Dissertation

submitted to the
Combined Faculties of Natural Sciences and Mathematics
of the
Ruperto-Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by:

Yashna Paul, M.Tech.

born in Delhi, India

Oral examination:

July 09th, 2020

Clonal Evolution Dynamics
and
Tumor Microenvironment Composition
of
Chronic Lymphocytic Leukemia

Referees: Prof. Dr. Benedikt Brors
Prof. Dr. Peter Lichter

Abstract

Chronic lymphocytic leukemia (CLL) is the most common adult leukemia in western world. This disease, with an indolent course and patients responding heterogeneously to recommended therapies, remains incurable. The Eμ-TCL1 mouse model, is a known useful tool for preclinical studies of CLL. In this thesis, I present a detailed *in-silico* view of CLL specific clonal heterogeneity and T cell tumor microenvironment (TME) as observed in spleen of Eμ-TCL1 mouse and patient lymph nodes during the course of the disease.

In the first part, I present clonal evolution orchestrated by dynamics of B cell receptor (BCR) rearrangements and somatic variations, using whole exome sequencing (WES) of serially transplanted Eμ-TCL1 mouse tumors. Low allele frequency mutations that were non-overlapping between mouse tumors were identified. 10 out of 13 tumors were identified to be oligoclonal. In addition, three distinct patterns of evolving SNV-defined and BCR clonotypes emerged as the disease aggressed from primary to secondary tumor. Interestingly, I identified stereotypic CLL mouse BCRs having *Ighv11* and *Ighv12* genes that are known to undergo chronic stimulation in response to autoantigens, hence potentially contributing to CLL pathogenesis. These observations signified the importance of clonotype information for accurate interpretation of CLL disease course and drug efficacy, especially during time-series experiments involving adoptively transferred Eμ-TCL1 mouse tumors. Also, trisomy 15 was observed, hypothesizing involvement of *Myc* overexpression during CLL development in Eμ-TCL1 mouse. It could be stated that, not just the overexpression of *Tcl1* gene but other factors also contribute to CLL malignancy in mice.

Following this, I investigated genetic (WES) and transcriptomic (RNA-seq) changes in monoclonal Eμ-TCL1 AT (adoptive transfer) mouse tumors, acquired as a result of ibrutinib resistance. Ibrutinib is widely used as a frontline treatment for CLL patients, some of which acquire resistance to the drug after showing an initial response. In mouse tumors, loss of therapeutic efficacy followed by uncontrolled tumor growth was observed at 6 weeks of treatment initiation. Ibrutinib was not able to inflict an observable selection pressure on BCR clonality as well as mutation profile of these tumors in 6 weeks. However, the transcriptional profile of ibrutinib resistant tumors was unique in contrast to untreated ones. From top upregulated genes identified to be putatively involved in ibrutinib resistance, *Tbet* gene, is currently being followed up for *in-vivo* studies as a therapeutic target.

In the second part of the thesis, I present subpopulations of CD3+ T cell compartment characteristically differentially expressed in the CLL TME as compared to that of controls. This analysis was the first of its kind to have utilized CLL patient lymph nodes (LN) for probing TME

at single cell level. Additionally, the patient's bone marrow (BM) and peripheral blood (PB); as well as the spleens from Eμ-TCL1 AT mice were investigated for CLL infiltrating T cell subpopulations.

Single cell (sc) CyTOF (mass cytometry) analysis using a panel of 32 surface protein markers revealed an increased abundance of exhausted phenotype in patient LNs as compared to BM and PB samples from the same patient. This observation raised uncertainty of PB and BM as the tissue of choice for studying CLL linked T cell exhaustion. Intriguingly, Eμ-TCL1 mouse T cell compartments showed presence of IFN-responders, absent from patient CD4+ cell type. 7 out of 12 mouse Cd4+ subpopulations showed expression of Tcytotoxic markers, which could indicate activated subpopulations.

The results presented in this thesis provide a detailed view of heterogeneity manifested by 1) Eμ-TCL1 mouse tumors in course of disease progression; 2) the transformed CLL TME in patients and mouse. These findings would prove valuable during mechanistic and drug treatment studies in Eμ-TCL1 mouse and to evaluate their translational potency in CLL clinical setting under the influence of CLL specific tumor niche.

Zusammenfassung

Chronische lymphatische Leukämie (CLL) ist die am häufigsten auftretende Leukämieerkrankung bei Erwachsenen in der westlichen Welt. Die CLL weist häufig einen langsamen Krankheitsverlauf auf und Therapieerfolge sind, abhängig vom Patienten, äußerst heterogen. Schlussendlich bleibt jedoch festzuhalten, dass CLL unheilbar ist. Das Eμ-TCL1-Maus-Mausmodell ist ein bekanntes nützliches Werkzeug für präklinische Studien zur CLL. In der vorliegenden Doktorarbeit präsentiere ich eine detaillierte *in-silico*-Ansicht der CLL-spezifischen klonalen Heterogenität und der T-Zell-Tumor-Mikroumgebung (TME), wie in Milz der Eμ-TCL1-Maus und Patientenlymphknoten im Verlauf der Krankheit.

Im ersten Teil meiner Arbeit habe ich die klonale Entwicklung der CLL analysiert, welche maßgeblich durch die Dynamik von B-Zellrezeptorrearrangements (BCR) und somatischen Mutationen beeinflusst wird. Zu diesem Zweck wurden vollständige Exomsequenzierungen (WES) von seriell transplantierten Eμ-TCL1-Mäusetumoren untersucht. Es zeigte sich, dass Mutationen mit niedriger Allelfrequenz tumorspezifisch sind. Zehn von 13 Mäusetumoren waren oligoklonal. Zusätzlich konnten drei unterschiedliche Entwicklungsmuster beobachtet werden, welche sich durch unterschiedliche SNVs und BCR-Klonotypen definierten und zu einer Progression der Krankheit führten. Interessanterweise wurden Eμ-TCL1 maustypische B-Zell-Rezeptorrearrangements unter Verwendung von zum Beispiel *Ighv11* und *Ighv12* identifiziert. Von diesen ist bekannt, dass sie als Reaktion auf Autoantigene eine chronische Stimulation erfahren. Mit diesen Ergebnissen konnte die Bedeutung von Klonotypinformationen für die genaue Interpretation des CLL-Krankheitsverlaufs und der Arzneimittelwirksamkeit, insbesondere durch die Analyse von Zeitreihenexperimenten bei denen CLL-Tumoren adoptiv übertragen wurden, hervorgehoben werden. Es wurde auch Trisomie 15 beobachtet, wobei angenommen wurde, dass die Myc-Überexpression während der CLL-Entwicklung in Eμ-TCL1-Mäusen beteiligt ist. Es kann festgestellt werden, dass nicht nur die Überexpression des *Tcl1*-Gens, sondern auch andere Faktoren zur CLL-Malignität bei Mäusen beitragen.

Im Anschluss daran, untersuchte ich Veränderungen in Genome sowie Transkriptome bei monoklonalen Eμ-TCL1 AT-Mäusetumoren (Adoptivtransfer) als Ergebnis einer erworbenen Ibrutinib-Resistenz. Ibrutinib wird häufig als Frontline-Behandlung für CLL-Patienten eingesetzt, von denen einige nach anfänglichem Ansprechen eine Resistenz gegen das Medikament entwickeln. Bei Mäusetumoren, Sechs Wochen nach Beginn der Behandlung, verlor die Therapie ihre Wirksamkeit und ein unkontrolliertes Tumorwachstum konnte beobachtet werden. Die Behandlung mit Ibrutinib war nicht in der Lage, innerhalb von 6 Wochen, einen für das Tumorwachstum nachteiligen

Selektionsdruck auf die BCR-Klonalität sowie Mutationsprofil auszuüben. Das Transkriptionsprofil von Ibrutinib-resistenten Tumoren war jedoch im Gegensatz zu unbehandelten einzigartig. Von den oben hochregulierten Genen, von denen festgestellt wurde, dass sie mutmaßlich an der Ibrutinib-Resistenz beteiligt sind, wird das *Tbet*-Gen derzeit für In-vivo-Studien als therapeutisches Ziel weiterverfolgt.

