., and Kellam, P. (2014). Capturing needles in haystacks: A comparison of B-cell receptor sequencing methods. *BMC Immunology*, 15:29.

[18] Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44.

[19] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In *Lecture Notes in Computer Science*, pages 217–235. Springer Berlin Heidelberg.

[20] Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., and Canaider, S. (2013). An estimation of the number of cells in the human body. *Annals of Human Biology*, 40(6):463–471.

[21] Bilal, S., Lie, K. K., Sæle, Ø., and Hordvik, I. (2018). T cell receptor alpha chain genes in the teleost Ballan wrasse (Labrus bergylta) are subjected to somatic hypermutation. *Frontiers in Immunology*, 9(MAY).

[22] Björklund, A. K., Forkel, M., Picelli, S., Konya, V., Theorell, J., Friberg, D., Sandberg, R., and Mjösberg, J. (2016). The heterogeneity of human CD127+ innate lymphoid cells revealed by single-cell RNA sequencing. *Nature Immunology*, 17(4):451–460.

[23] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

[24] Bolotin, D. A., Poslavsky, S., Davydov, A. N., Frenkel, F. E., Fanchi, L., Zolotareva, O. I., Hemmers, S., Putintseva, E. V., Obraztsova, A. S., Shugay, M., Ataullakhanov, R. I., Rudensky, A. Y., Schumacher, T. N., and Chudakov, D. M. (2017). Antigen receptor repertoire profiling from RNA-seq data. *Nature Biotechnology*, 35(10):908–911.

[25] Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., and Chudakov, D. M. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. *Nature Methods*, 12(5):380–381.

[26] Bonaventura, P., Shekarian, T., Alcazer, V., Valladeau-Guilemond, J., Valsesia-Wittmann, S., Amigorena, S., Caux, C., and Depil, S. (2019). Cold tumors: A therapeutic challenge for immunotherapy. *Frontiers in Immunology*, 10:168.

[27] Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., and Zaretskaya, I. (2013). BLAST: A more efficient report with usability improvements. *Nucleic Acids Research*, 41(Web Server issue).

[28] Borst, J., Ahrends, T., Bąbała, N., Melief, C. J. M., and Kastenmüller, W. (2018). CD4+ T cell help in cancer immunology and immunotherapy. *Nature Reviews Immunology*, 18(10):635–647.

[29] Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., Jones, C. D., Simen, B. B., Hanczaruk, B., Nguyen, K. D., Nadeau, K. C., Egholm, M., Miklos, D. B., Zehnder, J. L., and Fire, A. Z. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Science Translational Medicine*, 1(12).

[30] Brady, B. L., Steinel, N. C., and Bassing, C. H. (2010). Antigen receptor allelic exclusion: An update and reappraisal. *The Journal of Immunology*.

[31] Brahmer, J. R., Tykodi, S. S., Chow, L. Q., Hwu, W. J., Topalian, S. L., Hwu, P., Drake, C. G., Camacho, L. H., Kauh, J., Odunsi, K., Pitot, H. C., Hamid, O., Bhatia, S., Martins, R., Eaton, K., Chen, S., Salay, T. M., Alaparthy, S., Grosso, J. F., Korman, A. J., Parker, S. M., Agrawal, S., Goldberg, S. M., Pardoll, D. M., Gupta, A., and Wigginton, J. M. (2012). Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *New England Journal of Medicine*, 366(26):2455–2465.

[32] Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527.

[33] Brown, S. D., Raeburn, L. A., and Holt, R. A. (2015). Profiling tissue-resident T cell repertoires by RNA sequencing. *Genome Medicine*, 7(1).

[34] Bruce Alberts and Alexander Johnson and Julian Lewis and David Morgan and Martin Raff (2014). *Molecular Biology of the Cell (Sixth Edition)*. Taylor and Francis.

[35] Bruno, T. C., Ebner, P. J., Moore, B. L., Squalls, O. G., Waugh, K. A., Eruslanov, E. B., Singhal, S., Mitchell, J. D., Franklin, W. A., Merrick, D. T., McCarter, M. D., Palmer, B. E., Kern, J. A., and Slansky, J. E. (2017). Antigen-presenting intratumoral B cells affect CD4+ TIL phenotypes in non-small cell lung cancer patients. *Cancer Immunology Research*, 5(10):898–907.

[36] Buechler, M. B. and Turley, S. J. (2018). A short field guide to fibroblast function in immunity. *Seminars in Immunology*, 35:48–58.

[37] Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160.

[38] Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. (2017). f-scLVM: Scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biology*, 18(1).

[39] Calis, J. J. and Rosenberg, B. R. (2014). Characterizing immune repertoires by high throughput sequencing: Strategies and applications. *Trends in Immunology*, 35(12):581–590.

[40] Campbell, F. and Verbeke, C. S. (2013). *Pathology of the Pancreas*. Springer London.

[41] Cancer.Net (2019). Pancreatic Cancer: Statistics. www.cancer.net/cancer-types/pancreatic-cancer/statistics. [Accessed 22/11/2019].

[42] Canzar, S., Neu, K. E., Tang, Q., Wilson, P. C., and Khan, A. A. (2016). BASIC: BCR assembly from single cells. *Bioinformatics*, 33(3):btw631.

[43] Carreau, N. A. and Pavlick, A. C. (2019). Nivolumab and ipilimumab: Immunotherapy for treatment of malignant melanoma. *Future Oncology*, 15(4):349–358.

[44] Chen, H., Ye, F., and Guo, G. (2019). Revolutionizing immunology with single-cell RNA sequencing. *Cellular & Molecular Immunology*, 16(3):242–249.

[45] Clark, C. E., Hingorani, S. R., Mick, R., Combs, C., Tuveson, D. A., and Vonderheide, R. H. (2007). Dynamics of the immune reaction to pancreatic cancer from inception to invasion. *Cancer Research*, 67(19):9518–9527.

[46] Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

[47] Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771.

[48] Colbeck, E. J., Ager, A., Gallimore, A., and Jones, G. W. (2017). Tertiary lymphoid structures in cancer: Drivers of antitumor immunity, immunosuppression, or bystander sentinels in disease? *Frontiers in Immunology*, 8:1830.

[49] Coley, W. B. (1893). The treatment of malignant tumors by repeated inoculations of erysipelas: With a report of ten original cases. *The American Journal of the Medical Sciences*, page 487.

[50] CRUK (2019). A trial of plerixafor for cancer that has spread (CAM-PLEX). www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/a-trial-of-plerixafor-for-cancer-that-has-spread-cam-plex. [Accessed 29/09/2019].

[51] CRUK Cambridge Centre (2019). MelResist. crukcambridgecentre.org.uk/trials/melresist. [Accessed 10/10/2019].

[52] Darvin, P., Toor, S. M., Nair, V. S., and Elkord, E. (2018). Immune checkpoint inhibitors: recent progress and potential biomarkers. *Experimental & Molecular Medicine*, 50(12).

[53] Day, C. L., Kaufmann, D. E., Kiepiela, P., Brown, J. A., Moodley, E. S., Reddy, S., Mackey, E. W., Miller, J. D., Leslie, A. J., DePierres, C., Mncube, Z., Duraiswamy, J., Zhu, B., Eichbaum, Q., Altfeld, M., Wherry, E. J., Coovadia, H. M., Goulder, P. J. R., Klenerman, P., Ahmed, R., Freeman, G. J., and Walker, B. D. (2006). PD-1 expression on HIV-specific T cells is associated with T-cell exhaustion and disease progression. *Nature*, 443(7109):350–354.

[54] de Bruijn, N. (1946). A combinatorial problem. *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, 49(7):758–764.

[55] Desfrançois, J., Moreau-Aubry, A., Vignard, V., Godet, Y., Khammari, A., Dréno, B., Jotereau, F., and Gervois, N. (2010). Double positive CD4CD8 $\alpha\beta$ T cells: A new tumor-reactive population in human melanomas. *PLoS ONE*, 5(1):e8437.

[56] DiLillo, D. J., Yanaba, K., and Tedder, T. F. (2010). B cells are required for optimal CD4+ and CD8+ T cell tumor immunity: Therapeutic B cell depletion enhances B16 melanoma growth in mice. *Journal of immunology (Baltimore, Md. : 1950)*, 184(7):4006–4016.

[57] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

[58] Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J., and Schreiber, R. D. (2002). Cancer immunoediting: from immunosurveillance to tumor escape. *Nature Immunology*, 3(11):991–998.

[59] Dziubianau, M., Hecht, J., Kuchenbecker, L., Sattler, A., Stervbo, U., Rödelsperger, C., Nickel, P., Neumann, A. U., Robinson, P. N., Mundlos, S., Volk, H. D., Thiel, A., Reinke, P., and Babel, N. (2013). TCR repertoire analysis by next generation sequencing allows complex differential diagnosis of T cell-related pathology. *American Journal of Transplantation*, 13(11):2842–2854.

[60] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

[61] Ehrlich, P. (1909). Über den jetzigen stand der chemotherapie (The current state of carcinoma research). *Berichte der deutschen chemischen Gesellschaft*, 42(1):17–47.

[62] Elhanati, Y., Sethna, Z., Marcou, Q., Callan, C. G., Mora, T., and Walczak, A. M. (2015). Inferring processes underlying B-cell repertoire diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676).

[63] Engelhard, V. H., Rodriguez, A. B., Mauldin, I. S., Woods, A. N., Peske, J. D., and Slingluff, C. L. (2018). Immune cell infiltration and tertiary lymphoid structures as determinants of antitumor immunity. *The Journal of Immunology*, 200(2):432–442.

[64] Faint, J. M., Pilling, D., Akbar, A. N., Kitas, G. D., Bacon, P. A., and Salmon, M. (1999). Quantitative flow cytometry for the analysis of T cell receptor Vβ chain expression. *Journal of Immunological Methods*, 225(1-2):53–60.

[65] Feig, C., Jones, J. O., Kraman, M., Wells, R. J. B., Deonarine, A., Chan, D. S., Connell, C. M., Roberts, E. W., Zhao, Q., Caballero, O. L., Teichmann, S. A., Janowitz, T., Jodrell, D. I., Tuveson, D. A., and Fearon, D. T. (2013). Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti–PD-L1 immunotherapy in pancreatic cancer. *Proceedings of the National Academy of Sciences*, 110(50):20212–20217.

[66] Filbin, M. G., Tirosh, I., Hovestadt, V., Shaw, M. L., Escalante, L. E., Mathewson, N. D., Neftel, C., Frank, N., Pelton, K., Hebert, C. M., Haberler, C., Yizhak, K., Gojo, J., Egervari, K., Mount, C., Van Galen, P., Bonal, D. M., Nguyen, Q. D., Beck, A., Sinai, C., Czech, T., Dorfer, C., Goumnerova, L., Lavarino, C., Carcaboso, A. M., Mora, J., Mylvaganam, R., Luo, C. C., Peyrl, A., Popović, M., Azizi, A., Batchelor, T. T., Frosch, M. P., Martinez-Lage, M., Kieran, M. W., Bandopadhayay, P., Beroukhim, R., Fritsch, G., Getz, G., Rozenblatt-Rosen, O., Wucherpfennig, K. W., Louis, D. N., Monje, M., Slavc, I., Ligon, K. L., Golub, T. R., Regev, A., Bernstein, B. E., and Suvà, M. L. (2018). Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science*, 360(6386):331–335.

[67] Finotto, S., Neurath, M. F., Glickman, J. N., Qin, S., Lehr, H. A., Green, F. H. Y., Ackerman, K., Haley, K., Galle, P. R., Szabo, S. J., Drazen, J. M., De Sanctis, G. T., and Glimcher, L. H. (2002). Development of spontaneous airway changes consistent with human asthma in mice lacking T-bet. *Science*, 295(5553):336–338.

[68] Fletcher, A. L., Acton, S. E., and Knoblich, K. (2015). Lymph node fibroblastic reticular cells in health and disease. *Nature Reviews Immunology*, 15(6):350–361.

[69] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

[70] Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H., and Holt, R. A. (2009). Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Research*, 19(10):1817–1824.

[71] Fridman, W. H., Pagès, F., Sautès-Fridman, C., and Galon, J. (2012). The immune contexture in human tumours: Impact on clinical outcome. *Nature Reviews Cancer*, 12(4):298–306.

[72] Galon, J. and Bruni, D. (2019). Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nature Reviews Drug Discovery*, 18(3):197–218.

[73] Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., Zinzindohoué, F., Bruneval, P., Cugnenc, P. H., Trajanoski, Z., Fridman, W. H., and Pagès, F. (2006). Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313(5795):1960–1964.

[74] Garcia, K. C. and Adams, E. J. (2005). How the T cell receptor sees antigen—a structural view. *Cell*, 122(3):333–336.

[75] Gellert, M. (2002). V(D)J recombination: RAG proteins, repair factors, and regulation. *Annual Review of Biochemistry*, 71(1):101–132.

[76] Germain, C., Gnjatic, S., and Dieu-Nosjean, M.-C. (2015). Tertiary lymphoid structure-associated B cells are key players in anti-tumor immunity. *Frontiers in Immunology*, 6.

[77] Germain, C., Gnjatic, S., Tamzalit, F., Knockaert, S., Remark, R., Goc, J., Lepelley, A., Becht, E., Katsahian, S., Bizouard, G., Validire, P., Damotte, D., Alifano, M., Magdeleinat, P., Cremer, I., Teillaud, J. L., Fridman, W. H., Sautès-Fridman, C., and Dieu-Nosjean, M. C. (2014). Presence of B cells in tertiary lymphoid structures is associated with a protective immunity in patients with lung cancer. *American Journal of Respiratory and Critical Care Medicine*, 189(7):832–844.

[78] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

[79] Glick, B., Chang, T. S., and Jaap, R. G. (1956). The bursa of fabricius and antibody production. *Poultry Science*, 35(1):224–225.

[80] Goc, J., Fridman, W.-H., Sautès-Fridman, C., and Dieu-Nosjean, M.-C. (2013). Characteristics of tertiary lymphoid structures in primary cancers. *OncoImmunology*, 2(12):e26836.

[81] Gong, J., Chehrazi-Raffle, A., Reddi, S., and Salgia, R. (2018). Development of PD-1 and PD-L1 inhibitors as a form of cancer immunotherapy: A comprehensive review of registration trials and future considerations. *Journal for ImmunoTherapy of Cancer*, 6(8).

[82] Gooden, M. J. M., de Bock, G. H., Leffers, N., Daemen, T., and Nijman, H. W. (2011). The prognostic influence of tumour-infiltrating lymphocytes in cancer: A systematic review with meta-analysis. *British Journal of Cancer*, 105(1):93–103.

[83] Gorski, J., Piatek, T., Yassai, M., Gorski, J., and Maslanka, K. (1995). Improvements in repertoire analysis by CDR3 size spectratyping: Bifamily PCR. *Annals of the New York Academy of Sciences*, 756(1):99–102.

[84] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652.

[85] Graham, J. B. and Graham, R. M. (1959). The effect of vaccine on cancer patients. *Surgery, Gynecology & Obstetrics*, 24(5):535.

[86] Griss, J., Bauer, W., Wagner, C., Simon, M., Chen, M., Grabmeier-Pfistershammer, K., Maurer-Granofszky, M., Roka, F., Penz, T., Bock, C., Zhang, G., Herlyn, M., Glatz, K., Läubli, H., Mertz, K. D., Petzelbauer, P., Wiesner, T., Hartl, M., Pickl, W. F., Somasundaram, R., Steinberger, P., and Wagner, S. N. (2019). B cells sustain inflammation and predict response to immune checkpoint blockade in human melanoma. *Nature Communications*, 10(1):4186.

[87] Gros, A., Robbins, P. F., Yao, X., Li, Y. F., Turcotte, S., Tran, E., Wunderlich, J. R., Mixon, A., Farid, S., Dudley, M. E., Hanada, K.-i., Almeida, J. R., Darko, S., Douek, D. C., Yang, J. C., and Rosenberg, S. A. (2014). PD-1 identifies the patient-specific CD8+ tumor-reactive repertoire infiltrating human tumors. *The Journal of Clinical Investigation*, 124(5):2246–2259.

[88] Guimaraes, J. C. and Zavolan, M. (2016). Patterns of ribosomal protein expression specify normal and malignant human cells. *Genome Biology*, 17(1).

[89] Guo, X., Zhang, Y., Zheng, L., Zheng, C., Song, J., Zhang, Q., Kang, B., Liu, Z., Jin, L., Xing, R., Gao, R., Zhang, L., Dong, M., Hu, X., Ren, X., Kirchhoff, D., Roider, H. G., Yan, T., and Zhang, Z. (2018). Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nature Medicine*, 24(7):978–985.

[90] Gupta, R., Sinha, S., and Paul, R. N. (2018). The impact of microsatellite stability status in colorectal cancer. *Current Problems in Cancer*, 42(6):548–559.

[91] Gürel, M. and Biasci, D. (2018a). Technical report: Assessment of different methods to identify somatic mutations in ctDNA. www.sound-biomed.eu/wp-content/uploads/SOUND-D2.2.pdf. [Accessed 13/09/2019].

[92] Gürel, M. and Biasci, D. (2018b). Varbench: Open-source software for the standardized application of existing methods in liquid biopsy data analysis. www.sound-biomed.eu/wp-content/uploads/D2.1_UCAM.pdf. [Accessed 16/09/2019].

[93] Gürel, M., So, T., Biasci, D., Kania, K., Grenfell, R., Coupland, P., Smith, K., Jodrell, D., Tavaré, S., and Thaventhiran, J. (2018). Using single-cell transcriptomics of the novel agrsr-fate-mapping mouse to define T cell clonal dysfunction in cancer. Poster presented at the 2018 CRUK CI Retreat.

[94] Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427.

[95] Hall, B. M. (2015). T cells: Soldiers and spies—the surveillance and control of effector T cells by regulatory T cells. *Clinical Journal of the American Society of Nephrology*, 10(11):2050–2064.

[96] Heather, J. M., Ismail, M., Oakes, T., and Chain, B. (2018). High-throughput sequencing of the T-cell receptor repertoire: Pitfalls and opportunities. *Briefings in Bioinformatics*, 19(4):554–565.

[97] Heimberg, G., Bhatnagar, R., El-Samad, H., and Thomson, M. (2016). Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Systems*, 2(4):239–250.