Im zweiten Teil der Arbeit untersuche ich CD3-T-Zellsubpopulationen, welche im CLL TME im Vergleich zu Kontrollen ein charakteristisches Transkriptionsprofil aufweisen. Diese Analyse ist die erste ihrer Art, die CLL-Patientenlymphknoten (LN) zur Untersuchung von TME auf Einzelzellenebene verwendet. Zusätzlich das Knochenmark (BM) und das periphere Blut (PB) des Patienten; sowie die Milz des Eμ-TCL1-Mausmodells wurden auf infiltrierende T-Zell-Subpopulationen untersucht. Die Einzelzell (SC) -CyTOF-Analyse (Massenzytometrie), unter Verwendung eines Panels von 32 Oberflächenproteinmarkern, ergab eine erhöhte Fülle des *T-Zell erschöpften* Phänotyps in LNs im Vergleich zu BM- und PB-Proben desselben Patienten. Diese Beobachtung deutete darauf hin, dass PB und BM als Gewebe zur Untersuchung der CLL-verknüpften T-Zell-Erschöpfung möglicherweise ungeeignet ist. Interessanterweise waren die Eμ-TCL1-Maus-T-Zellkompartimente Interessanterweise zeigten Eμ-TCL1-Maus-T-Zellkompartimente das Vorhandensein von IFN-Respondern, die im CD4 + -Zelltyp des Patienten nicht vorhanden waren. 7 von 12 Maus-Cd4 + -Subpopulationen zeigten die Expression von Tcytotoxic-Markern, was auf aktivierte Subpopulationen hinweisen könnte.

Die in dieser Arbeit vorgestellten Resultate bieten einen detaillierten Überblick über die Heterogenität die sich durch 1) Eμ-TCL1-Mäusetumor im Verlauf des Krankheitsverlaufs manifestiert; 2) Das transformierte CLL TME bei Patienten und Mäusen. Diese Ergebnisse tragen dazu bei, Mausstudien für potenzielle Medikamente zur Behandlung von CLL besser zu verstehen und ihre Wirksamkeit im klinischen Einsatz unter dem Einfluss einer CLL-spezifischen Tumornische besser abschätzen zu können.

Contents

List of Figures

List of Abbreviations

Gene names are written in *italics* not included in this list.

ALL	Acute Lymphocytic Leukemia
AML	Acute Myelomonocytic Leukemia
AT	Adoptive Transfer
aTregs	activated regulatory T cells
BC	Breast Cancer
BCR	B Cell Receptor
BM	Bone Marrow
BTK	Bruton Tyrosine Kinase
CCA	Canonical Correlation Analysis
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
CMV	Cytomegalovirus
CNV	Copy Number Variations
controlLN	Control Lymph Node
CS	Class Switching
CyTOF	Mass Cytometry
EBV	Epstein-Barr virus
Eμ-TCL1	Enhancer(mu) - T Cell Leukemia 1
FCS	Flow Cytometry Standard
FDC	Follicular dendritic cells
GEM	Gel Bead in Emulsion
HSP	Heat Shock Proteins
I.P.	Intraperitoneal
Ib-L	Ibrutinib Late
Ighv	Immunoglobulin Heavy Chain
IL	Interleukins
Klr	Killer-cell Lectin
KNN	K-nearest Neighbour
LIH	Luxembourg Institute of Health
LN	Lymph Nodes
MDC	Monocyte-derived cells
MDS	Multi-Dimension Scaling
MHC	Major Histocompatibility Complex
MNN	Mutual Nearest Neighbour
NCP	National Cytometry Platform
NGS	Next Generation Sequencing
OS	Overall Survival
PB	Peripheral Blood
PCA	Principal Component Analysis
PD-1	Programmed Cell Death Protein 1

PFS	Progression Free Survival
PtC	Phosphatidylcholine
rLNs	reactive Lymph Node Samples
rTregs	resting regulatory T cells
sc	Single Cell
SHM	Somatic Hypermutation
SNV	Single Nucleotide Variation
Tc	Cytotoxic T cells
Tcm	Central Memory T cell
TCR	T cell receptors
Teff/Tef	Effector T cell
Tem	Effector Memory T cell
Tex	Exhausted T cell
Tfh	Follicular Helper T cells
Tg	Transgenic Mice
Th	Helper T cells
Tm	Memory T cells
TME	Tumor Microenvironment
Tnaive	Naïve T cell
TOF	Time of Flight
Tregs	Regulatory/Suppressor T cells
tSNE	t-distributed Stochastic Neighbor Embedding
tumorBM	Tumor sample from Bone Marrow
tumorLN	Tumor sample from Lymph Node
tumorPB	Tumor sample from Peripheral Blood
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique Molecular Identifier
V(D)J	Variable (Diversity) Joining
VAF	Variant Allele Frequency
Ve-E	Vehicle Early
Ve-L	Vehicle Late
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WHO	World Health Organization
WT	Wild Type

1. Introduction

1.1 Cancer

Cancer is the uncontrolled, abnormal growth and division of cells. These cells eventually infiltrate surrounding normal tissue and may spread to other parts of the body through the blood or the lymphatic system. It disrupts normal functioning of the body including the immune system's ability to counter the tumor. Presently cancer is diagnosed in more than 10 million new people every year. It is the second leading cause of death in the world after cardiovascular disorders (1, 2). In Europe alone, there were an estimated 3.9 million new cases of cancer and 1.9 million cancer related deaths in the year 2018. The most common cancers in Europe in the same year were that of breast, colorectal, lung and prostate (3). In addition to reducing the quality of life for the patient, cancer and its treatment result in financial losses, morbidity and premature death.

There are more than 200 types of known cancers. Broadly, they can be differentiated into 5 types based on the cell of origin. They are carcinomas (skin, tissue lining), sarcomas (connective tissue like muscle etc.), leukemias (blood or bone marrow), lymphomas and myelomas (lymphatic tissues and bone marrow respectively), and, brain and spinal cord (central nervous system). Tumors are also classified according to their degree of potency into benign and malignant. Benign tumors are slow growing, do not spread to other parts of the body and are non-cancerous. On the contrary, malignant tumors grow much faster, metastasize and are cancerous.

Risk factors that contribute to occurrence of cancer include aging, tobacco usage (smoking), prolonged exposure to sun (UV rays), internal exposure to radioactive materials and harmful chemicals persistent in polluted air and water, certain viruses (EBV, HPV, hepatitis B and hepatitis C virus etc.) and bacteria (H. pylori), hereditary and non-hereditary genetic mutations, over consumption of alcohol, lack of physical activity, and obesity (4). Several causes of cancer are preventable (tobacco usage, modifications in diet, controlled alcohol usage), whereas a family history of cancer or aging are unavoidable risk factors.

1.2 Hallmarks of Cancer

To sustain themselves, cancer cells must acquire several essential biological capabilities. These traits are shared between most cancers and drive their transformation from normal to malignant. In addition, these biological hallmarks capacitate cancer cells to acquire a tumor growth facilitating niche, become malignant and eventually metastatic. Figure 1.1 depicts eight groups into which the acquired complexities of cancer developing cells can be divided (Hanahan and Weinberg 2011).

Figure 1.1 Hallmarks of cancer (5).

Two factors that form the basis of these inherent properties of all cancers are genome instability and inflammation. *Genome instability* refers to increased tendency of acquiring mutations by the cancer cells over successive cell divisions in order to survive immunological control systems and evolve. Genome integrity is kept in check by DNA damage and repair machinery, and mitotic checkpoints. Defects in these processes predispose the DNA to genomic alterations including chromosomal aberrations and DNA strand breaks. Tumor heterogeneity can be a result of genomic instability (Yao and Dai 2014). *Chronic inflammation* induced as a result of viral or bacterial infections or even non-infectious agents can contribute to cancer development. Inflammation driven carcinogenesis promotes other hallmarks of

tumor development like escape of apoptotic signals, enhancing signals for angiogenesis and metastasis, etc (Multhoff, Molls et al. 2011).

These integral underlying components of the process of cancer development have provided a framework for understanding the biological complexities in the way of treating this disease. Owing to years of established research and understanding of underlying mechanisms of cancer phenotypic manifestations, treatments and medications are available for many cancer types. Early detection and advanced clinical interventions have provided successful remission including improved prognosis and overall survival for patients suffering from cancers of the breast, prostrate, thyroid, melanoma and cervix. On the other hand, cancers with worst survival include ones of the central nervous system, pancreatic cancer, lung cancer, gall bladder cancer and esophageal cancer with less than 20% patients surviving after 5 years of treatment (6). However, this number can vary even for the subtypes of same kind of cancer. For example, different types of leukemias have a varied percentage of patients surviving beyond 5 years post treatment (acute myelomonocytic leukemia (AML): 24% patients, chronic myeloid leukemia (CML): 66.9% patients, acute lymphocytic leukemia (ALL): 68.2% patients, chronic lymphocytic leukemia (CLL): 83.2% patients) (7).

Even though CLL has a better 5-year survival rate as compared to other leukemias, it remains the most common leukemia in western countries. Many CLL patients relapse by becoming resistant to available chemotherapy and antibody treatment options due to acquired mutations in *BTK* and *PLCG2* genes after the first treatment. Relapsed or progressive CLL is not curable except by allogenic stem cell transplantation (8). These patients however do respond to repeated palliative treatments, that prolong their life. It is well known that the course CLL disease development can vary between patients. Research in CLL is therefore focused to identify underlying genetic mutations, mechanisms and study therapeutic responses in as many patients as possible by enrolling them in clinical trials. The aim is hence to avoid relapse and disease progression into aggressive form.

CLL is the cancer of study in this thesis. The following sections describe it in detail, followed by the caveats in our knowledge of CLL and its treatment.

1.3 Chronic Lymphocytic Leukemia (CLL)

CLL is a slow growing malignancy characterized by the accumulation of CD19+ CD25+ CD23+ B cells in blood, bone marrow, spleen and secondary lymphoid organs. The median age at diagnosis is 70 years with the disease course as well as the genetic aberrations being quite heterogenous between patients. The indolent form of this disease generally goes undetected unless asymptomatic lymphocytosis is reported during incidental blood tests. Patients with progressive disease need treatment as they possess increasing lymphocyte count, adenopathy and hepatosplenomegaly. These patients also show infiltration in the bone marrow that may result in bone marrow failure and subsequent anemia and thrombocytopenia. To be diagnosed with CLL, a patient must possess at least 5,000 B cells per microliter in the peripheral blood (Garcia-Munoz, Galiacho et al. 2012).

Clinical attributes of CLL mentioned above are the basis for the following classification systems for prognosis and treatment (Grabowski, Hultdin et al. 2005):

1. Rai staging, and
2. Binet staging

In addition to the above, several biological and genetic markers are also of prognostic value. For example, chemoimmunotherapy resistant patients often manifest deletion in chromosome 17p and/or mutations in *TP53*. These patients constitute about 7% of relapsed CLL cases (Hallek 2019). 55% of CLL cases show recurrent deletions in chromosome 13q that includes the putative CLL driver *DLEU2*. Deletion in chromosome 11q with driver genes *ATM* and *BIRC3* and trisomy 12 make up 6%-18% and 12%-16% recurrent CLL cases respectively (Guieze and Wu 2015). Other genetic lesions reported in CLL that also aid in prognosis of the disease include mutations in *NOTCH1*, *SF3B1*, somatic hypermutation (SHM) in the *IGHV* gene and upregulation of cell surface markers CD38, CD49D, ZAP70 (Gaidano 2017).

In all, only the above recurring mutations of importance occur at a frequency of >5% in CLL. Majority of the mutations that are identified in CLL patients are biologically and clinically uncharacterized and occur at lower frequencies. This means there are few clonal and mostly sub-clonal mutations in CLL. The overall low somatic mutation rate in CLL (approx. 1/Mb) as

compared to UV/carcinogen induced tumors like melanoma, and the indolent course of the disease imply vastly heterogenous nature of CLL, without a universal genetic event common to all patients. The rate of inter- (between patients) as well as intra- (within the same tumor) tumoral heterogeneity have been reportedly high in CLL; and is dynamic throughout the course of the disease. This has clinical implications such as treatment failure or resistance (Guieze and Wu 2015).

Additionally, the BCR signaling pathway plays an important role in assessing CLL disease pathogenesis and is exploited as a therapeutic target. Other pathways of importance also influenced by CLL microenvironment include WNT signaling, NOTCH1 signaling, NF-kB nexus and CXCR4/CXCL12 signalling (Ferrer and Montserrat 2018).

1.4 B Cell Receptor (BCR) Clonal Dynamics in CLL

There are strong evidences that CLL arises due to chronic stimulation of the B cell receptor (BCR). The general process includes selection and expansion of the malignant clone resulting in B cells with aggressive BCR. Structurally BCRs are proteins with an N-terminal variable region that binds to the antigen and a C-terminal constant region with effector functions. Genetically the BCR is composed of segments of immunoglobulin V (variable), D (diversity), J (joining) and C (constant) genes. Functional BCR protein is formed when individual V, D and J segments undergo recombination during B cell maturation. These genes are highly polymorphic and are responsible for the germline BCR diversity of an individual. Clonal expansion occurs after a B cell recognizes an antigen, which is followed by SHM in the germinal centers and class switching (CS). In CLL, clonal expansion of B cells can occur both before and after SHM, causing accumulation of clonally related CD19+ CD5+ IgM+ IgD+ B cells.

Certain *IGHV* gene rearrangements are more common in CLL and are known as stereotypic clones, e.g. *IGHV3-21* is identified in approximately 30% CLL cases. The reason for stereotypic V(D)J rearrangements in CLL could be attributed as a response to common antigens or a shared mechanism of clonal expansion of CLL B cells that drives growth of the malignant clone (Petrova, Muir et al. 2018). Potential of expanded CLL clones to undergo SHM also makes way for the hypothesis that there could be modes other than antigenic stimulation for

malignant clonal expansion (Efremov 1996, Fais 1996). Such type of B cell response highly depends upon the sequence specific features of variable genes in V(D)J rearrangement (Duhren-von Minden, Ubelhart et al. 2012).

Sequencing BCR repertoires could hence aid in tracking and delineating evolution of B cell responses in CLL by characterizing the diversity or types of identifiable V(D)J rearrangements. V(D)J clonotype information can then be associated to CLL progression to identify aggressive BCRs. This can help clinicians in predicting disease course and treatment response for patients harboring similar clonotypes.

Eμ-TCL1 preclinical mouse model manifests a CLL like disease by overexpression of *Tcl1* gene placed downstream of the promoter sequence *Ighv* gene locus in mouse. This is explained later in detail. B cell repertoire studies in Eμ-TCL1 CLL mouse model have shown that both early generated B-1 B cells and stereotypic clonotypes contribute to the progression of CLL in mouse (Hayakawa, Formica et al. 2016). However, in general there is lack of evidence on the diversity of usage of V(D)J genes that leads to disease evolution and accelerated progression with each adoptively transferred (AT) CLL Eμ-TCL1 mouse. Understanding how BCR clonotype evolution correlates with CLL progression in mice is important for preclinical studies of CLL.

In addition, it has been observed that BCRs also evolve in response to microenvironment signals which are different in leukemia as compared to healthy individuals (Burger 2011). A leukemia supportive microenvironment probably signals evolution of relatively aggressive BCRs that can produce a sustained response adding to tumorigenicity.

Hence, to study evolution of CLL in patients and Eμ-TCL1 mouse model, it is very important to understand the dynamics of the disease microenvironment.

1.5 Tumor Microenvironment (TME) of CLL

CLL has a tumor supportive microenvironment that aids in homing, survival and proliferation of malignant B cells. Components of the CLL TME and the signaling mechanisms they employ to interact with B cells, could hence be potential targets to counter CLL progression. Components of CLL TME and the mechanisms by which they interact with CLL cells are shown in figure 1.2 adapted from (Ten Hacken and Burger 2016). Briefly, they are bystander cells like T cells, MDCs (monocyte-derived cells) and stromal cells like endothelial cells, pericytes and FDCs (follicular dendritic cells). Signals from these cells result in an immunotolerant milieu in CLL lymph nodes. In addition, such a niche promotes sustenance of neo-antigen expressing malignant B cells.

The communication between B cells and cells of the TME is regulated by:

1. *Interleukins (ILs)*: e.g. IL-4 and IL-21- promoting cell survival and proliferation; and 1L-10 facilitating immunosuppression
2. *Chemokines*: e.g. CCL2, CCL3, CCL4, and CLL22 are involved in chemo-attraction of cell towards the TME
3. *Growth factors*: e.g. IGF-1 (insulin-like growth factor one) promotes survival
4. Membrane bound factors on bystander cells aiding in cell survival like CD40L and integrins.
5. *Micro vesicles* and exosomes that are produced by both CLL and bystander cells help in signal transmission.
6. *Nucleoside adenosine*: renders the tumor immune niche useless leading to tumor resistance in CLL cells.

Since in this thesis I try to understand altered functional components of T cells from the CLL TME I review here normal T cell functions, subsets and then their established role in degenerating immune response in CLL including T cell exhaustion.

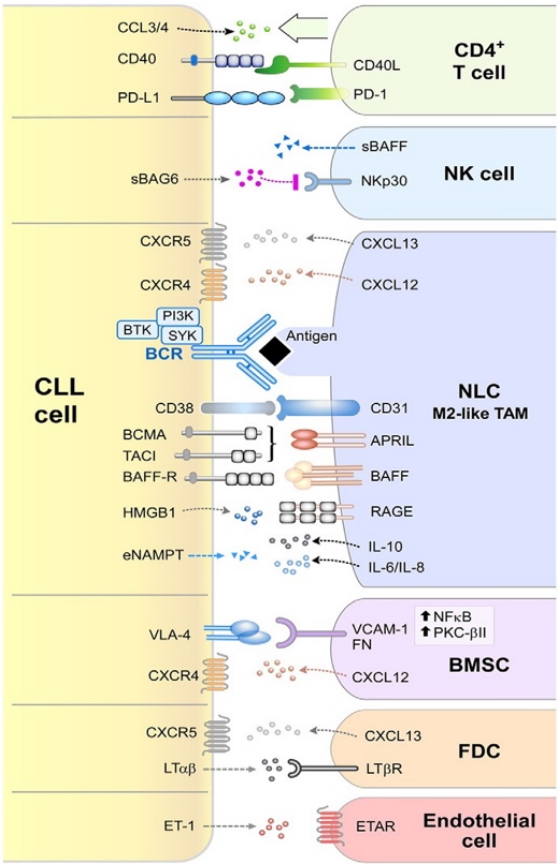


Figure 1.2: Interactions between B cells and the cells of the microenvironment in chronic lymphocytic leukemia (Ten Hacken and Burger 2016). NLC = nurse like cells, NK = natural killer cells, BMSCs = bone marrow stromal cells, FDCs = follicular dendritic cells.