[98] Hodi, F. S., O'Day, S. J., McDermott, D. F., Weber, R. W., Sosman, J. A., Haanen, J. B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J. C., Akerley, W., Van Den Eertwegh, A. J., Lutzky, J., Lorigan, P., Vaubel, J. M., Linette, G. P., Hogg, D., Ottensmeier, C. H., Lebbé, C., Peschel, C., Quirt, I., Clark, J. I., Wolchok, J. D., Weber, J. S., Tian, J., Yellin, M. J., Nichol, G. M., Hoos, A., and Urba, W. J. (2010). Improved survival with ipilimumab in patients with metastatic melanoma. *New England Journal of Medicine*, 363(8):711–723.

[99] Hopkins, A. C., Yarchoan, M., Durham, J. N., Yusko, E. C., Rytlewski, J. A., Robins, H. S., Laheru, D. A., Le, D. T., Lutz, E. R., and Jaffee, E. M. (2018). T cell receptor repertoire features associated with survival in immunotherapy-treated pancreatic ductal adenocarcinoma. *JCI Insight*, 3(13).

[100] Hruban, R. H. and Klimstra, D. S. (2014). Adenocarcinoma of the pancreas. *Seminars in Diagnostic Pathology*, 31(6):443–451.

[101] Hu, X., Zhang, J., Liu, J. S., Li, B., and Liu, X. S. (2018). Evaluation of immune repertoire inference methods from RNA-seq data. *Nature Biotechnology*, 36(11):1034–1034.

[102] Hu, X., Zhang, J., Wang, J., Fu, J., Li, T., Zheng, X., Wang, B., Gu, S., Jiang, P., Fan, J., Ying, X., Zhang, J., Carroll, M. C., Wucherpfennig, K. W., Hacohen, N., Zhang, F., Zhang, P., Liu, J. S., Li, B., and Liu, X. S. (2019). Landscape of B cell immunity and related immune evasion in human cancers. *Nature Genetics*, 51(3):560–567.

[103] Hughes, C. E., Benson, R. A., Bedaj, M., and Maffia, P. (2016). Antigen-presenting cells and antigen presentation in tertiary lymphoid organs. *Frontiers in Immunology*, 7.

[104] ICGC (2019). International Cancer Genome Consortium (ICGC). icgc.org. [Accessed 25/09/2019].

[105] Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17(1):29.

[106] IMGT (2019). IMGT/GENE-DB statistics. www.imgt.org/genedb/stats. [Accessed 25/09/2019].

[107] Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166.

[108] Jang, K.-J., Mano, H., Aoki, K., Hayashi, T., Muto, A., Nambu, Y., Takahashi, K., Itoh, K., Taketani, S., Nutt, S. L., Igarashi, K., Shimizu, A., and Sugai, M. (2015). Mitochondrial function provides instructive signals for activation-induced B-cell fates. *Nature Communications*, 6(1):6750.

[109] Jayaram, N., Bhowmick, P., and Martin, A. C. R. (2012). Germline VH/VL pairing in antibodies. *Protein Engineering, Design and Selection*, 25(10):523–530.

[110] Joachims, M. L., Chain, J. L., Hooker, S. W., Knott-Craig, C. J., and Thompson, L. F. (2006). Human $\alpha\beta$ and $\gamma\delta$ thymocyte development: TCR gene rearrangements, intracellular TCRβ expression, and $\gamma\delta$ developmental potential - differences between men and mice. *The Journal of Immunology*, 176(3):1543–1552.

[111] Joyce, J. A. and Fearon, D. T. (2015). T cell exclusion, immune privilege, and the tumor microenvironment. *Science*, 348(6230):74–80.

[112] June, C. H., O'Connor, R. S., Kawalekar, O. U., Ghassemi, S., and Milone, M. C. (2018). CAR T cell immunotherapy for human cancer. *Science*, 359(6382):1361–1365.

[113] Kaufmann, S. H. E. (2019). Immunology's coming of age. *Frontiers in Immunology*, 10:684.

[114] Khan, O., Giles, J. R., McDonald, S., Manne, S., Ngiow, S. F., Patel, K. P., Werner, M. T., Huang, A. C., Alexander, K. A., Wu, J. E., Attanasio, J., Yan, P., George, S. M., Bengsch, B., Staupe, R. P., Donahue, G., Xu, W., Amaravadi, R. K., Xu, X., Karakousis, G. C., Mitchell, T. C., Schuchter, L. M., Kaye, J., Berger, S. L., and Wherry, E. J. (2019). TOX transcriptionally and epigenetically programs CD8+ T cell exhaustion. *Nature*, 571(7764):211–218.

[115] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36.

[116] Kim, J. H., You, K. R., Kim, I. H., Cho, B. H., Kim, C. Y., and Kim, D. G. (2004). Over-expression of the ribosomal protein L36a gene is associated with cellular proliferation in hepatocellular carcinoma. *Hepatology*, 39(1):129–138.

[117] Kindt, T. J., Goldsby, R. A., Osborne, B. A., and Kuby, J. (2007). *Kuby immunology*. New York: W.H. Freeman.

[118] Kirsch, I., Vignali, M., and Robins, H. (2015). T-cell receptor profiling in cancer. *Molecular Oncology*, 9(10):2063–2070.

[119] Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(5):273–282.

[120] Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74.

[121] Klein, G. (1966). Tumor antigens. *Annual Review of Microbiology*, 20(1):223–252.

[122] Klein, L., Kyewski, B., Allen, P. M., and Hogquist, K. A. (2014). Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature Reviews Immunology*, 14(6):377–391.

[123] Kroeger, D. R., Milne, K., and Nelson, B. H. (2016). Tumor-infiltrating plasma cells are associated with tertiary lymphoid structures, cytolytic T-cell responses, and superior prognosis in ovarian cancer. *Clinical Cancer Research*, 22(12):3005–3015.

[124] Kuchroo, V. K., Anderson, A. C., and Petrovas, C. (2014). Coinhibitory receptors and CD8 T cell exhaustion in chronic infections. *Current Opinion in HIV and AIDS*, 9(5):439–445.

[125] Kumar, B. V., Connors, T. J., and Farber, D. L. (2018). Human T cell development, localization, and function throughout life. *Immunity*, 48(2):202–213.

[126] Kumar, V., Abbas, A. K., and Aster, J. C. (2014). *Robbins & Cotran Pathologic Basis of Disease (Robbins Pathology)*. Elsevier.

[127] Küppers, R. (2005). Mechanisms of B-cell lymphoma pathogenesis. *Nature Reviews Cancer*, 5(4):251–262.

[128] Ladányi, A., Kiss, J., Mohos, A., Somlai, B., Liszkay, G., Gilde, K., Fejős, Z., Gaudi, I., Dobos, J., and Tímár, J. (2011). Prognostic impact of B-cell density in cutaneous melanoma. *Cancer Immunology, Immunotherapy*, 60(12):1729–1738.

[129] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

[130] Larkin, J., Chiarion-Sileni, V., Gonzalez, R., Grob, J.-J., Rutkowski, P., Lao, C. D., Cowey, C. L., Schadendorf, D., Wagstaff, J., Dummer, R., Ferrucci, P. F., Smylie, M., Hogg, D., Hill, A., Márquez-Rodas, I., Haanen, J., Guidoboni, M., Maio, M., Schöffski, P., Carlino, M. S., Lebbé, C., McArthur, G., Ascierto, P. A., Daniels, G. A., Long, G. V., Bastholt, L., Rizzo, J. I., Balogh, A., Moshyk, A., Hodi, F. S., and Wolchok, J. D. (2019). Five-year survival with combined nivolumab and ipilimumab in advanced melanoma. *New England Journal of Medicine*, 381(16):1535–1546.

[131] Laydon, D. J., Bangham, C. R. M., and Asquith, B. (2015). Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1675).

[132] Lechner, A., Schlößer, H. A., Thelen, M., Wennhold, K., Rothschild, S. I., Gilles, R., Quaas, A., Siefer, O. G., Huebbers, C. U., Cukuroglu, E., Göke, J., Hillmer, A., Gathof, B., Meyer, M. F., Klussmann, J. P., Shimabukuro-Vornhagen, A., Theurich, S., Beutner, D., and von Bergwelt-Baildon, M. (2019). Tumor-associated B cells and humoral immune response in head and neck squamous cell carcinoma. *OncoImmunology*, 8(3):1535293.

[133] Lee, S. and Margolin, K. (2011). Cytokines in cancer immunotherapy. *Cancers*, 3(4):3856–3893.

[134] Lefranc, M.-P., Giudicelli, V., Duroux, P., Jabado-Michaloud, J., Folch, G., Aouinti, S., Carillon, E., Duvergey, H., Houles, A., Paysan-Lafosse, T., Hadi-Saljoqi, S., Sasorith, S., Lefranc, G., and Kossida, S. (2015). IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Research*, 43(D1):D413–D422.

[135] Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G., and Duroux, P. (2009). IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Research*, 37(Database):D1006–D1012.

[136] Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323.

[137] Li, B., Li, T., Pignon, J.-C., Wang, B., Wang, J., Shukla, S. A., Dou, R., Chen, Q., Hodi, F. S., Choueiri, T. K., Wu, C., Hacohen, N., Signoretti, S., Liu, J. S., and Liu, X. S. (2016). Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nature Genetics*, 48(7):725–732.

[138] Li, B., Li, T., Wang, B., Dou, R., Zhang, J., Liu, J. S., and Liu, X. S. (2017). Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA-seq data. *Nature Genetics*, 49(4):482–483.

[139] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[140] Li, H., van der Leun, A. M., Yofe, I., Lubling, Y., Gelbard-Solodkin, D., van Akkooi, A. C., van den Braber, M., Rozeman, E. A., Haanen, J. B., Blank, C. U., Horlings, H. M., David, E., Baran, Y., Bercovich, A., Lifshitz, A., Schumacher, T. N., Tanay, A., and Amit, I. (2019). Dysfunctional CD8 T cells form a proliferative, dynamically regulated compartment within human melanoma. *Cell*, 176(4):775–789.e18.

[141] Li, Q., Lao, X., Pan, Q., Ning, N., Yet, J., Xu, Y., Li, S., and Chang, A. E. (2011). Adoptive transfer of tumor reactive B cells confers host T-cell immunity and tumor regression. *Clinical Cancer Research*, 17(15):4987–4995.

[142] Liao, Y., Smyth, G. K., and Shi, W. (2013). The subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108.

[143] Lindahl, T., Satoh, M. S., Poirier, G. G., and Klungland, A. (1995). Post-translational modification of poly(ADP-ribose) polymerase induced by DNA strand breaks. *Trends in Biochemical Sciences*, 20(10):405–411.

[144] Lindeman, I., Emerton, G., Mamanova, L., Snir, O., Polanski, K., Qiao, S.-W., Sollid, L. M., Teichmann, S. A., and Stubbington, M. J. T. (2018). BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nature Methods*, 15(8):563–565.

[145] Liston, A., Lesage, S., Wilson, J., Peltonen, L., and Goodnow, C. C. (2003). Aire regulates negative selection of organ-specific T cells. *Nature Immunology*, 4(4):350–354.

[146] Liu, Q., Li, Z., Gao, J.-L., Wan, W., Ganesan, S., McDermott, D. H., and Murphy, P. M. (2015). CXCR4 antagonist AMD3100 redistributes leukocytes from primary immune organs to secondary immune organs, lung, and blood in mice. *European Journal of Immunology*, 45(6):1855–1867.

[147] Lossius, A., Johansen, J. N., Vartdal, F., Robins, H., Jūratė Šaltytė, B., Holmøy, T., and Olweus, J. (2014). High-throughput sequencing of TCR repertoires in multiple sclerosis reveals intrathecal enrichment of EBV-reactive CD8+ T cells. *European Journal of Immunology*, 44(11):3439–3452.

[148] Lun, A. T. L., Bach, K., and Marioni, J. C. (2016a). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(75).

[149] Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016b). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 5:2122.

[150] Lun, A. T. L., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., and Marioni, J. C. (2019). EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology*, 20(63).

[151] Mahmoud, S. M. A., Lee, A. H. S., Paish, E. C., Macmillan, R. D., Ellis, I. O., and Green, A. R. (2012). The prognostic significance of B lymphocytes in invasive carcinoma of the breast. *Breast Cancer Research and Treatment*, 132(2):545–553.

[152] Mak, T., Saunders, M., and Jett, B. (2014). *Primer to the Immune Response*. Academic Cell.

[153] Marcus, A., Gowen, B. G., Thompson, T. W., Iannello, A., Ardolino, M., Deng, W., Wang, L., Shifrin, N., and Raulet, D. H. (2014). Recognition of tumors by the innate immune system and natural killer cells. In *Advances in Immunology*, pages 91–128. Elsevier.

[154] Martinez-Lostao, L., Anel, A., and Pardo, J. (2015). How do cytotoxic lymphocytes kill cancer cells? *Clinical Cancer Research*, 21(22):5047–5056.

[155] McDade, T. W., Georgiev, A. V., and Kuzawa, C. W. (2016). Trade-offs between acquired and innate immune defenses in humans. *Evolution, Medicine, and Public Health*, 2016(1):1–16.

[156] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at arxiv.org/abs/1802.03426.

[157] Melief, C. J., van Hall, T., Arens, R., Ossendorp, F., and van der Burg, S. H. (2015). Therapeutic cancer vaccines. *Journal of Clinical Investigation*, 125(9):3401–3412.

[158] Messina, J. L., Fenstermacher, D. A., Eschrich, S., Qu, X., Berglund, A. E., Lloyd, M. C., Schell, M. J., Sondak, V. K., Weber, J. S., and Mulé, J. J. (2012). 12-chemokine gene signature identifies lymph node-like structures in melanoma: Potential for patient selection for immunotherapy? *Scientific Reports*, 2(765).

[159] Mittal, D., Gubin, M. M., Schreiber, R. D., and Smyth, M. J. (2014). New insights into cancer immunoediting and its three component phases—elimination, equilibrium and escape. *Current Opinion in Immunology*, 27:16–25.

[160] Moncada, R., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., Hajdu, C. H., Simeone, D. M., and Yanai, I. (2019). Integrating single-cell RNA-Seq with spatial transcriptomics in pancreatic ductal adenocarcinoma using multimodal intersection analysis. *bioRxiv*.

[161] Mose, L. E., Selitsky, S. R., Bixby, L. M., Marron, D. L., Iglesia, M. D., Serody, J. S., Perou, C. M., Vincent, B. G., and Parker, J. S. (2016). Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. *Bioinformatics*, 32(24):3729–3734.

[162] Murphy, K. and Weaver, C. (2016). *Janeway's Immunobiology*. Garland Science.

[163] Naito, Y., Saito, K., Shiiba, K., Ohuchi, A., Saigenji, K., Nagura, H., and Ohtani, H. (1998). CD8+ T cells infiltrated within cancer cell nests as a prognostic factor in human colorectal cancer. *Cancer Research*, 58(16):3491–3494.

[164] National Cancer Institute (2019a). The cancer genome atlas program. www.cancer.gov/tcga. [Accessed 24/09/2019].

[165] National Cancer Institute (2019b). TCGA-SKCM. portal.gdc.cancer.gov/projects/TCGA-SKCM. [Accessed 25/09/2019].

[166] Nguyen, Q. P., Deng, T. Z., Witherden, D. A., and Goldrath, A. W. (2019). Origins of CD4+ circulating and tissue-resident memory T-cells. *Immunology*, 157(1):3–12.

[167] Nielsen, J. S., Sahota, R. A., Milne, K., Kost, S. E., Nesslinger, N. J., Watson, P. H., and Nelson, B. H. (2012). CD20+ tumor-infiltrating lymphocytes have an atypical CD27- memory phenotype and together with CD8+ T cells promote favorable prognosis in ovarian cancer. *Clinical Cancer Research*, 18(12):3281–3292.

[168] Nurieva, R. I., Chung, Y., Martinez, G. J., Yang, X. O., Tanaka, S., Matskevitch, T. D., Wang, Y. H., and Dong, C. (2009). Bcl6 mediates the development of T follicular helper cells. *Science*, 325(5943):1001–1005.

[169] O'Donnell, J. S., Teng, M. W. L., and Smyth, M. J. (2018). Cancer immunoediting and resistance to T cell-based immunotherapy. *Nature Reviews Clinical Oncology*, 16(3):151–167.

[170] Oliveira, A. F., Bretes, L., and Furtado, I. (2019). Review of PD-1/PD-L1 inhibitors in metastatic dMMR/MSI-h colorectal cancer. *Frontiers in Oncology*, 9.

[171] Overgaard, N. H., Jung, J.-W., Steptoe, R. J., and Wells, J. W. (2015). CD4+/CD8+ double-positive T cells: More than just a developmental stage? *Journal of Leukocyte Biology*, 97(1):31–38.

[172] Padera, T. P., Meijer, E. F., and Munn, L. L. (2016). The lymphatic system in disease processes and cancer progression. *Annual Review of Biomedical Engineering*, 18(1):125–158.

[173] Paley, M. A., Kroy, D. C., Odorizzi, P. M., Johnnidis, J. B., Dolfi, D. V., Barnett, B. E., Bikoff, E. K., Robertson, E. J., Lauer, G. M., Reiner, S. L., and Wherry, E. J. (2012). Progenitor and terminal subsets of CD8+ T cells cooperate to contain chronic viral infection. *Science*, 338(6111):1220–1225.

[174] Parham, P. (2009). *The Immune System, 3rd Edition*. Garland Science.

[175] Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., and Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401.

[176] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419.

[177] Perfetto, S. P., Chattopadhyay, P. K., and Roederer, M. (2004). Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunology*, 4(8):648–655.

[178] Pesce, S., Greppi, M., Grossi, F., Zotto, G. D., Moretta, L., Sivori, S., Genova, C., and Marcenaro, E. (2019). PD/1-PD-Ls checkpoint: Insight on the potential role of NK cells. *Frontiers in Immunology*, 10.

[179] Petersen-Mahrt, S. K., Harris, R. S., and Neuberger, M. S. (2002). AID mutates e. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature*, 418(6893):99–104.

[180] Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753.

[181] Philip, M. and Schietinger, A. (2019). Heterogeneity and fate choice: T cell exhaustion in cancer and chronic infections. *Current Opinion in Immunology*, 58:98–103.

[182] Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In *Computer and Information Sciences - ISCIS 2005*, pages 284–293. Springer Berlin Heidelberg.