1.6 T Cell Subsets and their Functions in the TME

The TME of CLL is characterized by increased numbers and compromised functionality of T lymphocytes (both CD4+ and CD8+). There are several types of T cells functionally active in the TME and they are defined briefly below. Since each subpopulation of T cells can be identified by specific markers, after understanding their role in the CLL TME, specific subpopulations could be potential targets for therapy.

T lymphocytes arise in the bone marrow and mature in the thymus. The spectra of responses generated by T cells are together called as cell-mediated immune responses. T cell receptors (TCRs) on T cells recognize antigen presented to them by MHC (major histocompatibility complex) molecules. In addition to the TCR, T cells express either the CD8 or the CD4

glycoproteins on their surface, and are then called cytotoxic or helper T cells, respectively. Based on the type of surface marker they express and function, there are four types of T cells (9):

1. *Cytotoxic T cells (Tc)*: Also called killer T cells, they express CD8 marker and can induce an infected cell to undergo apoptosis, hence killing it.
2. *Helper T cells (Th)*: These are CD4+ T cells, and they proliferate to activate B cells or CD8+ T cells. They function after transformation to effector cells upon activation.
3. *Memory T cells (Tm)*: They can be both CD4 and CD8 cells and help induce quick secondary responses upon interaction with the same antigenic stimuli that has also been experienced earlier.
4. *Regulatory/suppressor T cells (Tregs)*: as the name suggests, they help generate controlled immune T cell responses and avoid self-damage.

Characterized by their unique cytokine profiles, CD4+ T cells can also be divided into Th1, Th2, Th9, Th17, Th22 and Tfh (follicular helper T cells).

T cells get activated upon pathogenic antigen presentation to the naïve T (Tnaive) cell. These activated T cell proliferate and differentiate into effector T (Teff/Tef) cells, which get recruited to the site of infection to eliminate pathogens. Although the life of Teff cells is short, a subset transitions into long term memory cells, that retain memory of the pathogen in case of restimulation. Memory T (Tm) cells can either be in secondary lymphoid organs and there they are called central memory T (Tcm) cells; or they can be located at the site of freshly infected tissue and are there called as effector memory T (Tem) cells. Upon re-exposure to the same antigen, Tm cells can produce an immune response that is faster and stronger than the primary immune response. The unique features of Tm cells are hence, producing a more effective secondary immune response and antigen-independent self-renewal driven by IL-7 and IL-15, and development of memory after absence of ongoing antigenic stimulation. The process of CD4+ and CD8+ T cell differentiation is shown in figure 1.3. Tregs are the components unique to CD4+ T cells.

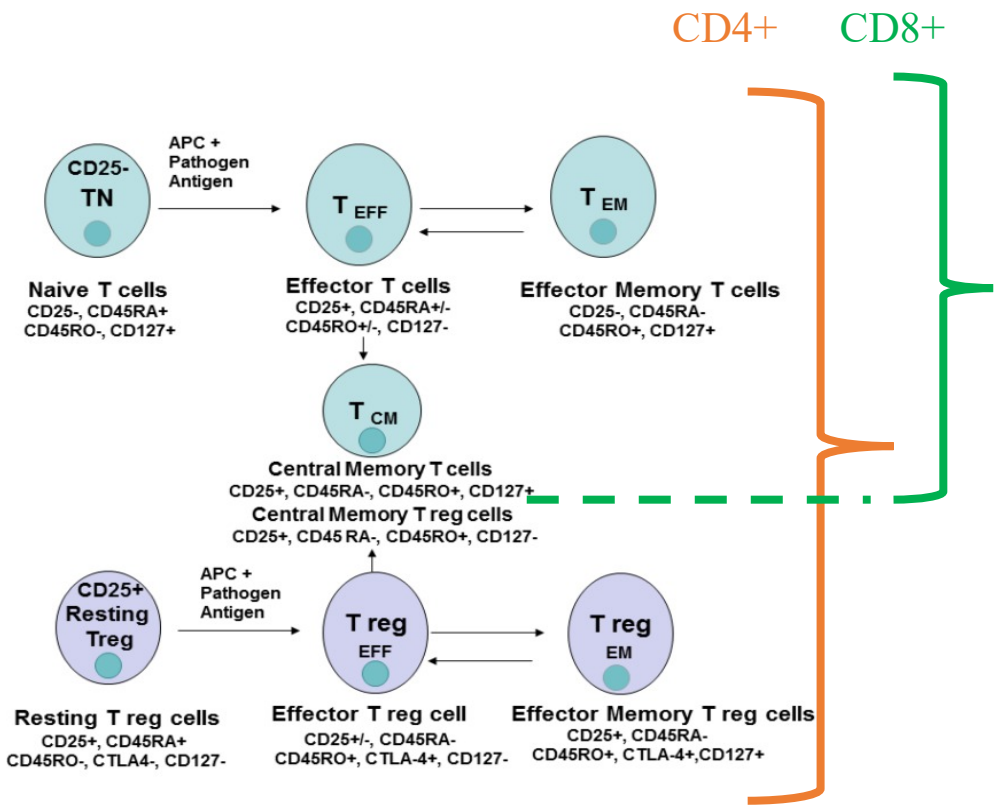
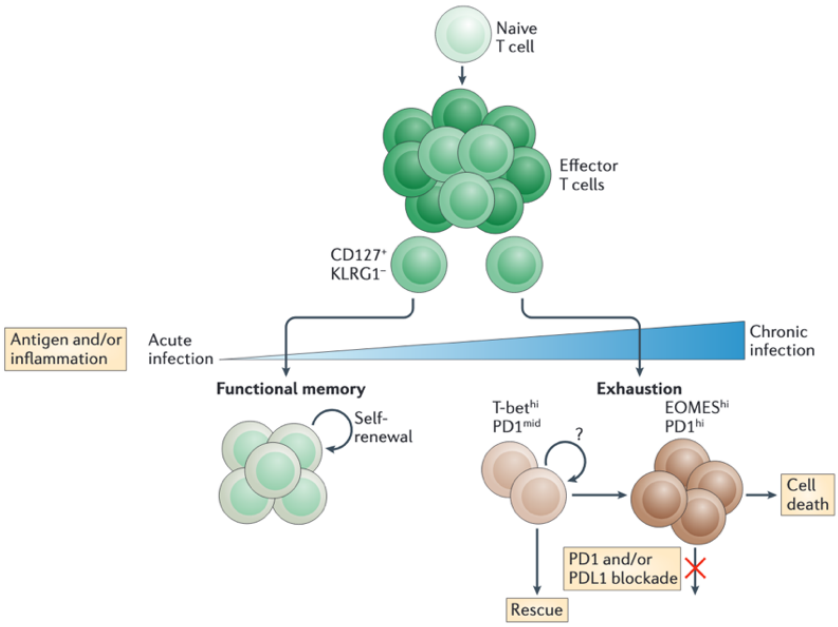


Figure 1.3: Differentiation process of CD4+ and CD8+ T cells into Teff, Tem, Tcm on encountering an antigen. Modified from (Golubovskaya and Wu 2016).

1.7 T Cell Exhaustion and its Impact on CLL

Persistent antigenic stimulation in chronic diseases and cancers, alters the differentiation mechanism of Tm cells, causing Teffs to become exhausted T (Tex) cells instead. Exhausted T cells have a transcriptionally distinct state as a result of stepwise loss of effector functions, constant upregulation, expression of inhibitory receptors (*PDCD1*, *LAG3*) and use of certain key transcription factors (*MAF*, *TOX*, *EOMES*). These altered effector cells, are unable to transition into memory state after absence of infection and are chronically stimulated (Wherry and Kurachi 2015). Figure 1.4 entails the differences in surface marker expression, cytokine production, proliferation and antigen dependency of the memory and exhausted T cell states. The table under the figure reports that exhausted cell types have characteristic reduced proliferation and cytokine production, increased expression of inhibitory receptors and no capacity of self-renewal. Exhaustion hence seems like a roadblock in the process of acquired immunity.



Feature	Functional memory T cell	Exhausted T cell
Proliferative potential	+++	+/–
Cytokine production	+++	+/–
Memory markers (e.g. CD44, CD62L, CD127 or CXCR3)	+++	+/–
Inhibitory receptors (e.g. PD1, LAG3, CD160 or 2B4)	–	+++
IL-7- and/or IL-15-driven self-renewal	++	–
Antigen dependency	–	+

Figure 1.4: Difference in surface marker expression of Tm cell and Tex cell during acute and chronic infections respectively (Wherry and Kurachi 2015).

If the infection persists way longer as is usually the case with cancers, effector T cells that were once functional against the cancer are completely lost. All that is left is inhibitory markers expressing exhausted T cells, incapable of producing cytokines like IFN-γ, TNF (tumor necrosis factor), and IL-2 (Wherry 2011).

Loss of effector functions renders the immune niche of an infection or a tumor (in case of cancer) functionless. However, the exhausted immune environment can be rescued by controlling and modulating inhibitors and pathways over expressed in exhaustion. The PD1-PD-L1 nexus is well explored with respect to rescuing effector functions of chronically stimulated T cells. PD-1 (programmed cell death protein 1) is an inhibitory receptor present on the surface of activated T cells. Its function is to assure controlled immune response and avoid chronic autoimmune inflammation. When PD1 binds to its ligand PD-L1, it signals attenuation of T cell functions leading to T cell depletion. T cells respond against cancer cells

via effector functions, and this also leads to binding of PD-L1 expressed in high amounts on cancer cells to PD1 on effector T cells (figure 1.5 a). Once this happens, the effector T cells lose their functions. To rescue this anti-PD1/anti-PD-L1 antibodies have been proposed that block either PD1/PD-L1 epitopes and keep the T cell effector functions intact (figure 1.5 b). This results in resumed immune response against the tumor (Marchetti, Di Lorito et al. 2017, Angelousi, Chatzellis et al. 2018).

Anti-PD1 therapy has shown promising results against melanoma (stage III/IV), metastatic renal cell carcinoma, refractory Hodgkin's lymphoma, chronic lymphocytic leukemia and ovarian cancer to name a few. Anti-CTLA4 (e.g. Ipilimumab) in combination with anti-PD1 therapy (Nivolumab) has been effective in treating melanoma (state III/IV), and non-small cell lung cancer (Seidel, Otsuka et al. 2018). Anti-LAG3 also holds potential for patients with CLL, gastric cancer and prostate cancer among others (Long 2018). These therapies could be used individually or in combination for best results. We as well as others have shown that anti-PD-L1 (alone or in combination with ibrutinib monotherapy) as well as dual PD1/LAG3 blockade prevents immune dysfunction and suppresses leukemia in the CLL preclinical mouse model Eμ-TCL1 (Wierz 2018, Hanna, Yazdanparast et al. 2020).

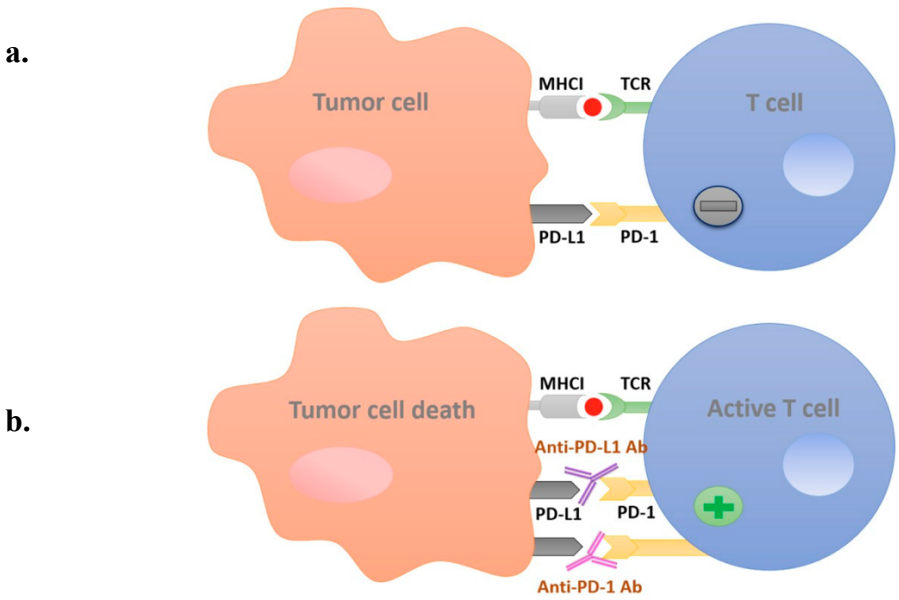


Figure 1.5: (a) PD-L1 on cancer cell interacts with PD1 on T cell to inhibit its activity (b) Anti-PD1 and Anti-PD-L1 antibodies block the epitope of PD1. This hinders the interaction of PD1 with PD-L1 on the cancer cell. This leads to a potential increase in T cell effector functions and immune response against the tumor (Abdin, Zaher et al. 2018).

In our group there have been studies on the dynamics of the PD1-PDL1 pathway in CLL tumor microenvironment in the Eμ-TCL1 mouse model (McClanahan, Riches et al. 2015). Researchers from our group have shown that effector CD8+ T cells from the CLL TME of both Eμ-TCL1 mouse spleens and patient lymph nodes and peripheral blood, are composed of two phenotypically and transcriptionally distinct populations separated by high ($PD1^{hi}$) and intermediate ($PD1^{int}$) expression of the *PD1* gene. The subpopulation with $PD1^{hi}$ expression was found to resemble exhausted T cells with respect to compromised cytotoxicity and increased expression of inhibitory receptors like TIGIT and LAG3 and transcription factors like EOMES. However, CD8+ $PD1^{int}$ population still displayed potential for effector functions. It was observed that the balance between $PD1^{hi}$ and $PD1^{int}$ CD8+ T cell populations was tightly regulated by 1L10/STAT3 signaling; which when blocked shifted the proportion of $PD1^{int}$ population towards $PD1^{hi}$. Hence IL10 was associated with shifting the CLL TME towards an immunocompromised one (Hanna 2020). This mechanism was hence suggested as a potential therapeutic target for CLL.

Effective therapies and their combinations have been useful for the treatment of CLL. Some of them target the highly upregulated BCR signaling pathway, and others work on the components of the tumor microenvironment to combat CLL. Presently, research on CLL involves deciphering the effectiveness of these therapies individually and in combination, both in the Eμ-TCL1 mouse as well as in CLL patients.

1.8 Treating Chronic Lymphcytic Leukemia

There is a plethora of treatment options available for chronic lymphocytic leukemia. Parameters like the stage and grade of CLL, symptoms manifested by the patient, previous infections/fitness of the patient, previous response to a treatment, remnants from previous treatment, whether it's a primary or relapse disease state, availability of the drug in consideration and its economic burden on the patient (Burger 2011), all affect the choice of treatment for a patient.

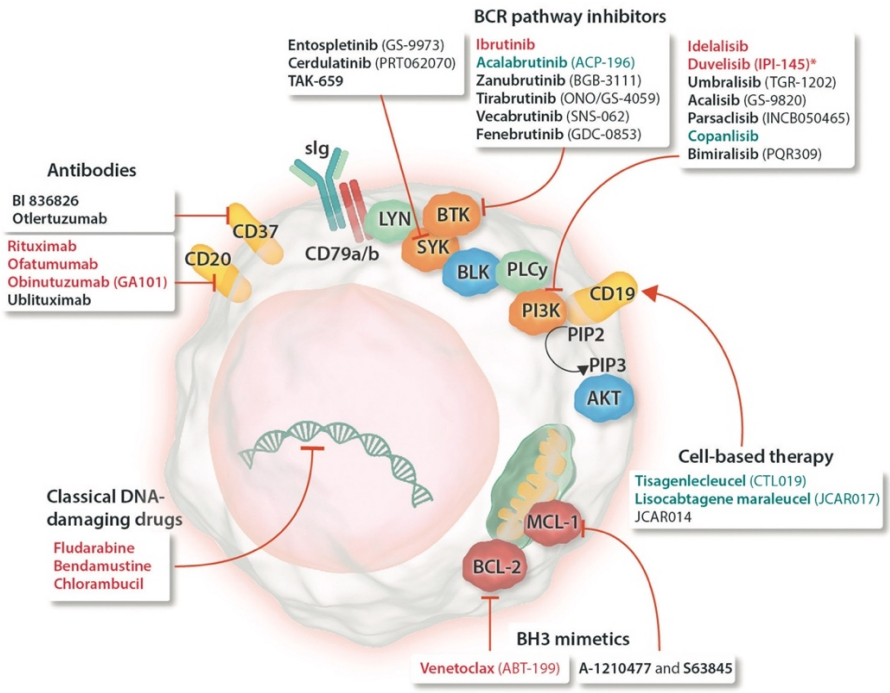


Figure 1.6: Treatment options for CLL and the molecule/pathway they work on to inhibit progression of chronic lymphocytic leukemia (Yosifov, Wolf et al. 2019).