[183] Protein Atlas (2019). Dictionary - Expression: PTPRC - The Human Protein Atlas. www.proteinatlas.org/learn/dictionary/expression/PTPRC. [Accessed 14/07/2019].

[184] Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C. L., Mroz, E. A., Emerick, K. S., Deschler, D. G., Varvares, M. A., Mylvaganam, R., Rozenblatt-Rosen, O., Rocco, J. W., Faquin, W. C., Lin, D. T., Regev, A., and Bernstein, B. E. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624.e24.

[185] Pylayeva-Gupta, Y., Das, S., Handler, J. S., Hajdu, C. H., Coffre, M., Koralov, S. B., and Bar-Sagi, D. (2016). IL35-producing B cells promote the development of pancreatic neoplasia. *Cancer Discovery*, 6(3):247–255.

[186] Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

[187] Rawla, P., Sunkara, T., and Gaduputi, V. (2019). Epidemiology of pancreatic cancer: Global trends, etiology and risk factors. *World Journal of Oncology*, 10(1):10–27.

[188] Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., Hodi, F. S., Martín-Algarra, S., Mandal, R., Sharfman, W. H., Bhatia, S., Hwu, W.-J., Gajewski, T. F., Slingluff, C. L., Chowell, D., Kendall, S. M., Chang, H., Shah, R., Kuo, F., Morris, L. G., Sidhom, J.-W., Schneck, J. P., Horak, C. E., Weinhold, N., and Chan, T. A. (2017). Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell*, 171(4):934–949.e16.

[189] Ribas, A. and Wolchok, J. D. (2018). Cancer immunotherapy using checkpoint blockade. *Science*, 359(6382):1350–1355.

[190] Rizzetto, S., Koppstein, D. N. P., Samir, J., Singh, M., Reed, J. H., Cai, C. H., Lloyd, A. R., Eltahla, A. A., Goodnow, C. C., and Luciani, F. (2018). B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. *Bioinformatics*, 34(16):2846–2847.

[191] Robert, C., Schachter, J., Long, G. V., Arance, A., Grob, J. J., Mortier, L., Daud, A., Carlino, M. S., McNeil, C., Lotem, M., Larkin, J., Lorigan, P., Neyns, B., Blank, C. U., Hamid, O., Mateus, C., Shapira-Frommer, R., Kosh, M., Zhou, H., Ibrahim, N., Ebbinghaus, S., and Ribas, A. (2015). Pembrolizumab versus ipilimumab in advanced melanoma. *New England Journal of Medicine*, 372(26):2521–2532.

[192] Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., Riddell, S. R., Warren, E. H., and Carlson, C. S. (2009). Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood*, 114(19):4099–4107.

[193] Rosati, E., Dowds, C. M., Liaskou, E., Henriksen, E. K. K., Karlsen, T. H., and Franke, A. (2017). Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnology*, 17(1):61.

[194] Rosvall, M. and Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331.

[195] Royal, R. E., Levy, C., Turner, K., Mathur, A., Hughes, M., Kammula, U. S., Sherry, R. M., Topalian, S. L., Yang, J. C., Lowy, I., and Rosenberg, S. A. (2010). Phase 2 trial of single agent ipilimumab (anti-CTLA-4) for locally advanced or metastatic pancreatic adenocarcinoma. *Journal of Immunotherapy*, 33(8):828–833.

[196] Rudensky, A. Y. (2011). Regulatory T cells and FOXP3. *Immunological Reviews*, 241(1):260–268.

[197] Sade-Feldman, M., Yizhak, K., Bjorgaard, S. L., Ray, J. P., de Boer, C. G., Jenkins, R. W., Lieb, D. J., Chen, J. H., Frederick, D. T., Barzily-Rokni, M., Freeman, S. S., Reuben, A., Hoover, P. J., Villani, A.-C., Ivanova, E., Portell, A., Lizotte, P. H., Aref, A. R., Eliane, J.-P., Hammond, M. R., Vitzthum, H., Blackmon, S. M., Li, B., Gopalakrishnan, V., Reddy, S. M., Cooper, Z. A., Paweletz, C. P., Barbie, D. A., Stemmer-Rachamimov, A., Flaherty, K. T., Wargo, J. A., Boland, G. M., Sullivan, R. J., Getz, G., and Hacohen, N. (2018). Defining T cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*, 175(4):998–1013.e20.

[198] Saeys, Y., Gassen, S. V., and Lambrecht, B. N. (2016). Computational flow cytometry: Helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16(7):449–462.

[199] Sakuishi, K., Apetoh, L., Sullivan, J. M., Blazar, B. R., Kuchroo, V. K., and Anderson, A. C. (2010). Targeting Tim-3 and PD-1 pathways to reverse T cell exhaustion and restore anti-tumor immunity. *The Journal of Experimental Medicine*, 207(10):2187–2194.

[200] Sant'Angelo, D. B., Lucas, B., Waterbury, P. G., Cohen, B., Brabb, T., Goverman, J., Germain, R. N., and Janeway, C. A. (1998). A molecular map of T cell development. *Immunity*, 9(2):179–186.

[201] Sautès-Fridman, C., Petitprez, F., Calderaro, J., and Fridman, W. H. (2019). Tertiary lymphoid structures in the era of cancer immunotherapy. *Nature Reviews Cancer*, 19(6):307–325.

[202] Savas, P., Virassamy, B., Ye, C., Salim, A., Mintoff, C. P., Caramia, F., Salgado, R., Byrne, D. J., Teo, Z. L., Dushyanthen, S., Byrne, A., Wein, L., Luen, S. J., Poliness, C., Nightingale, S. S., Skandarajah, A. S., Gyorki, D. E., Thornton, C. M., Beavis, P. A., Fox, S. B., Darcy, P. K., Speed, T. P., MacKay, L. K., Neeson, P. J., and Loi, S. (2018). Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nature Medicine*, 24(7):986–993.

[203] Schadendorf, D., Hodi, F. S., Robert, C., Weber, J. S., Margolin, K., Hamid, O., Patt, D., Chen, T.-T., Berman, D. M., and Wolchok, J. D. (2015). Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma. *Journal of Clinical Oncology*, 33(17):1889–1894.

[204] Schietinger, A., Philip, M., Krisnawan, V. E., Chiu, E. Y., Delrow, J. J., Basom, R. S., Lauer, P., Brockstedt, D. G., Knoblaugh, S. E., Hämmerling, G. J., Schell, T. D., Garbi, N., and Greenberg, P. D. (2016). Tumor-specific T cell dysfunction is a dynamic antigen-driven differentiation program initiated early during tumorigenesis. *Immunity*, 45(2):389–401.

[205] Schreiber, R. D., Old, L. J., and Smyth, M. J. (2011). Cancer immunoediting: Integrating immunity's roles in cancer suppression and promotion. *Science*, 331(6024):1565–1570.

[206] Schumacher, K., Haensch, W., Röefzaad, C., and Schlag, P. M. (2001). Prognostic significance of activated CD8(+) T cell infiltrations within esophageal carcinomas. *Cancer Research*, 61(10):3932–3936.

[207] Scialdone, A., Natarajan, K. N., Saraiva, L. R., Proserpio, V., Teichmann, S. A., Stegle, O., Marioni, J. C., and Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61.

[208] Scott, A. C., Dündar, F., Zumbo, P., Chandran, S. S., Klebanoff, C. A., Shakiba, M., Trivedi, P., Menocal, L., Appleby, H., Camara, S., Zamarin, D., Walther, T., Snyder, A., Femia, M. R., Comen, E. A., Wen, H. Y., Hellmann, M. D., Anandasabapathy, N., Liu, Y., Altorki, N. K., Lauer, P., Levy, O., Glickman, M. S., Kaye, J., Betel, D., Philip, M., and Schietinger, A. (2019). TOX is a critical regulator of tumour-specific T cell differentiation. *Nature*, 571(7764):270–274.

[209] See, P., Lum, J., Chen, J., and Ginhoux, F. (2018). A single-cell sequencing guide for immunologists. *Frontiers in Immunology*, 9:2425.

[210] Seidel, J. A., Otsuka, A., and Kabashima, K. (2018). Anti-PD-1 and anti-CTLA-4 therapies in cancer: Mechanisms of action, efficacy, and limitations. *Frontiers in Oncology*, 8:86.

[211] Seo, H., Chen, J., González-Avalos, E., Samaniego-Castruita, D., Das, A., Wang, Y. H., López-Moyado, I. F., Georges, R. O., Zhang, W., Onodera, A., Wu, C.-J., Lu, L.-F., Hogan, P. G., Bhandoola, A., and Rao, A. (2019). TOX and TOX2 transcription factors cooperate with NR4a transcription factors to impose CD8+ T cell exhaustion. *Proceedings of the National Academy of Sciences*, 116(25):12410–12415.

[212] Shah, D. K. and Zúñiga-Pflücker, J. C. (2014). An overview of the intrathymic intricacies of T cell development. *The Journal of Immunology*, 192(9):4017–4023.

[213] Shalapour, S., Font-Burgada, J., Di Caro, G., Zhong, Z., Sanchez-Lopez, E., Dhar, D., Willimsky, G., Ammirante, M., Strasner, A., Hansel, D. E., Jamieson, C., Kane, C. J., Klatte, T., Birner, P., Kenner, L., and Karin, M. (2015). Immunosuppressive plasma cells impede T-cell-dependent immunogenic chemotherapy. *Nature*, 521(7550):94–98.

[214] Sharma, P., Hu-Lieskovan, S., Wargo, J. A., and Ribas, A. (2017). Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell*, 168(4):707–723.

[215] Shcherbo, D., Murphy, C. S., Ermakova, G. V., Solovieva, E. A., Chepurnykh, T. V., Shcheglov, A. S., Verkhusha, V. V., Pletnev, V. Z., Hazelwood, K. L., Roche, P. M., Lukyanov, S., Zaraisky, A. G., Davidson, M. W., and Chudakov, D. M. (2009). Far-red fluorescent tags for protein imaging in living tissues. *Biochemical Journal*, 418(3):567–574.

[216] Shimabukuro-Vornhagen, A., Schlößer, H. A., Gryschok, L., Malcher, J., Wennhold, K., Garcia-Marquez, M., Herbold, T., Neuhaus, L. S., Becker, H. J., Fiedler, A., Scherwitz, P., Koslowsky, T., Hake, R., Stippel, D. L., Hölscher, A. H., Eidt, S., Hallek, M., Theurich, S., and von Bergwelt-Baildon, M. S. (2014). Characterization of tumor-associated B-cell subsets in patients with colorectal cancer. *Oncotarget*, 5(13):4651–4664.

[217] Shin, H., Blackburn, S. D., Intlekofer, A. M., Kao, C., Angelosanto, J. M., Reiner, S. L., and Wherry, E. J. (2009). A role for the transcriptional repressor Blimp-1 in CD8+ T cell exhaustion during chronic viral infection. *Immunity*, 31(2):309–320.

[218] Shitaoka, K., Hamana, H., Kishi, H., Hayakawa, Y., Kobayashi, E., Sukegawa, K., Piao, X., Lyu, F., Nagata, T., Sugiyama, D., Nishikawa, H., Tanemura, A., Katayama, I., Murahashi, M., Takamatsu, Y., Tani, K., Ozawa, T., and Muraguchi, A. (2018). Identification of tumoricidal TCRs from tumor-infiltrating lymphocytes by single-cell analysis. *Cancer Immunology Research*, 6(4):378–388.

[219] Simoni, Y., Becht, E., Fehlings, M., Loh, C. Y., Koo, S.-L., Teng, K. W. W., Yeong, J. P. S., Nahar, R., Zhang, T., Kared, H., Duan, K., Ang, N., Poidinger, M., Lee, Y. Y., Larbi, A., Khng, A. J., Tan, E., Fu, C., Mathew, R., Teo, M., Lim, W. T., Toh, C. K., Ong, B.-H., Koh, T., Hillmer, A. M., Takano, A., Lim, T. K. H., Tan, E. H., Zhai, W., Tan, D. S. W., Tan, I. B., and Newell, E. W. (2018). Bystander CD8+ T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature*, 557(7706):575–579.

[220] Singer, A., Adoro, S., and Park, J.-H. (2008). Lineage fate and intense debate: myths, models and mechanisms of CD4- versus CD8-lineage choice. *Nature Reviews Immunology*, 8(10):788–801.

[221] Singer, M., Wang, C., Cong, L., Marjanovic, N. D., Kowalczyk, M. S., Zhang, H., Nyman, J., Sakuishi, K., Kurtulus, S., Gennert, D., Xia, J., Kwon, J. Y., Nevin, J., Herbst, R. H., Yanai, I., Rozenblatt-Rosen, O., Kuchroo, V. K., Regev, A., and Anderson, A. C. (2016). A distinct gene module for dysfunction uncoupled from activation in tumor-infiltrating T cells. *Cell*, 166(6):1500–1511.e9.

[222] Six, A., Mariotti-Ferrandiz, M. E., Chaara, W., Magadan, S., Pham, H. P., Lefranc, M. P., Mora, T., Thomas-Vaslin, V., Walczak, A. M., and Boudinot, P. (2013). The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Frontiers in Immunology*, 4:413.

[223] Slaney, C. Y., Kershaw, M. H., and Darcy, P. K. (2014). Trafficking of T cells into tumors. *Cancer Research*, 74(24):7168–7174.

[224] Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

[225] Spear, S., Candido, J. B., McDermott, J. R., Ghirelli, C., Maniati, E., Beers, S. A., Balkwill, F. R., Kocher, H. M., and Capasso, M. (2019). Discrepancies in the tumor microenvironment of spontaneous and orthotopic murine models of pancreatic cancer uncover a new immunostimulatory phenotype for B cells. *Frontiers in Immunology*, 10:542.

[226] Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145.

[227] Stubbington, M. J. T., Lönnberg, T., Proserpio, V., Clare, S., Speak, A. O., Dougan, G., and Teichmann, S. A. (2016). T cell fate and clonality inference from single-cell transcriptomes. *Nature Methods*, 13(4):329–332.

[228] Sun, C., Mezzadra, R., and Schumacher, T. N. (2018). Regulation and function of the PD-L1 checkpoint. *Immunity*, 48(3):434–452.

[229] Swann, J. B. and Smyth, M. J. (2007). Immune surveillance of tumors. *Journal of Clinical Investigation*, 117(5):1137–1146.

[230] Takagi, M., Absalon, M. J., McLure, K. G., and Kastan, M. B. (2005). Regulation of p53 translation and induction after DNA damage by ribosomal protein L26 and nucleolin. *Cell*, 123(1):49–63.

[231] Teillaud, J.-L. and Dieu-Nosjean, M.-C. (2017). Tertiary lymphoid structures: An anti-tumor school for adaptive immune cells and an antibody factory to fight cancer? *Frontiers in Immunology*, 8:830.

[232] Thibult, M.-L., Mamessier, E., Gertner-Dardenne, J., Pastor, S., Just-Landi, S., Xerri, L., Chetaille, B., and Olive, D. (2012). PD-1 is a novel regulator of human B-cell activation. *International Immunology*, 25(2):129–137.

[233] Thommen, D. S. and Schumacher, T. N. (2018). T cell dysfunction in cancer. *Cancer Cell*, 33(4):547–562.

[234] Thompson, E. D., Enriquez, H. L., Fu, Y.-X., and Engelhard, V. H. (2010). Tumor masses support naive T cell infiltration, activation, and differentiation into effectors. *The Journal of Experimental Medicine*, 207(8):1791–1804.

[235] Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G. K., Kazer, S. W., Gaillard, A., Kolb, K. E., Villani, A.-C., Johannessen, C. M., Andreev, A. Y., Allen, E. M. V., Bertagnolli, M., Sorger, P. K., Sullivan, R. J., Flaherty, K. T., Frederick, D. T., Jane-Valbuena, J., Yoon, C. H., Rozenblatt-Rosen, O., Shalek, A. K., Regev, A., and Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196.

[236] Topalian, S. L., Drake, C. G., and Pardoll, D. M. (2015). Immune checkpoint blockade: A common denominator approach to cancer therapy. *Cancer Cell*, 27(4):450–461.

[237] Topalian, S. L., Hodi, F. S., Brahmer, J. R., Gettinger, S. N., Smith, D. C., McDermott, D. F., Powderly, J. D., Carvajal, R. D., Sosman, J. A., Atkins, M. B., Leming, P. D., Spigel, D. R., Antonia, S. J., Horn, L., Drake, C. G., Pardoll, D. M., Chen, L., Sharfman, W. H., Anders, R. A., Taube, J. M., McMiller, T. L., Xu, H., Korman, A. J., Jure-Kunkel, M., Agrawal, S., McDonald, D., Kollia, G. D., Gupta, A., Wigginton, J. M., and Sznol, M. (2012). Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *New England Journal of Medicine*, 366(26):2443–2454.

[238] Topham, N. J. and Hewitt, E. W. (2009). Natural killer cell cytotoxicity: How do they pull the trigger? *Immunology*, 128(1):7–15.

[239] Toubal, A., Nel, I., Lotersztajn, S., and Lehuen, A. (2019). Mucosal-associated invariant T cells and disease. *Nature Reviews Immunology*, 19(10):643–657.

[240] Trautmann, L., Janbazian, L., Chomont, N., Said, E. A., Gimmig, S., Bessette, B., Boulassel, M.-R., Delwart, E., Sepulveda, H., Balderas, R. S., Routy, J.-P., Haddad, E. K., and Sekaly, R.-P. (2006). Upregulation of PD-1 expression on HIV-specific CD8+ T cells leads to reversible immune dysfunction. *Nature Medicine*, 12(10):1198–1202.

[241] Tsou, P., Katayama, H., Ostrin, E. J., and Hanash, S. M. (2016). The emerging role of B cells in tumor immunity. *Cancer Research*, 76(19):5597–5601.

[242] Tumeh, P. C., Harview, C. L., Yearley, J. H., Shintaku, I. P., Taylor, E. J., Robert, L., Chmielowski, B., Spasic, M., Henry, G., Ciobanu, V., West, A. N., Carmona, M., Kivork, C., Seja, E., Cherry, G., Gutierrez, A. J., Grogan, T. R., Mateus, C., Tomasic, G., Glaspy, J. A., Emerson, R. O., Robins, H., Pierce, R. H., Elashoff, D. A., Robert, C., and Ribas, A. (2014). PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*, 515(7528):568–571.