Some of the common treatments are explained below in detail and also detailed in figure 1.6.

1. A common front-line treatment is monotherapy with ibrutinib. It is also used for CLL patients with *TP53* mutation. This is a small molecule BTK (Bruton tyrosine kinase) inhibitor that covalently binds to BTK leading to downstream inactivation of cell survival pathways like NF-κB and MAP kinase, important for BCR signaling. 82.1% of the patients (total 84 patients) treated with ibrutinib showed signs of reduced tumor proliferation and cell death resulting in improved survival (Ahn, Underbayev et al. 2017). However, other 17.1% of the patients after showing an initial decrease in malignant B cells, stop responding and relapse. These patients that show resistance to ibrutinib treatment have a mutation either in the gene *BTK* or *PLCG2*.

   There is a keen interest now in understanding the mechanism of CLL growth and evolution through acquired mutations and dynamic transcriptional changes that make way to a highly aggressive ibrutinib resistant disease.

2. The mechanism of BCR pathway inactivation by a BTK inhibitor (ibrutinib and acalabrunitib) is shown in figure 1.6. The figure also shows drugs (eg. Kinase inhibitors Idelasib, develisib targeting PI3K) used to inhibit different components of the BCR signaling pathway, with an aim to contain it and restrict B cell proliferation.

3. Venetoclax is a Bcl-2 inhibitor that blocks the prosurvival functions of the Bcl-2 protein. A combination of ibrutinib and venetoclax is also recommended, so that B cells that delocalize to the peripheral blood from secondary lymphoid tissues as a response to ibrutinib are fairly short lived (Souers, Leverson et al. 2013).

Other treatment options employ Anti-CD20 antibodies that target CD20 surface protein expressed by malignant B cells (Huhn, von Schilling et al. 2001), or a combination therapy using fludarabine with cyclophosphamide (FC) (Eichhorst, Busch et al. 2006).

The mechanism of action of treatment options for CLL, before undergoing clinical trials with enrolled patients, is studied in the pre-clinical mouse model of CLL, the Eμ-TCL1 model. This necessitates the study of genetic and transcriptomic make up of this mouse. Mouse models can help study treatment response rates and mechanism of action faster and reliably if the progression of the disease in mouse is comparable to that in human patients. I therefore review next, known facts about the Eμ-TCL1 mouse model and its suitability to study CLL.

1.9 Eμ-TCL1 mouse: The Preclinical Model for studying CLL

TCL1 is an oncogene activated in T-cell prolymphocytic leukemia. It is activated due to repeated reciprocal translocations occurring at the chromosome segment 14q32.1. It was also shown to be expressed in B cells in CLL, and not on the normal mature B cells (Yuille 2001). This gene was then placed under the control of the promoter of the mouse immunoglobulin heavy chain (*Ighv*) gene in B cells, to enhance its production (Bichi, Shinton et al. 2002). In two months, this led to polyclonal proliferation of CD5+ B cells in the peritoneal cavity of the mice. At 3-5 months these cells could be detected in the spleen and at 5-8 months in the bone marrow. At 8-9 months it was possible to detect high numbers of monoclonal B cells, and at 13-18 months CLL like symptoms of enlarged spleen and marked

lymphoadenopathy was confirmed. This phenotype was established as a prolymphocytic transformation of CLL and was henceforth used in many other studies a relevant medium to investigate CLL.

The Eμ-TCL1 mouse model, hence, manifests a disease like the aggressive form of chronic lymphocytic leukemia. Since its development 17 years ago, the mouse model has been extensively used for studying CLL development, progression and pathogenesis, clonal evolution in CLL, screening drugs and elucidating their mechanism of action figure 1.7.

This model and its adoptively transferred form (Eμ-TCL1-AT) has been used in our lab to understand the pathophysiology of CLL. The adoptive transfer model is one in which B cells from the spleen of Eμ-TCL1 mouse manifesting CLL, are taken and injected into the peritoneal cavity of another Eμ-TCL1 mouse. The secondary tumor in the AT mouse develops into CLL much faster than the primary tumor, i.e. in about 5-6 months. This approach also develops a more aggressive CLL.

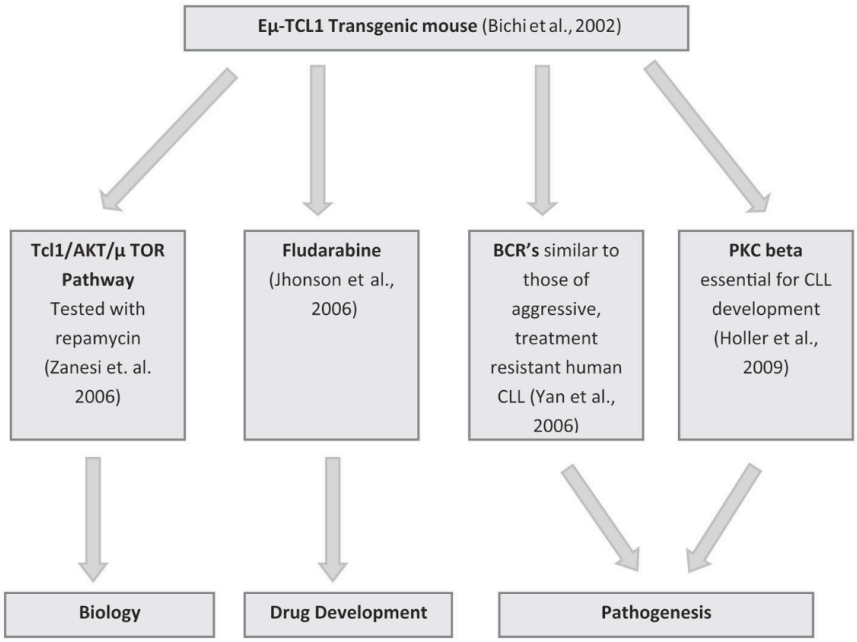


Figure 1.7: The Eμ-TCL1 mouse and its usefulness in testing disrupted biological pathways, drug development and treatment studies and evolution and progression of CLL (Pekarsky, Drusco et al. 2015).

It is still needed to study the kind and mechanism of B cell clonal changes in the secondary tumor that causes faster development of the disease. This could answer whether there is a specific stage in B cell development that when transformed develops into a more aggressive CLL. Gap also remains to understand the suitability of the TCL1 mouse in studying the CLL tumor microenvironment. CLL progression is highly dependent upon signals from its microenvironment. In addition, it also needs to be studied how the genetic background of the mouse impacts CLL development.

This thesis digs into genomic heterogeneity of cells from CLL patients and the above described mouse model by means of computational and bioinformatics approaches. Next Generation Sequencing (NGS) approaches were used to assess the mutational landscape, BCR dynamics (whole exome sequencing: WES) as well as transcriptional changes due to drug treatment (RNA-sequencing). In addition, a major part of the thesis comprehends the tumor microenvironment (TME) of CLL at single cell proteome and transcriptome using high throughput CyTOF (mass cytometry) as well as RNA-sequencing (scRNA-seq) respectively. These NGS and single cell measurement approaches are discussed next.

1.10 Whole Exome Sequencing (WES)

WES is targeted sequencing of the protein-coding regions in a genome. This means that only the exons of all the gene are captured and sequenced, hence the name “exome sequencing”. The protein coding region constitutes about 1% of the entire human genome. In order to identify genes that alter protein sequence, structure and eventually their function, genetic variants underlying exonic regions of approximately 30 million base pairs need to be assessed. Therefore, to identify only the protein altering genetic variants, sequencing the entire genome is both more cost and time consuming as compared to sequencing the exome. The first step in this process is to capture the target regions of interest by either capture oligos or hybridization arrays. This is followed by high throughput sequencing of the captured regions. In addition, the costs saved can be used to achieve higher coverage while sequencing only the targeted regions, and hence be useful for variant calls with higher read support and reliability (10).

Agilent SureSelect kits for human as well as mouse are common protocols for capturing targeted regions. They use approximately 120 base RNA probes to capture coding sequences. Briefly, the steps include fragmenting, denaturing and then hybridizing the DNA to capture oligos. Captured sequences are then labeled with paramagnetic beads conjugated with streptavidin. These are amplified further before sequencing on Illumina machine (Chen 2015). Illumina HiSeq 4000 platform was used for WES in this thesis. 4 samples were multiplexed per lane generating a total of 235 million paired reads. WES can be used to identify Single Nucleotide Variants (SNVs) and Copy Number Variations (CNVs) altering the coding region of the genome. WES has applications in identifying rare variants with higher coverage, identifying candidate genes for Mendelian disorders and other complex diseases, and to design panels for clinical diagnostics.