[243] Upadhyay, A. A., Kauffman, R. C., Wolabaugh, A. N., Cho, A., Patel, N. B., Reiss, S. M., Havenar-Daughton, C., Dawoud, R. A., Tharp, G. K., Sanz, I., Pulendran, B., Crotty, S., Lee, F. E.-H., Wrammert, J., and Bosinger, S. E. (2018). BALDR: a computational pipeline for paired heavy and light chain immunoglobulin reconstruction in single-cell RNA-seq data. *Genome Medicine*, 10(1):20.

[244] Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nature Methods*, 14(6):565–571.

[245] van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

[246] Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., De Jager, P. L., Rozenblatt-Rosen, O., Lane, A. A., Haniffa, M., Regev, A., and Hacohen, N. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335).

[247] Wang, C., Sanders, C. M., Yang, Q., Schroeder, H. W., Wang, E., Babrzadeh, F., Gharizadeh, B., Myers, R. M., Hudson, J. R., Davis, R. W., and Han, J. (2010). High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proceedings of the National Academy of Sciences*, 107(4):1518–1523.

[248] Weinstein, J. A., Jiang, N., White, R. A., Fisher, D. S., and Quake, S. R. (2009). High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324(5928):807–810.

[249] Wherry, E. J., Ha, S.-J., Kaech, S. M., Haining, W. N., Sarkar, S., Kalia, V., Subramaniam, S., Blattman, J. N., Barber, D. L., and Ahmed, R. (2007). Molecular signature of CD8+ T cell exhaustion during chronic viral infection. *Immunity*, 27(4):670–684.

[250] WHO (2019). WHO Fact Sheet - Cancer 2019. www.who.int/news-room/fact-sheets/detail/cancer. [Accessed 02/09/2019].

[251] Wittrup, K. D. (2017). Antitumor antibodies can drive therapeutic T cell responses. *Trends in Cancer*, 3(9):615–620.

[252] Wolchok, J. D., Chiarion-Sileni, V., Gonzalez, R., Rutkowski, P., Grob, J.-J., Cowey, C. L., Lao, C. D., Wagstaff, J., Schadendorf, D., Ferrucci, P. F., Smylie, M., Dummer, R., Hill, A., Hogg, D., Haanen, J., Carlino, M. S., Bechter, O., Maio, M., Marquez-Rodas, I., Guidoboni, M., McArthur, G., Lebbé, C., Ascierto, P. A., Long, G. V., Cebon, J., Sosman, J., Postow, M. A., Callahan, M. K., Walker, D., Rollin, L., Bhore, R., Hodi, F. S., and Larkin, J. (2017). Overall survival with combined nivolumab and ipilimumab in advanced melanoma. *New England Journal of Medicine*, 377(14):1345–1356.

[253] Wolchok, J. D., Kluger, H., Callahan, M. K., Postow, M. A., Rizvi, N. A., Lesokhin, A. M., Segal, N. H., Ariyan, C. E., Gordon, R.-A., Reed, K., Burke, M. M., Caldwell, A., Kronenberg, S. A., Agunwamba, B. U., Zhang, X., Lowy, I., Inzunza, H. D., Feely, W., Horak, C. E., Hong, Q., Korman, A. J., Wigginton, J. M., Gupta, A., and Sznol, M. (2013). Nivolumab plus ipilimumab in advanced melanoma. *New England Journal of Medicine*, 369(2):122–133.

[254] Wouters, M. C. and Nelson, B. H. (2018). Prognostic significance of tumor-infiltrating B cells and plasma cells in human cancer. *Clinical Cancer Research*, 24(24):6125–6135.

[255] Xu, C. and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980.

[256] Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6(1):30750.

[257] Yao, C., Sun, H.-W., Lacey, N. E., Ji, Y., Moseman, E. A., Shih, H.-Y., Heuston, E. F., Kirby, M., Anderson, S., Cheng, J., Khan, O., Handon, R., Reilley, J., Fioravanti, J., Hu, J., Gossa, S., Wherry, E. J., Gattinoni, L., McGavern, D. B., O'Shea, J. J., Schwartzberg, P. L., and Wu, T. (2019). Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection. *Nature Immunology*, 20(7):890–901.

[258] Yates, A. J. (2014). Theories and quantification of thymic selection. *Frontiers in Immunology*, 5.

[259] Ye, J., Ma, N., Madden, T. L., and Ostell, J. M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41(W1):W34–W40.

[260] Yuen, G. J., Demissie, E., and Pillai, S. (2016). B lymphocytes and cancer: A love–hate relationship. *Trends in Cancer*, 2(12):747–757.

[261] Yuseff, M.-I., Pierobon, P., Reversat, A., and Lennon-Duménil, A.-M. (2013). How B cells capture, process and present antigens: a crucial role for cell polarity. *Nature Reviews Immunology*, 13(7):475–486.

[262] Zajac, A. J., Blattman, J. N., Murali-Krishna, K., Sourdive, D. J., Suresh, M., Altman, J. D., and Ahmed, R. (1998). Viral immune evasion due to persistence of activated T cells without effector function. *Journal of Experimental Medicine*, 188(12):2205–2213.

[263] Zamora, A. E., Crawford, J. C., and Thomas, P. G. (2018). Hitting the target: How T cells detect and eliminate tumors. *The Journal of Immunology*, 200(2):392–399.

[264] Zhang, L., Conejo-Garcia, J. R., Katsaros, D., Gimotty, P. A., Massobrio, M., Regnani, G., Makrigiannakis, A., Gray, H., Schlienger, K., Liebman, M. N., Rubin, S. C., and Coukos, G. (2003). Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. *New England Journal of Medicine*, 348(3):203–213.

[265] Zhang, L., Yu, X., Zheng, L., Zhang, Y., Li, Y., Fang, Q., Gao, R., Kang, B., Zhang, Q., Huang, J. Y., Konno, H., Guo, X., Ye, Y., Gao, S., Wang, S., Hu, X., Ren, X., Shen, Z., Ouyang, W., and Zhang, Z. (2018). Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature*, 564(7735):268–272.

[266] Zhao, J., Chen, A. X., Gartrell, R. D., Silverman, A. M., Aparicio, L., Chu, T., Bordbar, D., Shan, D., Samanamud, J., Mahajan, A., Filip, I., Orenbuch, R., Goetz, M., Yamaguchi, J. T., Cloney, M., Horbinski, C., Lukas, R. V., Raizer, J., Rae, A. I., Yuan, J., Canoll, P., Bruce, J. N., Saenger, Y. M., Sims, P., Iwamoto, F. M., Sonabend, A. M., and Rabadan, R. (2019). Immune and genomic correlates of response to anti-PD-1 immunotherapy in glioblastoma. *Nature Medicine*, 25(3):462–469.

[267] Zhao, L., Zhao, H., and Yan, H. (2018). Gene expression profiling of 1200 pancreatic ductal adenocarcinoma reveals novel subtypes. *BMC Cancer*, 18(603).

[268] Zheng, C., Zheng, L., Yoo, J.-K., Guo, H., Zhang, Y., Guo, X., Kang, B., Hu, R., Huang, J. Y., Zhang, Q., Liu, Z., Dong, M., Hu, X., Ouyang, W., Peng, J., and Zhang, Z. (2017a). Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell*, 169(7):1342–1356.e16.

[269] Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017b). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(14049).

Appendix A

Reconstructed clonotype sequences

A.1 Chapter 2 - Melanoma

```
> P2.BCR1.H{IGHV5-51:IGHD1-26:IGHJ3:IGHG2}
TCTTGCATCATACTTCTTTTCTTATATGGGGGAGTCTCCCTCACTGCCCAGCTGGGATCTCAGGGCTTCA
TTTTCTGTCCTCCGCCATCATGGGGTCAACCGCCATCCTCGCCCTCCTCCTGGCTGTTCTCCAGGGAGTC
TGTGGCGAGGTGCAGCTGGAGCAGTCTGGAGGAGAGGTGAAGAAGCCGGGGGAGTCTCTGAAGATCTCCT
GTAAGGCTTCTGGATACAGATTTACCAGCTTTTGGATCGTCTGGGTGCGCCAGATGCCCGGAAAAGGCCT
GGAGTGGGTGGGGATCATCCATCCTGGTGACTCCGATATTAGTTACAGCCCGTCTTTTGAAGGCCACGTC
ACCCTGTCAGCCGACAGGTCCAGCACCACCGCCTACCTGCAGTGGGACAGCCTGAAGGCCTCGGACAGCG
CCATGTATTACTGTGCGAGAAGGGTGGGAGCTACCTCTGCTTTTGATATCTGGGGCCTAGGGACACTGGT
CACCGTCTCTTCAGCCTCCACCAAGGGCCCATCGGTCTTCCCCCTGGCGCCCTGCTCCAGGAGCACCTCC
GAGAGCACAGCGGCCCTGGGCTGCCTGGTCAAGGACTACTTCCCCGAACCGGTGACGGTGTCGTGGAACT
CAGGCGCTCTGACCAGCGGCGTGCACACCTTCCCAGCTGTCCTACAGTCCTCAGGACTCTACTCCCTCAG
CAGCGTGGTGACCGTGCCCTCCAGCAACTTCGGCACCCAGACC
> P2.BCR1.L{IGLV3-25:IGLJ2:IGLC3}
GATCCTGCTTCTTCTTCTTCCTCTTGTCGCCTGTGCTGCCCCCACAGCTGGTTTGGGTGACATCTCTCCA
GGAGGAGTCCCAGAGGAAGTAAATTTGCATAAACACCAAACACTGACTACCCTAAAAAGCCTGAGAGAGA
ATAAGAGAGGCCTGGGGAGCCTAGCTGTGCTGTGGGTCCAGGAGGCAGAACTCTGGGTGTCTCACCATGG
CCTGGATCCCTCTACTTCTCCCCCTCCTCACTCTCTGCACAGACTCTGAGGCCGCCCATGAGTTGACACA
GCCACCCTCGGTGTCAGTGTCCCCAGGACAGACGGCCAGAATCACCTGCTCTGGAGATGGATTGTCAAAG
CAGTATGTTCATTGGTACCAGGCGAAGCCAGGCCAGGCCCCTGTCTTGGTGATATATAAAGACACTGAGA
GGCCCCCAGGAATCCCTGAGCGATTCTCTGCCTCCAGCTCAGCGACGACAGTCACATTGACCATTAGTGG
AGTCCAGGCAGAGGACGAGGCTGACTATTATTGTCAATCGGGAGACAGCCGTCTTACTTTTGTGGTTTTT
GGCGGCGGGACCAAGCTGACCGTCCTACGTCAGCCCAAGGCTGCCCCCTCGGTCACTCTGTTCCCGCCCT
CCTCTGAGGAGCTTCAAGCCAACAAGGCCACACTGGTGTGTCTCATAAGTGACTTCTACCCGGGAGCCGT
GACAGTGGCCTGGAAGGCAGATAGCAGCCCCGTCAAGGCGGGAGTGGAGACCTCCACACCCTCCAAACAA
AGCAACAACAAGTACGCGGCCAGCAGCTATCTGAGCCTGACGCCTGAGCAGTGGAAGTCCCACA
```

Listing A.1 Sequences of tumor-antigen specific antibody candidate

A.2 Chapter 3 - CRC

```
> Pat02.H{IGHV1-69:IGHD3-22:IGHJ4:IGHG1}
TCTAAAGAAGCCCCTGGGAGCACAGCTCATCACCATGGACTGGACCTGGAGGTTCCTCTTTGTGGTGGCA
GCAGCTACAGGTGTCCAATCCCAGGTCCAGTTGCTACAATCTGGGGCTGAGGTGAAGAAGCCTGGGTCCT
CGGTGAAGGTCTCCTGCAAGGCTTCTGGAGGCATCCTCAGCACCGATAGTATCAGCTGGGTGCGACAGGC
CCCGGGACAAGGGCTTGAGTGGATGGGAAGGATCATCCCTATCCTTGGTAGAGCAAACTACGCACAGACG
TTCCAGGGCAGAGTCACAATTATCGCGGACAGATTTACGAACACAGTCGAGATGGAGCTGAGCAGCCTGA
GATCTGAGGACTCGGCCGTCTATTTCTGTGTGAACGGCATATCAGATAGTCGTGGTCGCTACTACTTTGA
CTACTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG
> Pat02.L{IGKV1-39:IGKJ1:IGKC}
TGCAACCTGAAGATTTTGCAACTTACTACTGTCAACAGAGTTACAGTACCTGGTGGACGTTCGGCCAAGG
GACCAAGGTGGAAATCAAA
```

Listing A.2 Sequences of tumor-antigen specific antibody candidate reconstructed from patient 2

```
> Pat11.H{IGHV3-30:IGHD1-7:IGHJ4:IGHA2}
CACGCTGTATCTGCAAATGAACAGTCTGAGAGGTGAGGACTCGGCTGTGTATTACTGTGCGAAAGAAGAA
CTACGACGAGGGCCCCATGACTACTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG
> Pat11.L{IGKV1-39:IGKJ1:IGKC}
TCTGCAACCTGAAGATTTTGCAACTTACTTCTGTCAACAGAGTTACATTACCCCTCGGACGTTCGGCCAA
GGGACCAAGGTGGAAATCAAA
```

Listing A.3 Sequences of tumor-antigen specific antibody candidate reconstructed from patient 11

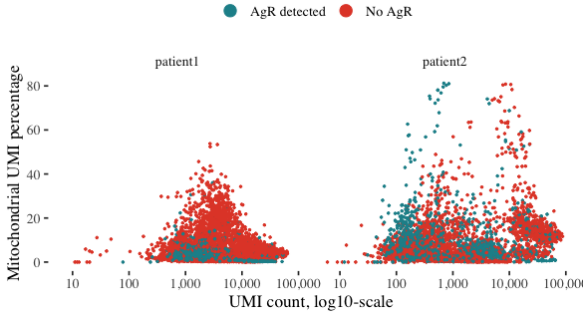
A.3 Chapter 3 - PDAC

```
> Pat25.H{IGHV3-30:IGHD6-13:IGHJ4:IGHG2}
GAACACGCTGTTTCTGCAAATGAACAGCCTGAGAAGTGAGGACACGTCTATGTATCACTGTGCGAAGGAT
TTGAGAGGCAACAGCTGGTCTTTTGACTACTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAGCCTCCA
CCAA
> Pat25.L{IGLV2-14:IGLJ3:IGLC2}
TCTCAGGAGGCAGCGCTCTCGGGACGTCTCCACCATGGCCTGGGCTCTGCTATTCCTCACCCTCCTCACT
CAGGGCACAGGGTCCTGGGCCCAGTCTGCCCTGACTCAGCCTGCCTCCGTGTCTGGGTCTCCTGGACAGT
CGATCACCATCTCCTGCACTGGAACCAGCAGTGACGTTGGTGGTTATAATTATGTCTCCTGGTACCAACA
GCACCCCGGCAAAGCCCCCAAACTCATAATTTTTGATGTCAATGATCGGCCCTCAGGGGTTTCTAATCGC
TTCTCTGGCTCCAAGTCTGGCAACACGGCCTCCCTGACCATCGCTGGGCTCCAGGCTGAGGACGAGGCTC
ATTATTACTGCAGCTCATATGCAATCACTAATTCTCTCGTTTTCGGCGGAGGGACCGAGCTGACCGTCCT
A
```

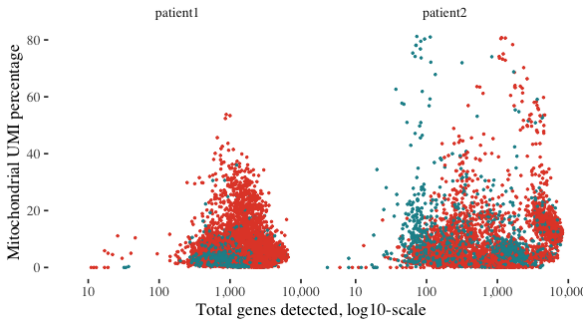
Listing A.4 Sequences of tumor-antigen specific antibody candidate reconstructed from patient 25

Appendix B

Supplementary Figures



(a)



(b)

Fig. B.1 (a) plots the percentage of UMI originating from mitochondrial genes against the library size and (b) against the detected gene counts of the *patient1* and *patient2* samples. Each point represents a cell. Color denotes whether an AgR was detected in a cell or not.

(a) Communities detected with WalkTrap

(b) Communities detected with Louvain

Fig. B.2 The *patient2* sample projected onto 2D space with UMAP where each point corresponds to a cell. Clusters are colored by membership as detected by the (a) WalkTrap and (b) Louvain algorithms. There are 27 distinct clusters in (a), and 15 in (b).

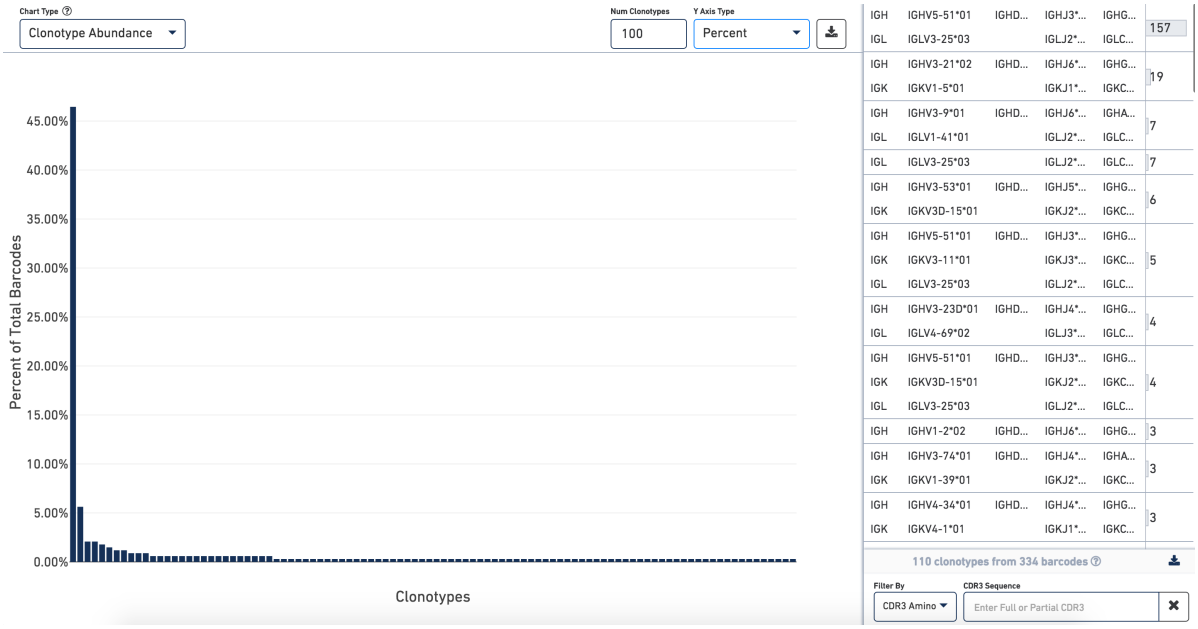


Fig. B.3 Screenshot of the Loupe VDJ browser opened with the BCR repertoire of the *patient2* sample.

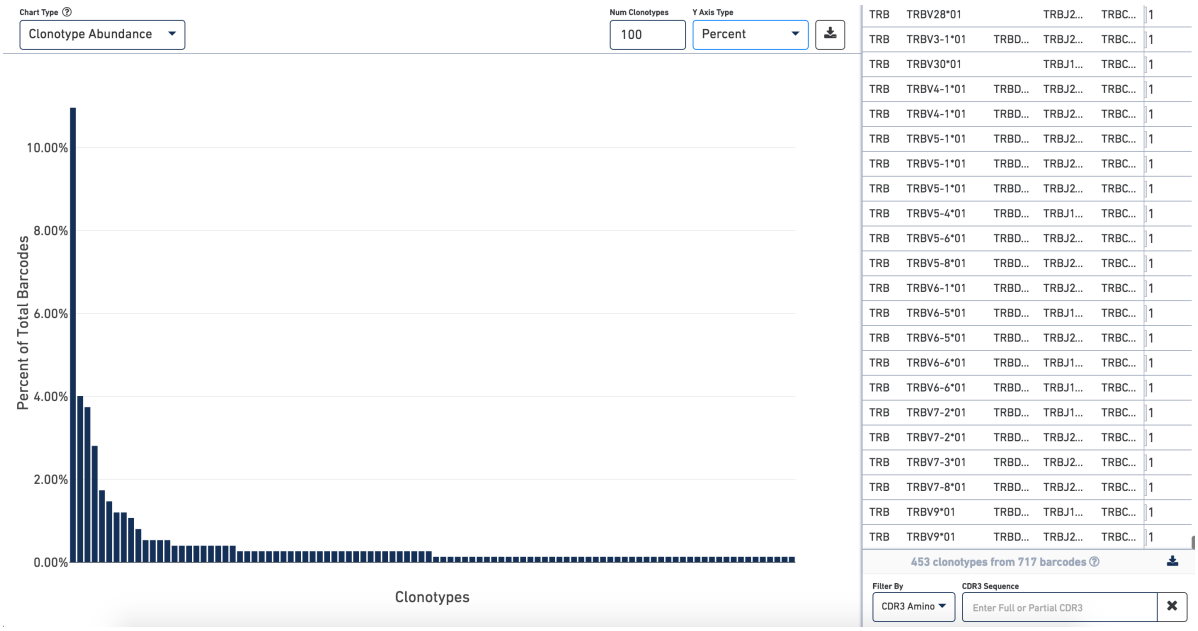


Fig. B.4 Screenshot of the Loupe VDJ browser opened with the TCR repertoire of the *patient2* sample.

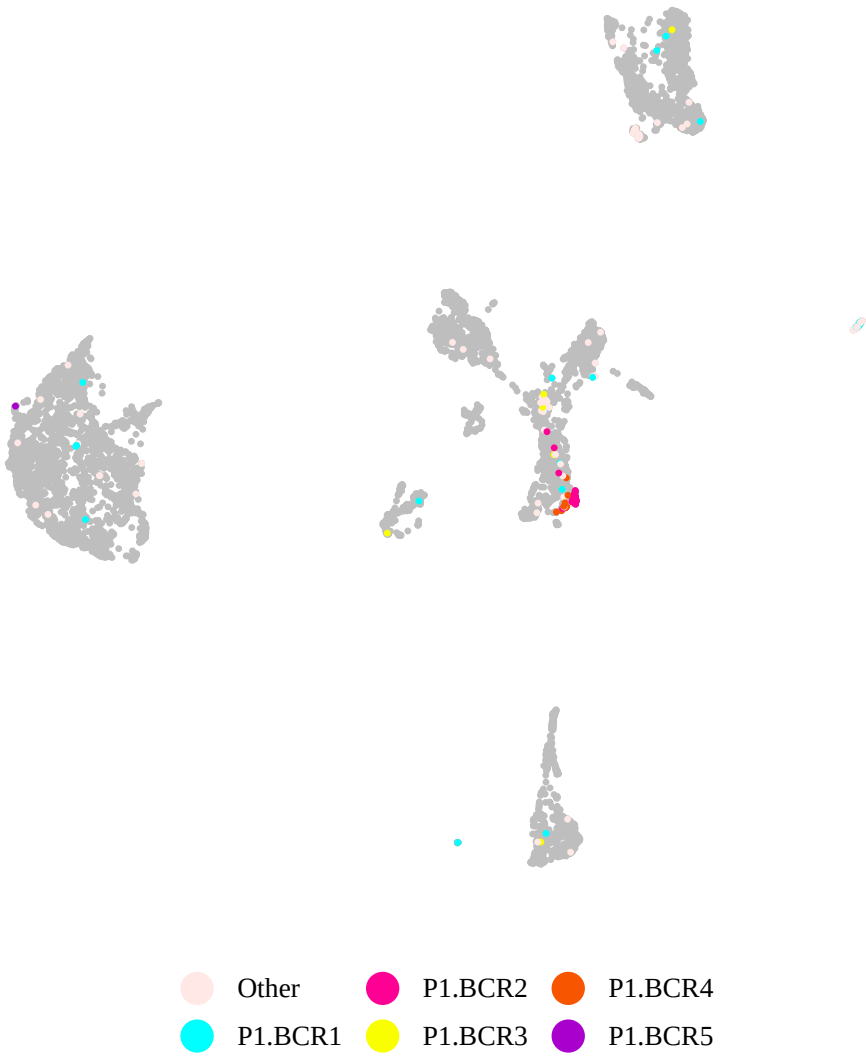
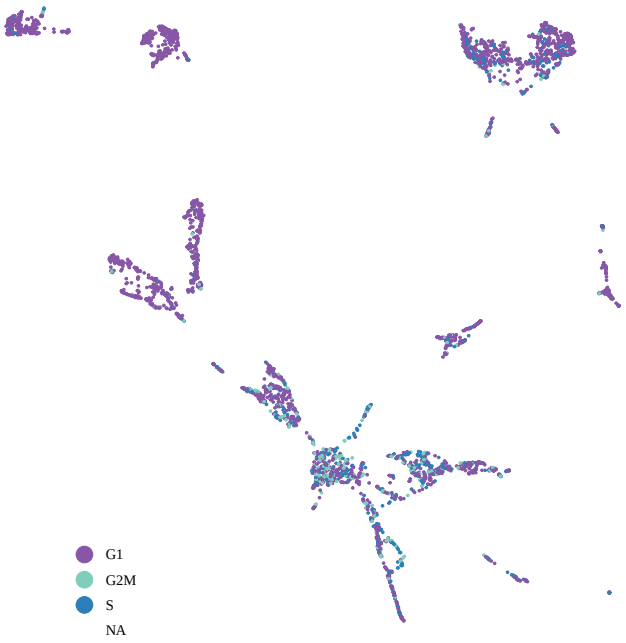
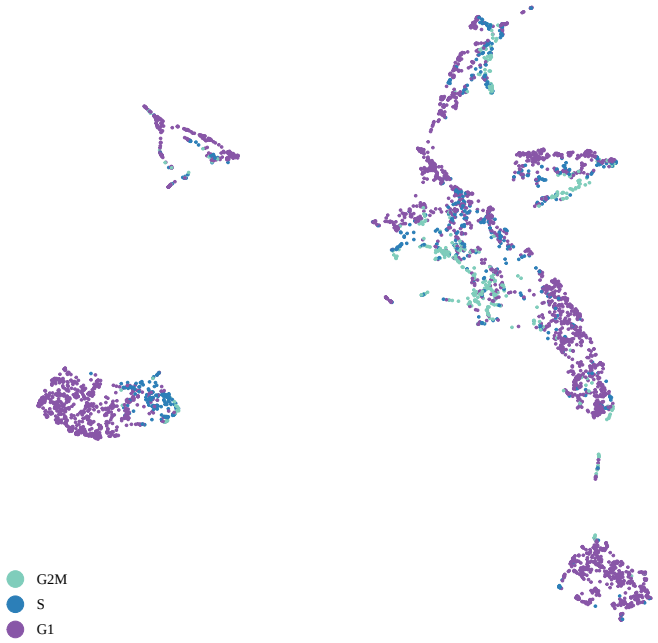


Fig. B.5 Clonotype projection of the *patient1* sample. UMAP embedded data is colored by clonotype membership. A BCR was not detected in cells colored with dark gray.

Fig. B.6 Clonotype projection of the *patient2* sample. UMAP embedded data is colored by clonotype membership. A BCR was not detected in cells colored with dark gray.



(a) Before correction



(b) After correction

Fig. B.7 Overlays of cell cycle phase on the UMAP of the *patient2* sample before and after cell cycle correction of gene expression. Each point represents a cell.

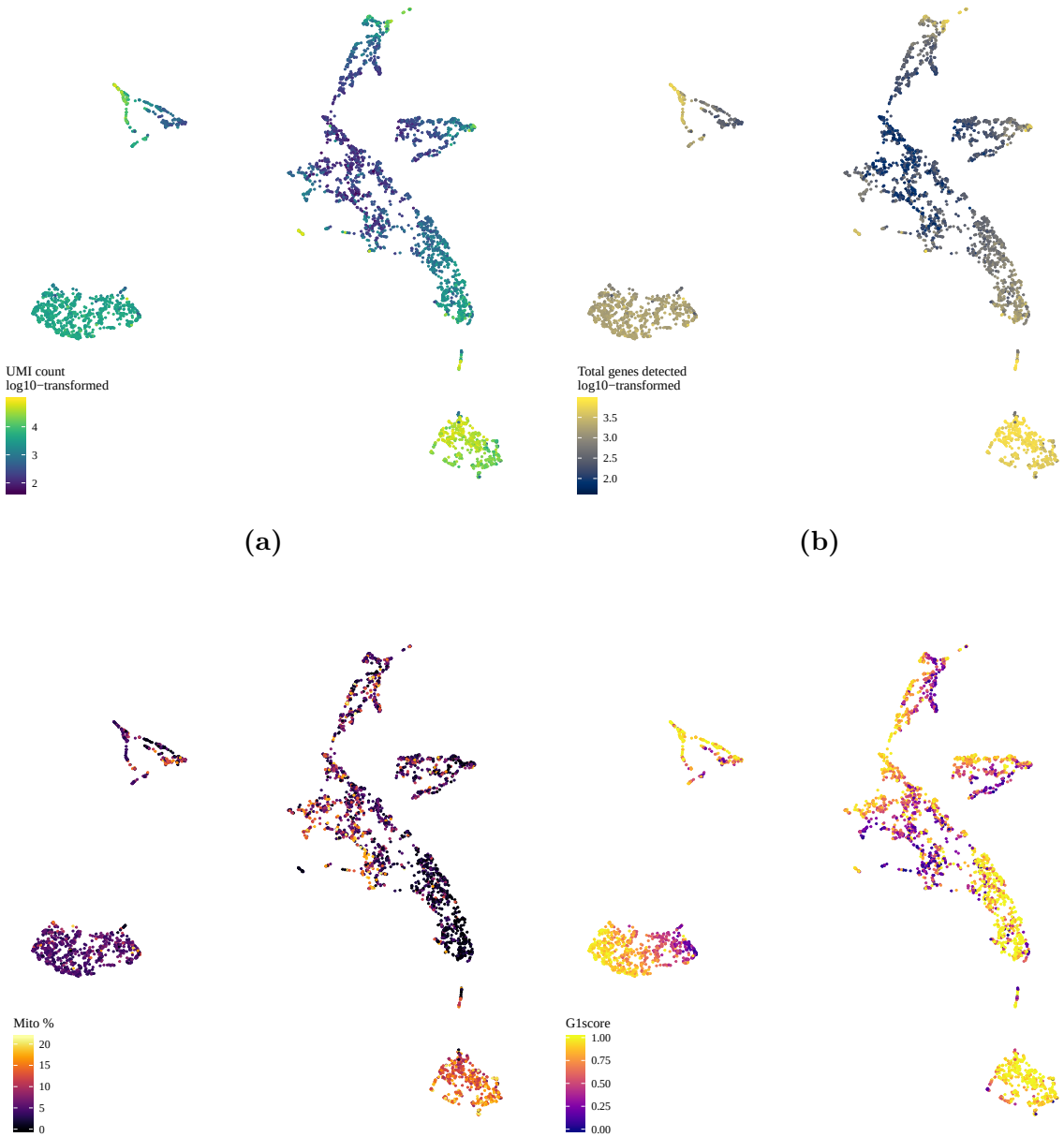


Fig. B.8 Overlays of each metadata on the UMAP of the *patient2* sample after cell cycle correction of gene expression. Each point represents a cell.

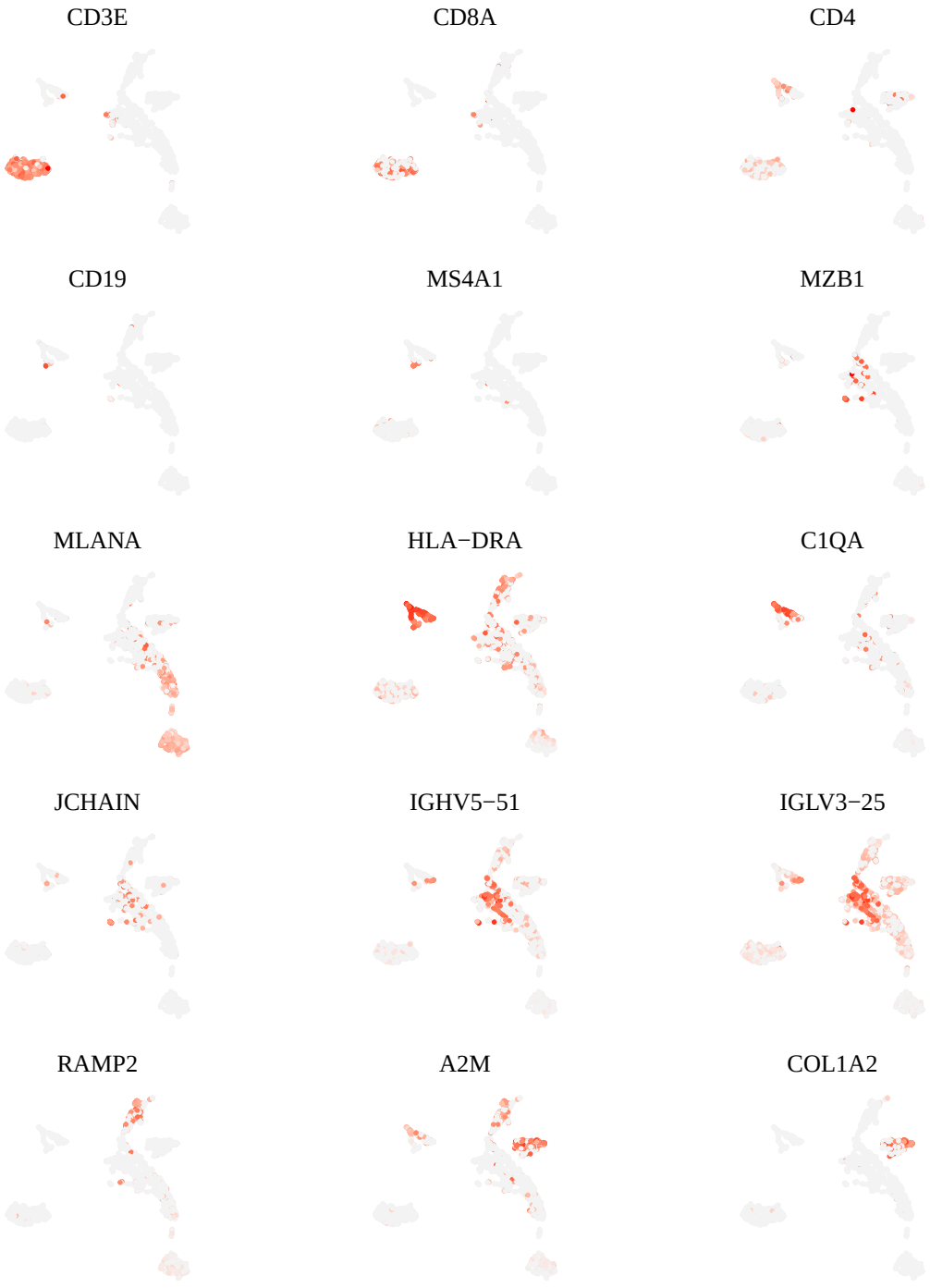
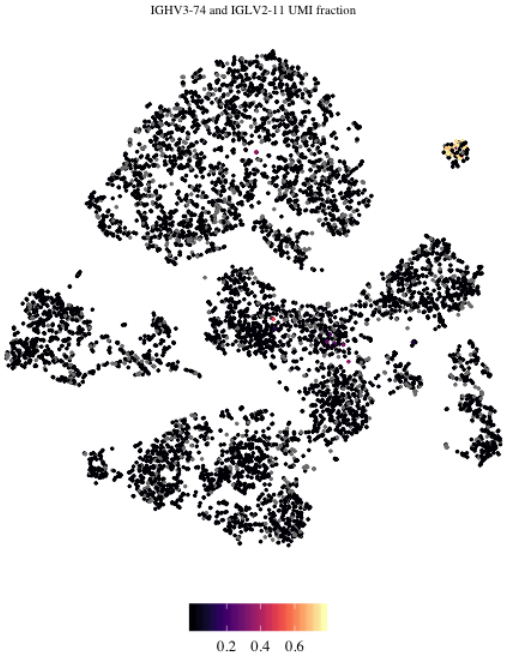


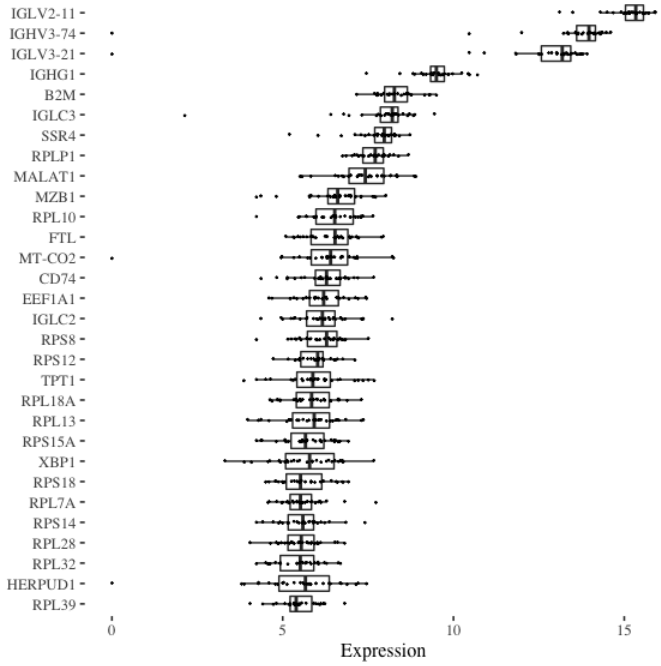
Fig. B.9 Overlays of selected gene expressions, colored from low (gray) to high (red), on the UMAP of the *patient2* sample after cell cycle correction. Each point represents a cell.