1.11 RNA-sequencing (RNA-seq)

RNA-seq is an NGS technique used to quantify the transcriptome of a cell. Changes in gene expression over time and in between conditions can hence be assessed. In addition, isoforms of a transcript as a result of alternate splicing, gene fusions and post translational modifications can be identified using RNA sequencing. In cancer biology, RNA sequencing often aids in quantifying the changes in transcriptional profile of the disease state in contrast to the control/normal state. Small RNAs as well as non-coding RNAs can also be sequenced, applying varied library preparation methods. In contrast to bulk RNA sequencing where RNA from the entire dissociated tissue is sequenced, newer technologies to isolate single cell and process their transcriptome are becoming increasingly common. These fall under the category of single cell RNA sequencing (scRNA-seq) and is explained below. Isolating and sorting single cells specific to a tissue or condition and quantifying their transcriptome prevents contamination from non-tissue sources, which is unavoidable in bulk sequencing. Bulk RNA-sequencing and scRNA-sequencing NGS approached have both been used in this thesis to answer separate biological questions.

1.12 Single cell droplet-based transcriptome profiling

ScRNA sequencing has proved to be beneficial in studying tumor clonality and heterogeneity, track development and interrogate immune system cell types and diversity at single cell level. To study immune T cell populations infiltrating the CLL niche, scRNA sequencing paired with targeted single cell T cell receptor sequencing was used. There are several kinds of single cell library preparation techniques useful for different biological purposes. scRNA-seq methods differ based upon either amplifying full-length cDNA or cDNA at either 5' or 3' end attached with a unique molecular identifier (UMI). SMART-seq and SMART-seq2 are techniques that employ full length transcript. Protocols using UMIs include MARS-seq, STRT, CEL-seq and ones that used droplet-based platforms (Drop-seq, inDrop, 10X chromium).

10x genomics chromium scRNA sequencing

The 10X platform uses gel bead in emulsion (GEM) approach. Each cell gets encapsulated in a gel bead labeled with oligonucleotides. These oligonucleotides consist of a unique cell barcode (10bp UMI), sequencing adapters and primers, and a 30bp oligo-dT (See, Lum et al. 2018). Steps involved in scRNA-seq using 10X genomics method is described in figure 1.8.

Droplet strategy greatly increases the throughput to thousands of cells being profiled at once. Up to 8 samples can be simultaneously processed on the 10X microfluidics chip, at acceptable costs and minimal time (Zhang, Li et al. 2019). The current detection limit of this technique is 500-1500 genes per cell on an average. However, there is little control over the number of cells analysed, the throughput being 50% of input cell number. This may lead to incorrect representation of systems inside the cells and may prevent detection of rare subpopulations. Importantly, 10X transcriptome profiling method can be used in combination to determine TCR repertoires of single cells in case of T cells or in general with cellular indexing of epitopes for multiplexed quantification of thousands of protein markers in each cell. This can be very useful for large scale immunophenotyping and studying post-transcriptional modifications at single cell level. Paired scRNA and TCR profiling have been used in this thesis.

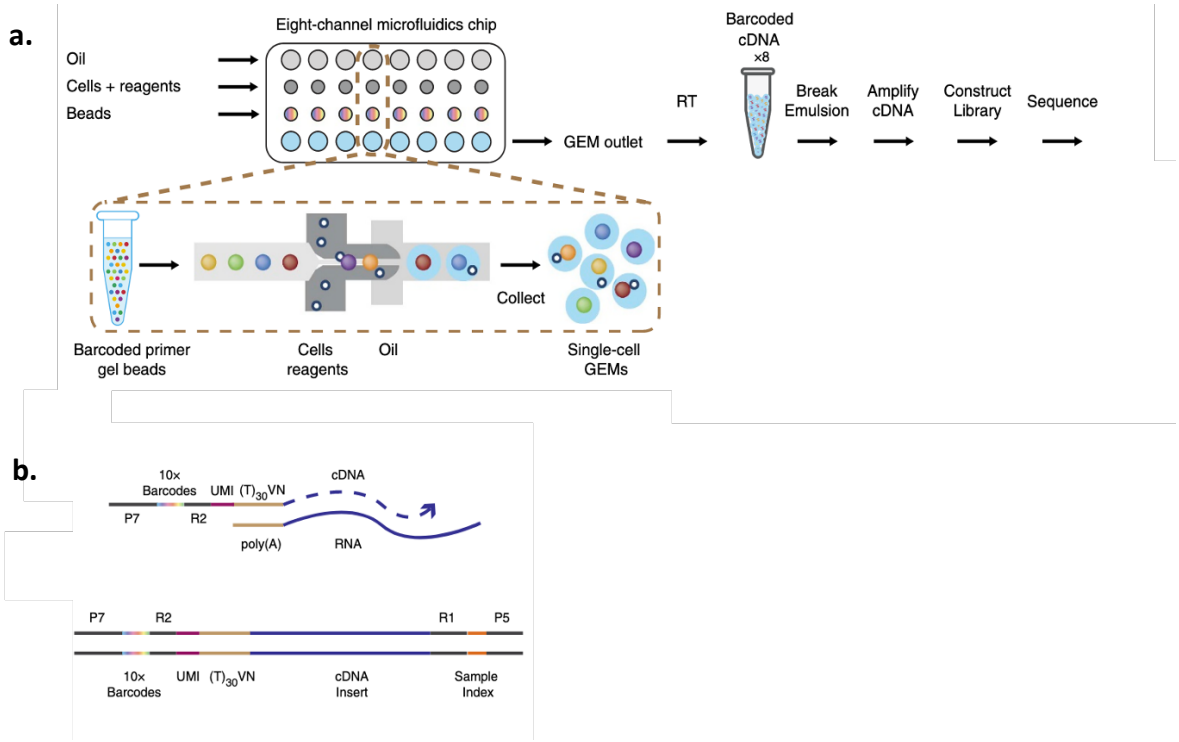


Figure 1.8: (a) Steps involved in 10x chromium based scRNA-sequencing. (b) Structure of the oligonucleotide in each gel bead, and binding of a single poly(A) mRNA per oligo (above), and the final transcribed cDNA with UMI, ready for sequencing (below) (Zheng, Terry et al. 2017).

Single cell libraries are sequenced on conventional Illumina machines depending upon the coverage required. Even the other NGS measurements used in this thesis like WES (whole exome sequencing) and RNA-sequencing, after library preparation employ Illumina machines for amplification. Illumina uses sequencing by synthesis approach with the important step of bridge amplification to create clonal clusters of the same fragment of DNA (11, 12).

1.13 CyTOF (Mass Cytometry)

Mass cytometry is the fusion of *mass* spectrometry and *flow* cytometry and is capable of recording over 40 parameters at the level of single cell resolution on an "-omics" scale. Conventional flow cytometry utilizes fluorophores as reporters, limiting measurement of several molecules together due to overlap of their fluorophore emission spectra. Mass cytometry utilizes interaction between the proteins to be measured their specific antibodies. The technique enables investigation, by coupling antibodies to heavy-metal isotopes, whose

reported quantity in a mass channel quantifies the molecular expression of several surface and intra-cellular proteins in one go, and that too with little signal overlap.

The procedure of mass cytometry is represented in figure 1.10. Briefly, cells of interest (e.g. T cells) are first incubated with a cocktail of antibodies conjugated to stable heavy metal isotopes, carefully selected to study the phenotype of interest (e.g. T cell exhaustion) (Lou 2007, Ornatsky 2008). These antibody probes are designed to bind proteins of interest on the surface or within the cells, so that the attached metal ions can quantify the expression level of the target proteins. Creation of single cell suspensions is then initialized, followed by their transmission through a nebulizer to place the cells within droplets ready to be introduced into the mass cytometer. The cells pass through an argon plasma after entering the mass cytometer. Here, free atoms are produced by disintegration of covalent bonds. The ion cloud of atoms that get charged in the process then passes through a quadrupole that enriches for heavy-metal reporter ions. These ions are then separated by their mass-to-charge ratio in a time of flight (TOF) mass spectrometer. The output is a data matrix (.fcs) of converted ion counts into electrical signals. Each column of the data matrix is a unique isotope and each row represents the scanned mass in every cell.