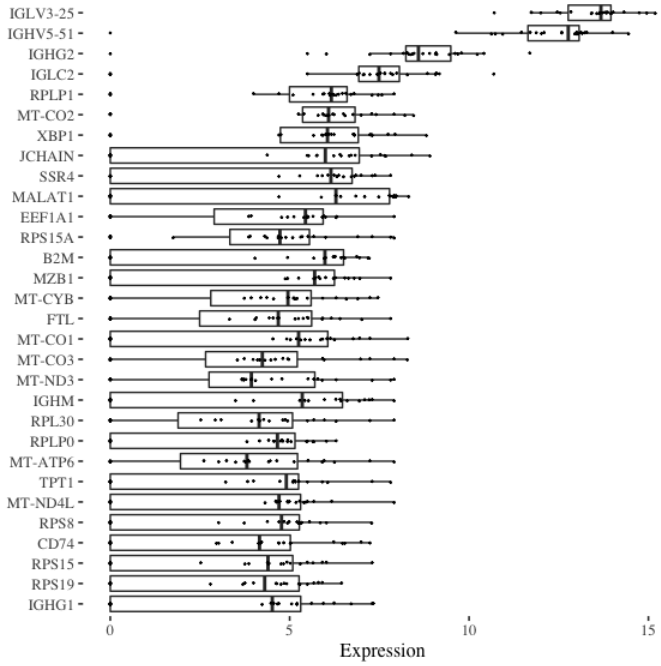
(a) *patient1* sample

(b) *patient2* sample

Fig. B.10 t-SNE projections of the most clonal antibody's heavy and light V genes' UMI fractions.



(a) *patient1* sample



(b) *patient2* sample

Fig. B.11 Most highly expressed genes in the cells with high (> 0.5) IGLV and IGHV expression. Gene expressions are normalized and log2-transformed.

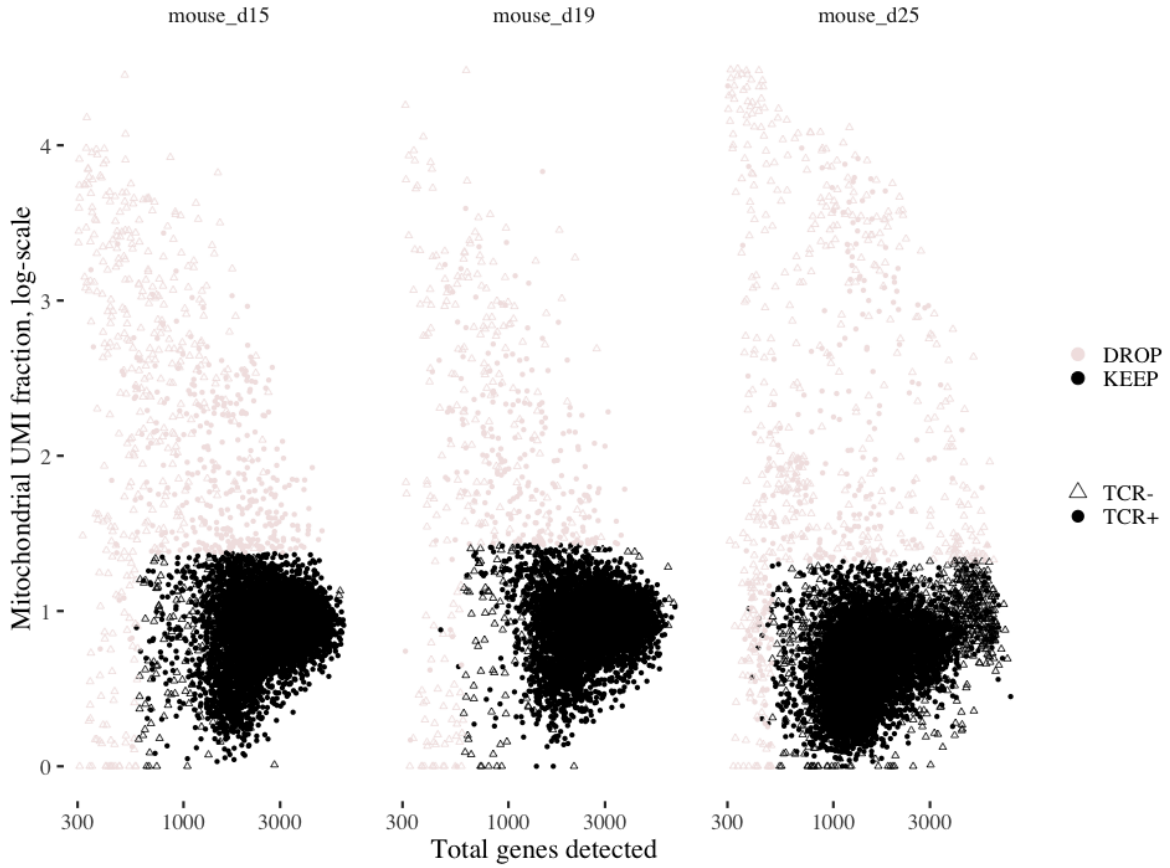


Fig. B.12 I remove cells with mitochondrial content greater than 3 MADs and with total genes detected less than 600, 600, and 500, in samples *mouse_d15*, *mouse_d19*, and *mouse_d25* respectively.

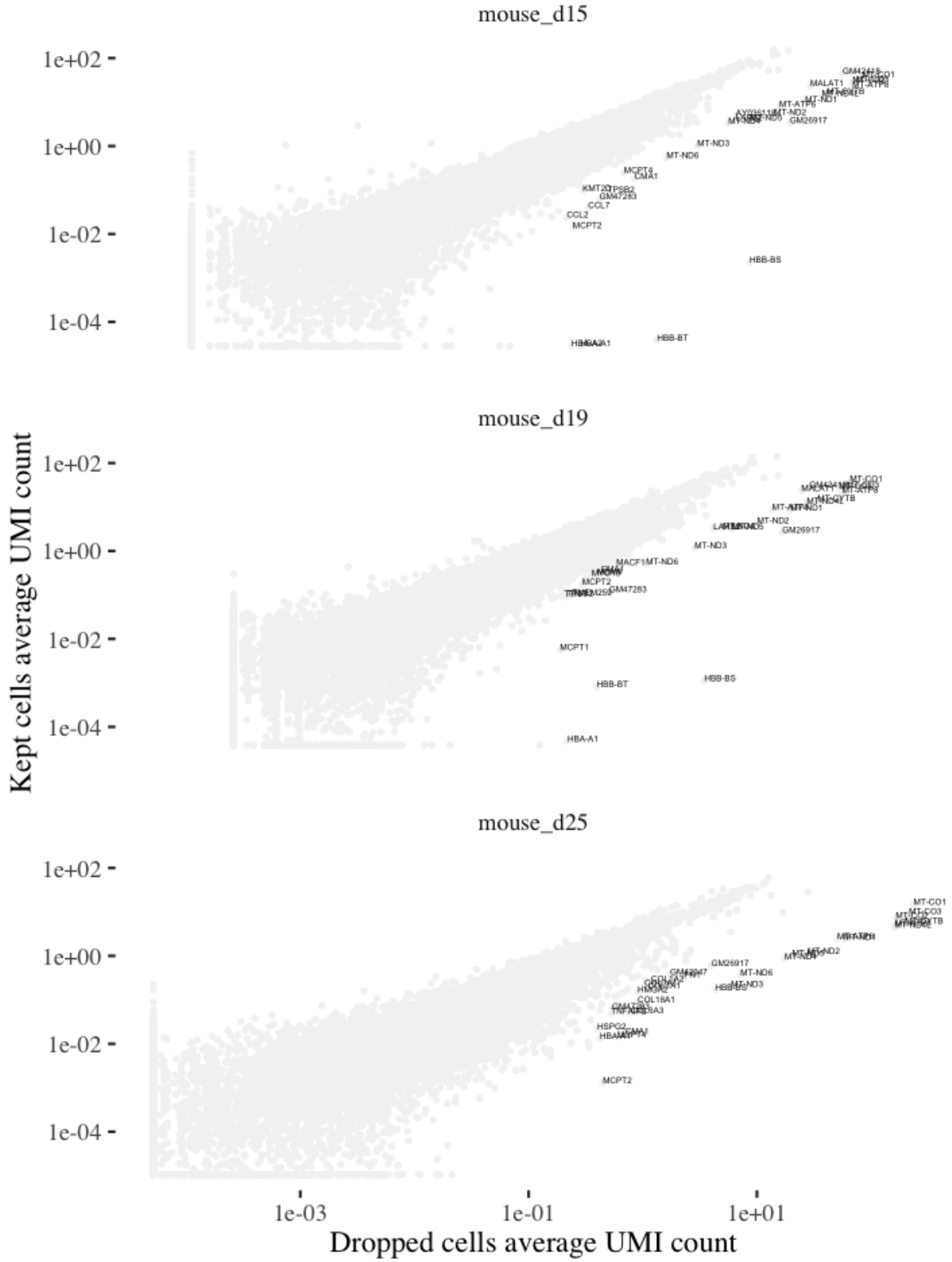


Fig. B.13 Average UMI count correlation between kept and dropped cells.

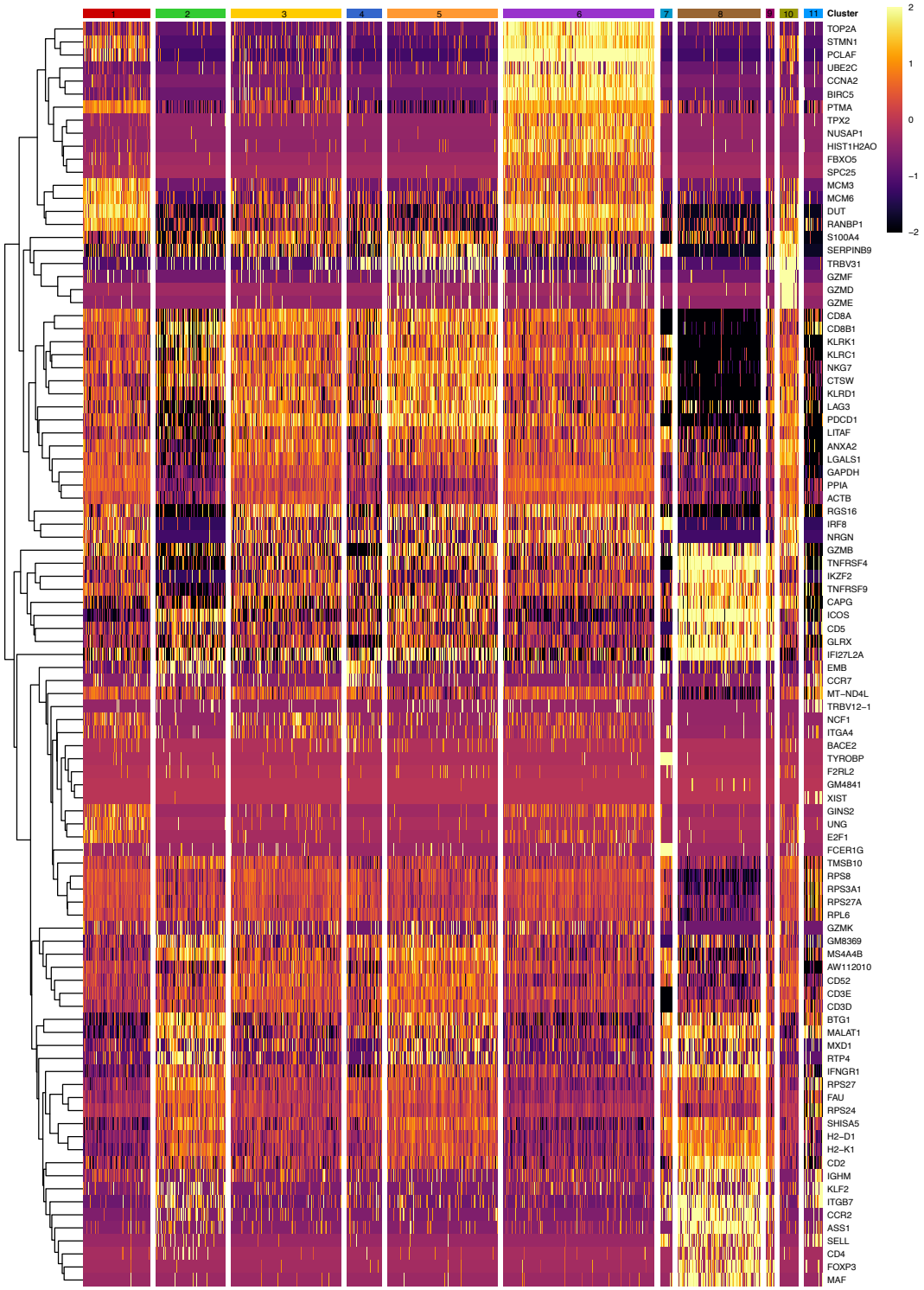


Fig. B.14 Cluster biomarkers of *mouse_d15*.

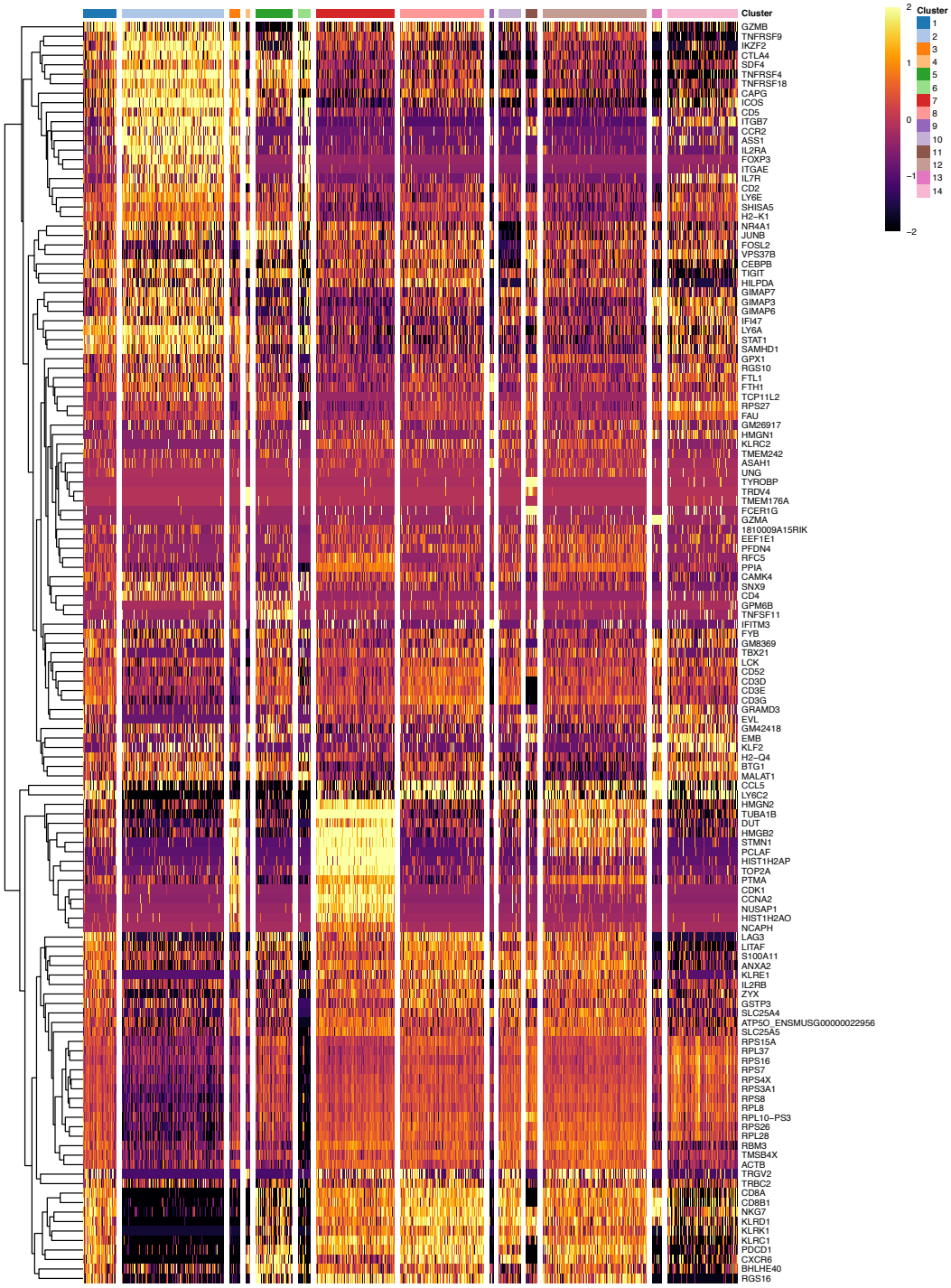


Fig. B.15 Cluster biomarkers of *mouse_d19*.

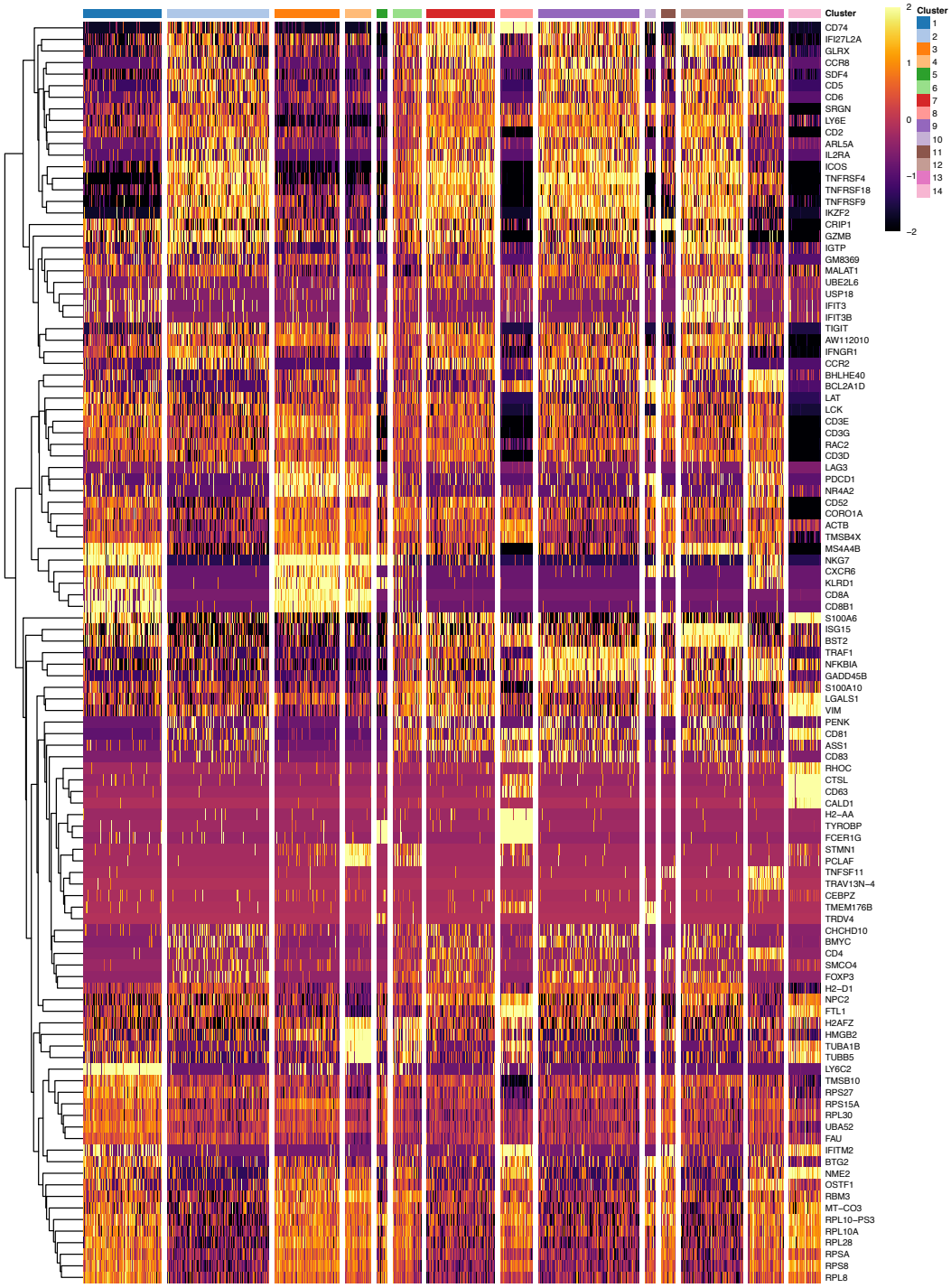


Fig. B.16 Cluster biomarkers of *mouse_d25*.

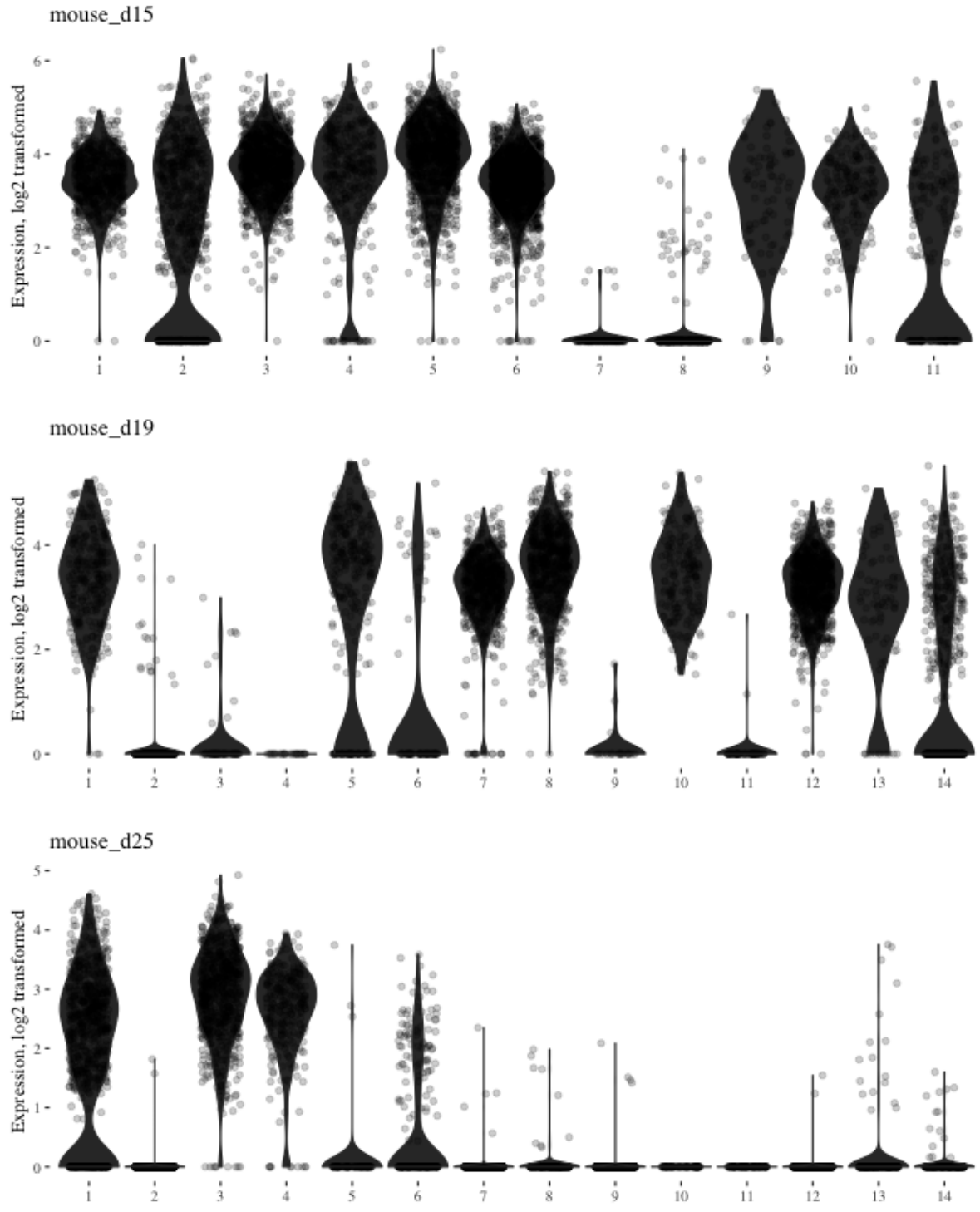


Fig. B.17 CD8 expression in the clusters of tumor samples. In *mouse_d15*, I subsetted on clusters 1,2,3,4,5,6,9,10,11 (top panel), in *mouse_d19* on clusters 1,5,7,8,10,12,13,14 (middle panel), and in *mouse_d25* on 1,3,4,6 (bottom panel). From each cluster, I only took the cells with a positive UMI count for the CD8 gene.

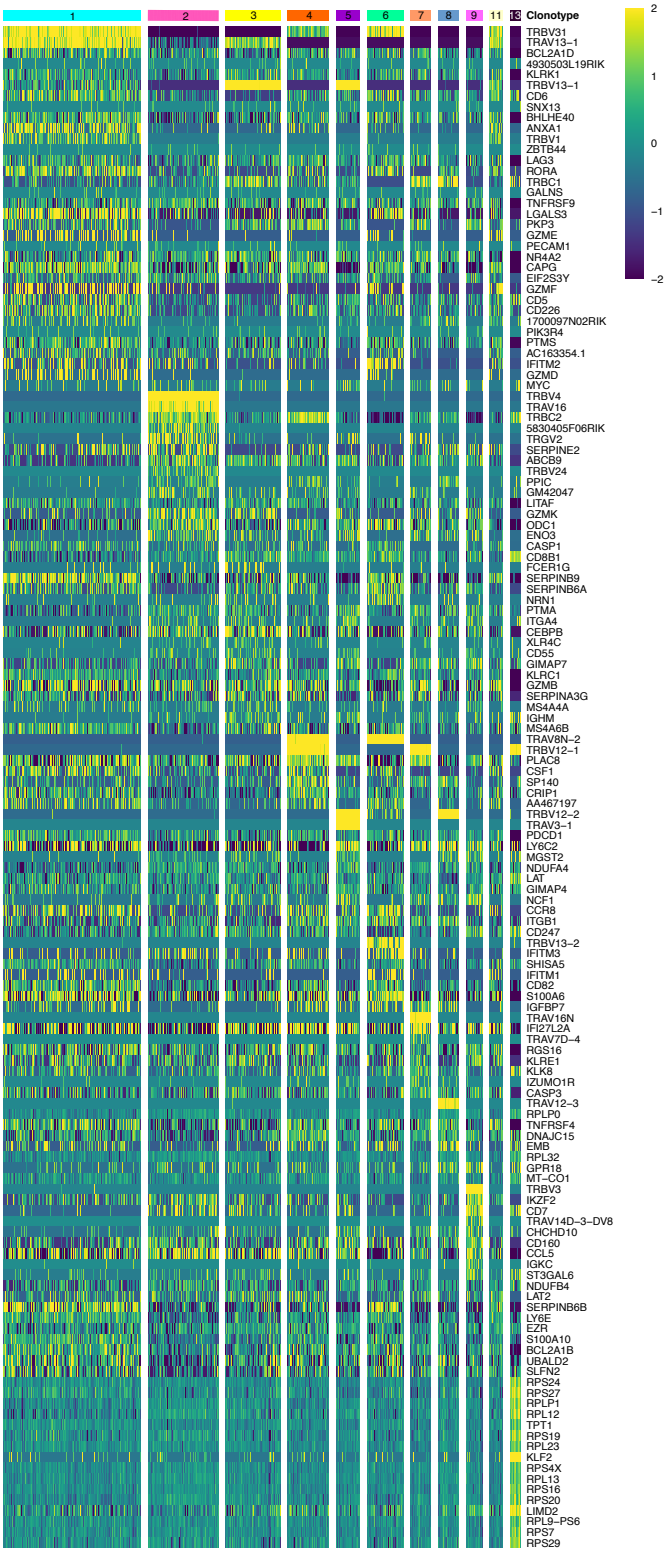


Fig. B.18 Clonotype biomarkers of *mouse_d15* calculated using the Welch t-test to identify differentially expressed genes.

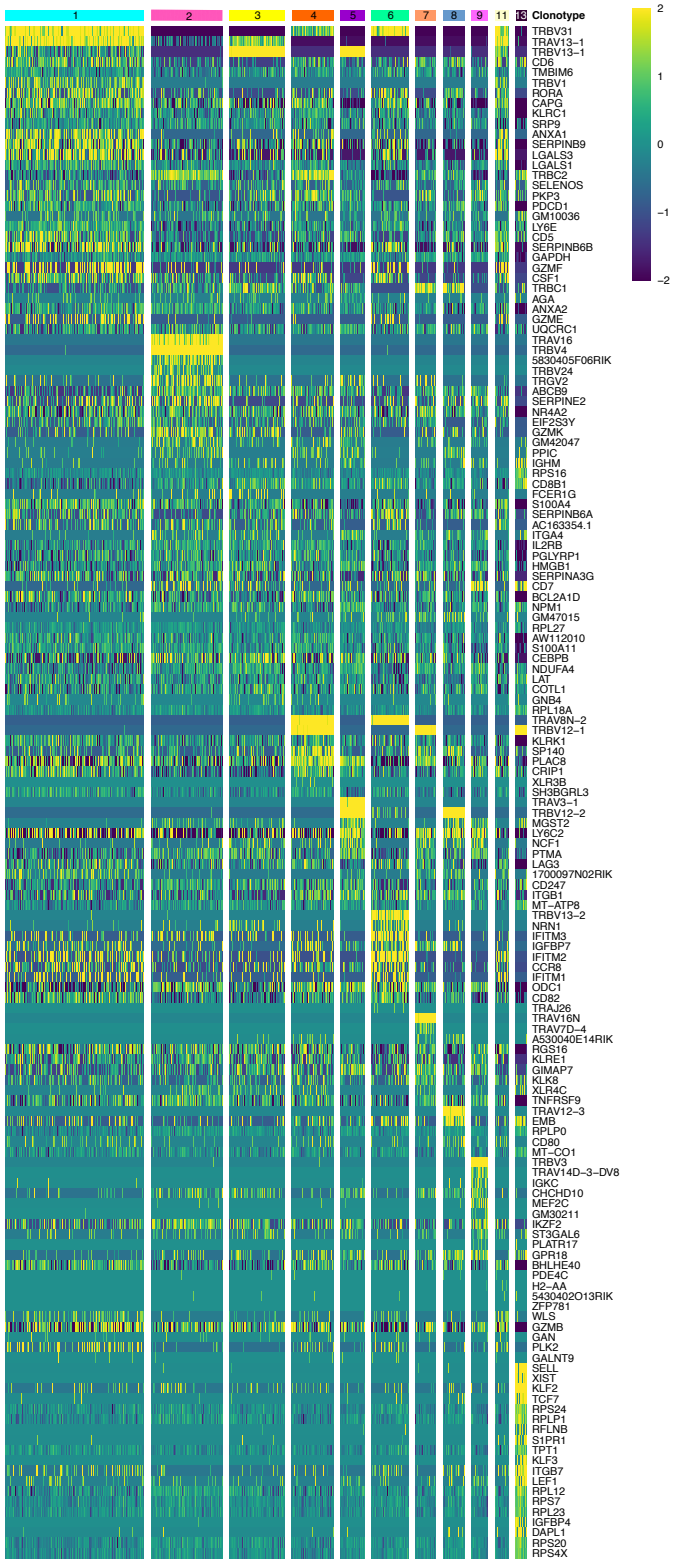


Fig. B.19 Clonotype biomarkers of *mouse_d15* calculated using the Wilcoxon rank sum test to identify differentially expressed genes.

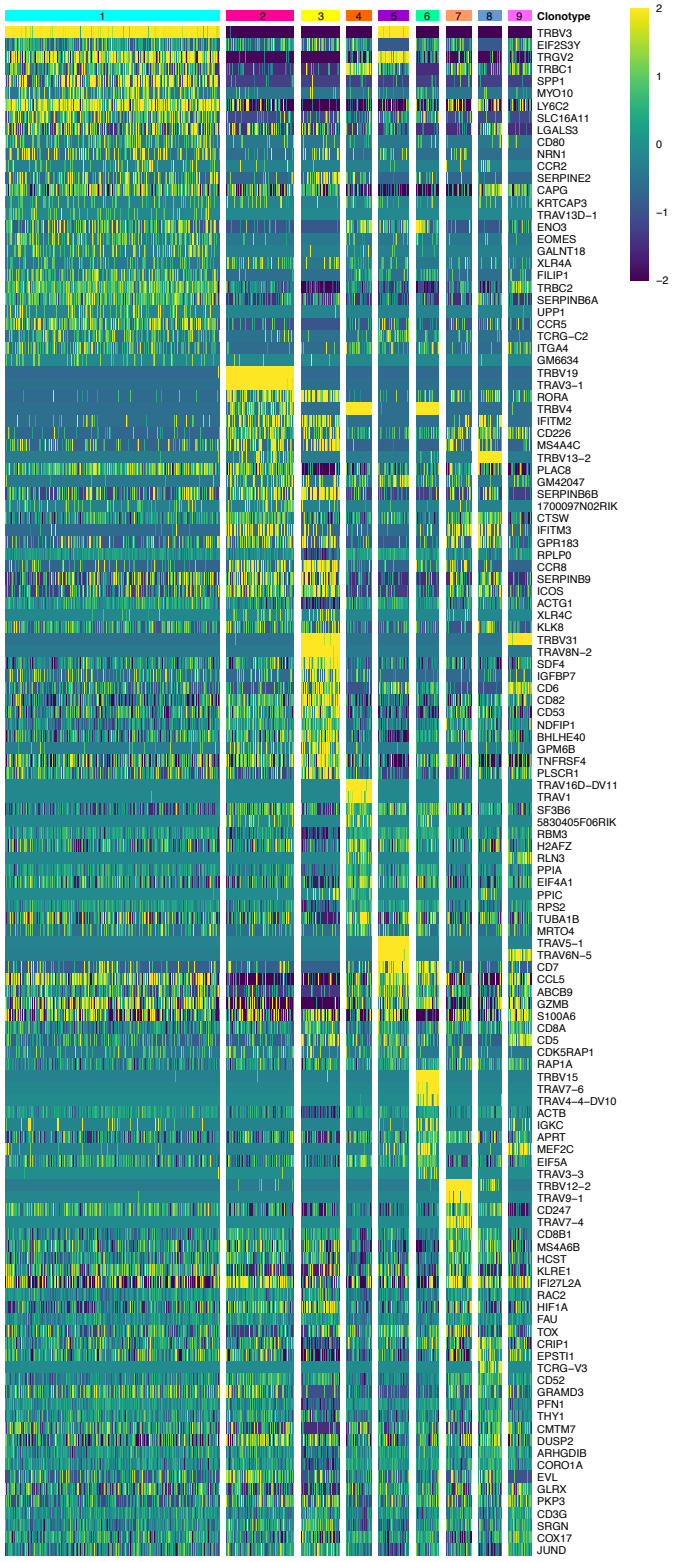


Fig. B.20 Clonotype biomarkers of *mouse_d19* calculated using the Welch t-test to identify differentially expressed genes.

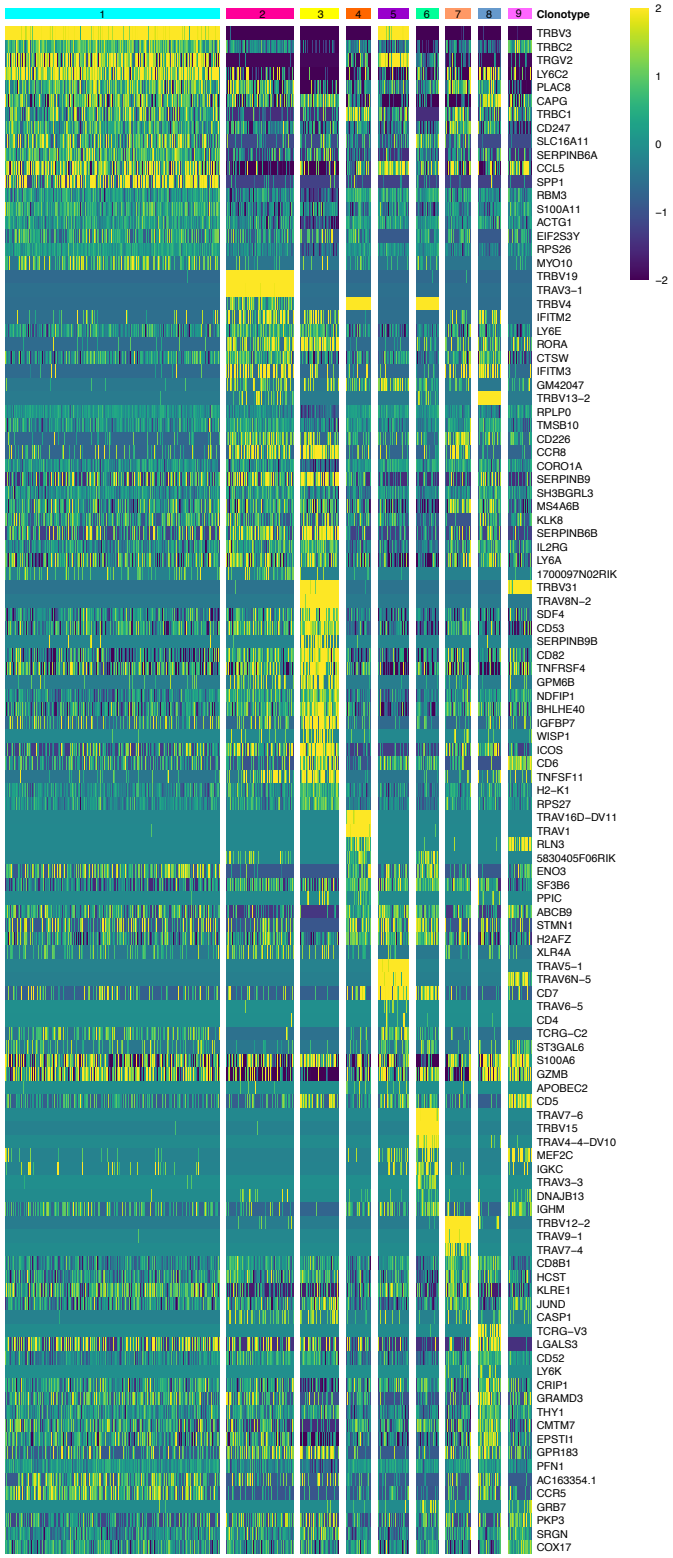


Fig. B.21 Clonotype biomarkers of *mouse_d19* calculated using the Wilcoxon rank sum test to identify differentially expressed genes.

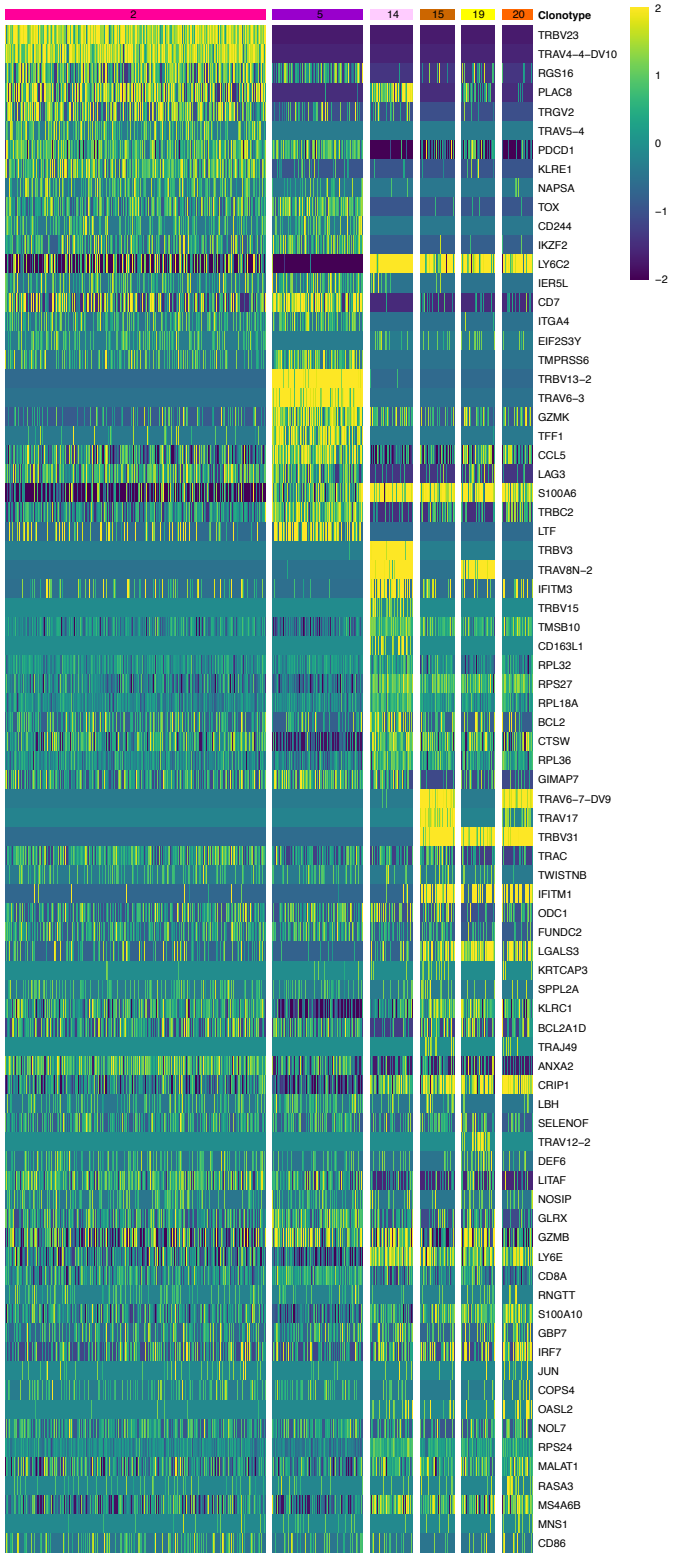


Fig. B.22 Clonotype biomarkers of *mouse_d25* calculated using the Welch t-test to identify differentially expressed genes.

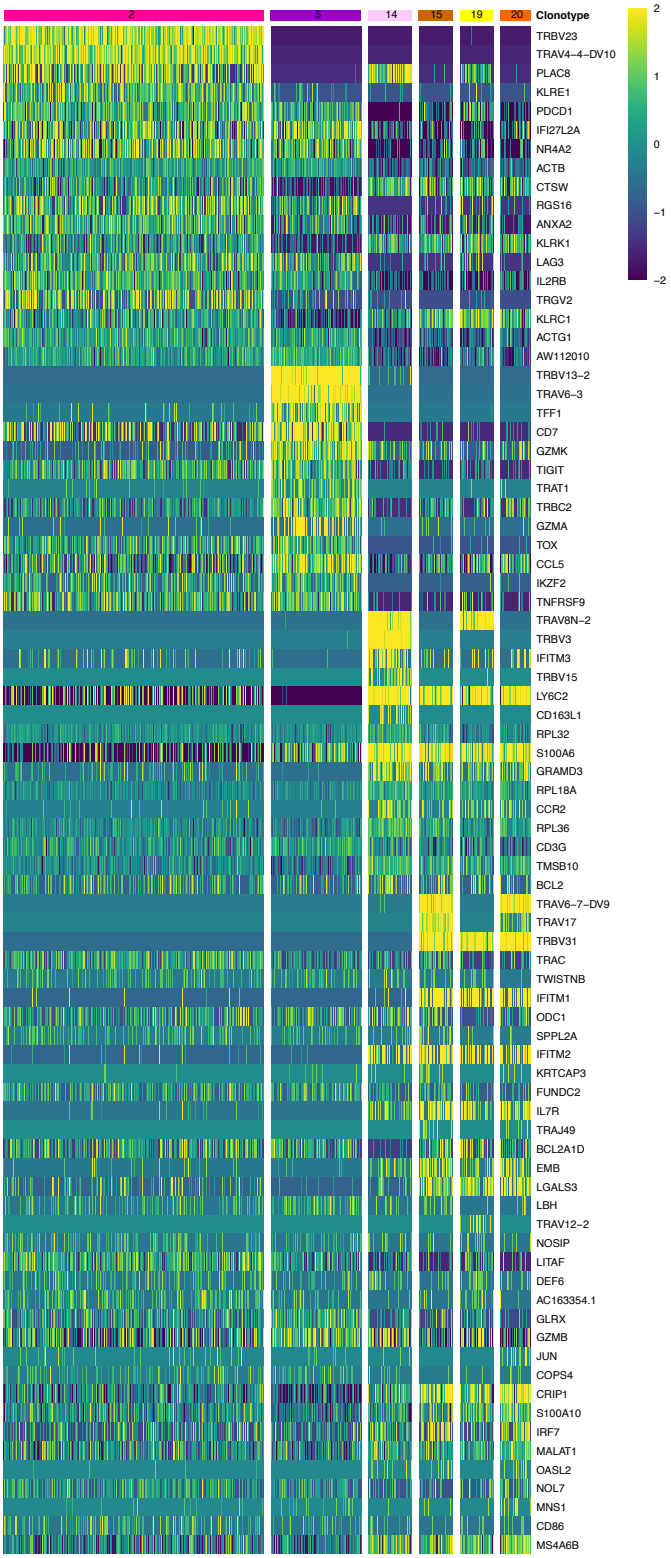
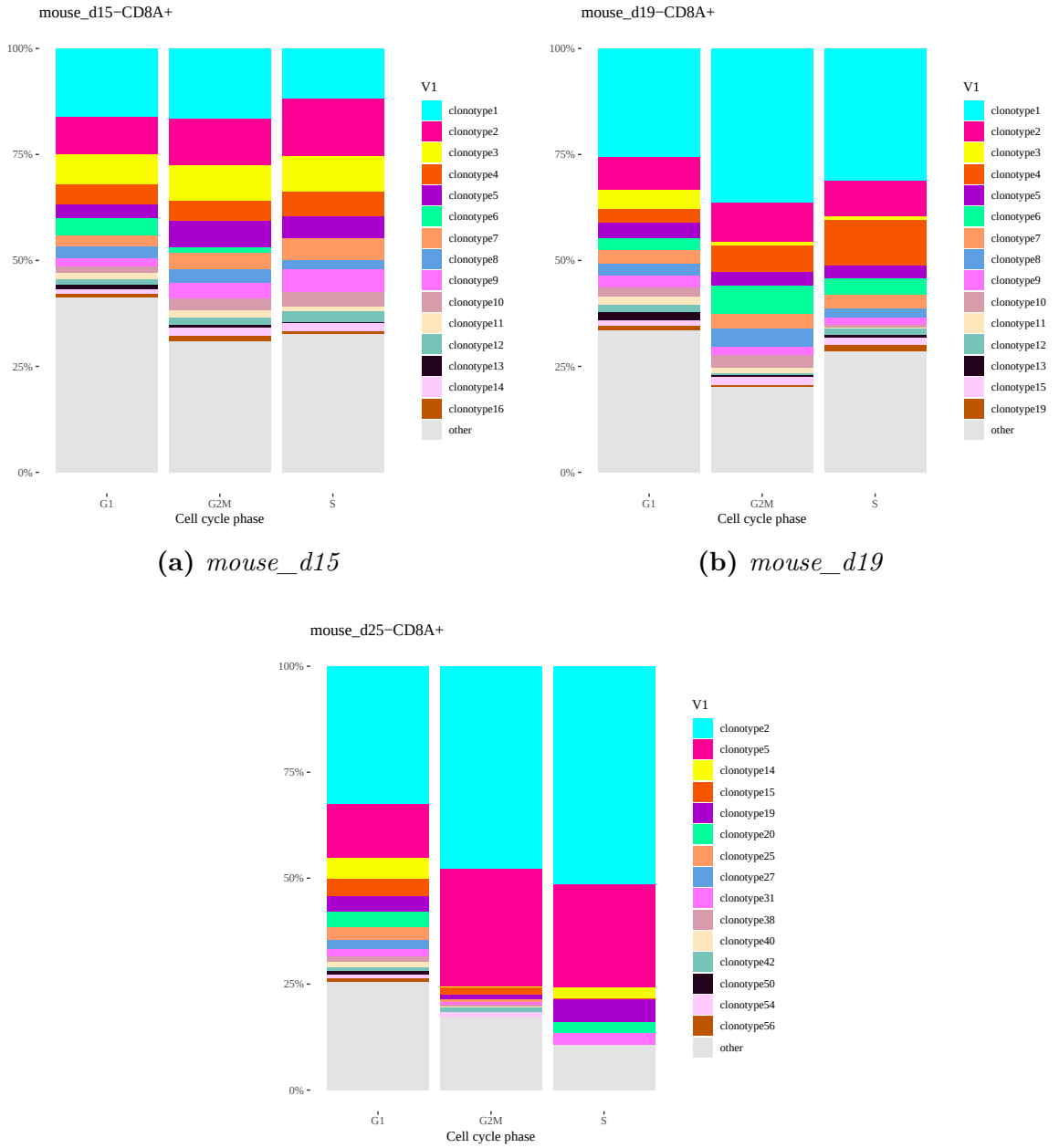


Fig. B.23 Clonotype biomarkers of *mouse_d25* calculated using the Wilcoxon rank sum test to identify differentially expressed genes.



(a) *mouse_d15* **(b)** *mouse_d19*

(c) *mouse_d25*

Fig. B.24 Clonotype proportions within cell cycle phases.

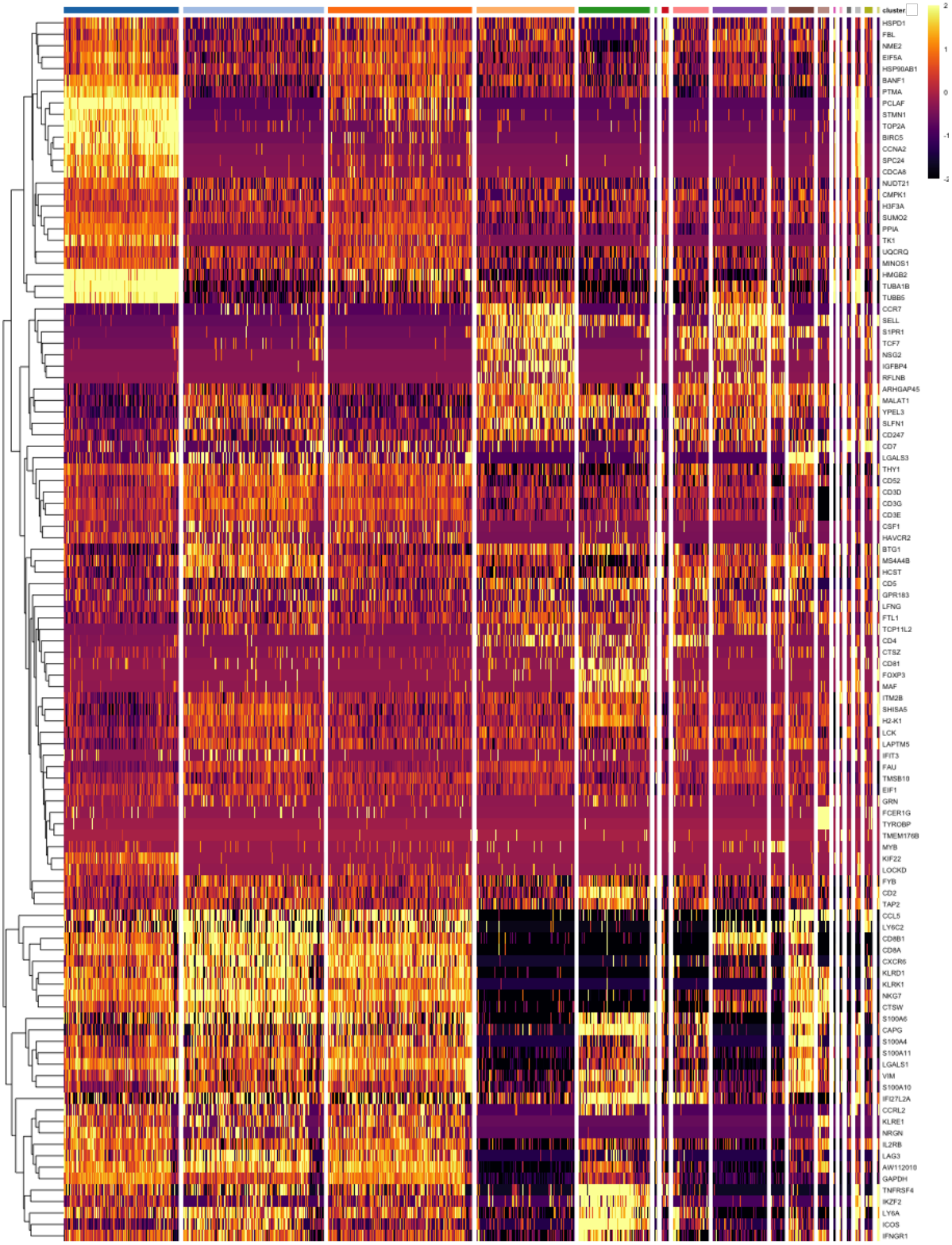


Fig. B.25 *mouse_d15* biomarkers.