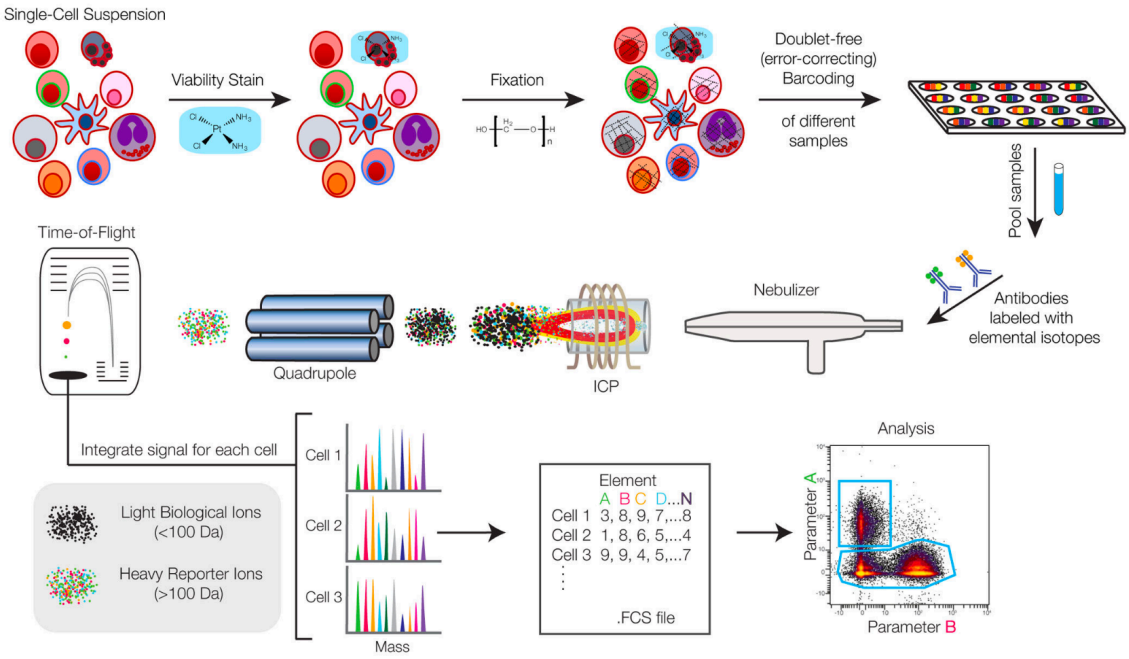


Figure 1.9: Steps involved in mass cytometry. Adapted from (Spitzer and Nolan 2016).

Mass cytometry can be used to study cellular complexities simultaneously at several levels, e.g. measurement of proliferation markers, cell signaling molecules, activation and adhesion molecules together can report cellular behavior and stage during a biological process (Bendall, Davis et al. 2014). The technique also allows interrogating several cell types in the same assay. Number of parameters measured can also be scaled to approximately 100 in certain cases. This can help quantify cell-cell interactions in cancer and other scenarios to reveal coregulation and crosstalk between molecules.

However, the procedure is limited in a way, that live cells are not feasible to recover because of ionization. In addition, compared to other single cell measurement techniques the throughput of mass cytometry is low. Some of the heavy isotopes are reportedly less sensitive than the conventional fluorophores making it difficult to measure features expressed at very low levels. Mass cytometry output can be analysed by conventional genomics and proteomics tools available for clustering cells to define subpopulations, visualization, and identifying differentially abundant populations between phenotypes. Certain software and workflows integrate specialized steps to normalize CyTOF intensities and interpret protein marker expression, e.g. HDCytoData, Citrus, SPADE, FLOW-MAP etc.

1.14 Computational Approaches Used

MiXCR

MiXCR was used to identify B cell receptor (BCR) rearrangements from WES data of Eμ-TCL1 mouse tumors in the present thesis. It can quantify B cell receptor (BCR) and T cell receptor (TCR) clonotype information from raw RNA or DNA, paired or single-end reads (Bolotin, Poslavsky et al. 2015). In addition, it handles sequencing and PCR errors, and identifies germline hypermutations. MiXCR taken in raw sequencing files (.fastq) and target regions (.gtf) to output the exact V, D, J gene segments, complementarity-determining region 3 (CDR3) amino acid and nucleotide sequence, and frequency of the identified clonotype rearrangement. The thresholds used are explained in detail in the methods section. MiXCR was preferred over other available tools for identifying BCR and TCR rearrangements as it has previously been shown to be more flexible and accurate with a detailed documentation and

adjustable parameters to suit the analysis. It was proven reliable especially for analysis of UMI-based reads like the ones used in the present study (Afzal 2019).

Mutect2

Mutect2 was used to identify somatic variants from deeply sequenced (approximately 200x) WES data of Eμ-TCL1 mouse tumors and matched controls in the present thesis. Mutect2 can identify somatic variants from both WES (whole exome sequencing) and WGS (whole genome sequencing) data with higher sensitivity and precision as compared to other variant callers (Benjamin, Sato et al. 2019). It applies local assembly and alignment, a Bayesian somatic genotyping model, and new filtering methods to identify variants. Mutect2 is particularly suited for identifying variants with low allele frequencies and hence, subclonal mutations; and was decided to be best suited for studying the mutational landscape of CLL that manifests most variants at less than 5% allele frequency (also cited before in this thesis).

CNVkit

CNVkit was used to call copy number variations (CNVs) from WES data (Talevich, Shain et al. 2016) of Eμ-TCL1 mouse tumors using a reference of pooled matched controls in the present thesis. CNVkit is easy to implement and produces detailed information on copy number segments and ranges. Most importantly it is able to infer CNVs using reads mapping to both on and off-target genomic regions. This was an important consideration for the present thesis because of availability of only WES mouse tumor data for analysis. Also, CNVkit allowed pooling of control samples, to infer CNVs accurately. It allows flexibility with commands 'scatter' and 'metrics' to produce copy number visualizations and converting segmentation into copy number states for easy interpretation.

PyClone for Estimating Cancer Cell Fraction (CSF)

PyClone uses information about copy number states and allelic frequencies of SNVs as input to hierarchical Bayesian non-parametric statistical model to estimate clonal changes in tumors (Orbanz and Teh , Roth, Khattra et al. 2014). Accounting for change in copy number states, it corrects for allelic imbalances. PyClone has been validated to be accurate in predicting cellular prevalence of clonal clusters in tumors from time series experiments, especially with deeply sequenced genomes. Such an analysis helps to reflect tumor growth

dynamics. Beta-binomial emission densities uniquely used by PyClone accurately model the variance in allelic prevalence over time. Another distinguishing feature of PyClone is its ability to estimate clonal clusters even with relatively low number of SNVs from WES data. This was helpful as there were low number of high confidence mutations (read depth = 5 and allele frequency > 10%) identified in the mouse tumor cohort analysed in this thesis. Other filtering steps are detailed in the methods section.

Limma: linear analysis for microarray data

Though limma was developed to analyse microarray data, it can be used to fit many kinds of data types (expression, methylation, counts) into a broad class of data models (linear model, generalized linear model, generalized linear mixed model) to understand linear regression, as well as impact of covariates on the data by building contrast matrices (Ritchie, Phipson et al. 2015). It uses Bayes moderation for differential testing. This kind of linear modelling, that internally normalizes variation in library sizes apparent in the data has been used twice in this thesis. It was used for studying 1) differential transcriptome changes by ibrutinib treatment on Eμ-TCL1 mouse tumors; and 2) differential subpopulation abundances between tumor LNs and control LNs. It was necessary in both the cases to either subtract or add effects of covariates like proliferation, time and treatment, gender respectively on the parameter (transcription profile and abundances) being evaluated. Therefore, the flexibility with which limma analysis can be modulated and its suitability with even 3-4 cases in each group was utilized in the present thesis (Law, Alhamdoosh et al. 2016).

Clustering and Integrating single cell data across species and technologies

Graph-based clustering

Identifying similar subpopulations/groups of cells from scRNA-seq data involved a 2 step process. These were applied using functions in the Seurat workflow that are detailed in the methods. In the first step, Euclidean distances in a PCA space are used to construct a KNN (K-nearest neighbour) graph. Edge weights between a pair of cells are refined based upon their locally shared overlap or Jaccard similarity (15). Graph based approaches (16) have been popularly used for high dimension data previously as well (Levine, Simonds et al. 2015, Xu and Su 2015, Liu, Song et al. 2019).

The second step, in addition applies a modularity approach, to assess the similarity of the connections in a group cell in comparison to random connections. Seurat uses the Louvain approach for this, which is popular over many domains (17) (Blondel 2008).

Integrating scRNA-seq datsets

To integrate scRNA-seq data across multiple datasets (CLL and Beast Cancer data set in the present thesis), species or technologies Seurat workflow first identifies common anchors between the datasets using canonical correlation analysis (CCA) and mutual nearest neighbour approach (MNN) (Haghverdi, Lun et al. 2018). The method works by calculating correlations between highly variable features identified in the datasets to be combined. This preserves the biological structure of each dataset. MNN pairs/anchors identified between datasets to be combined are then scored and corrections are applied. This method was found to outperform other existing data integration methods in terms of accuracy, preserving the original biological structure, and integrating diverse scRNA-seq datasets. This method also preserves rare subpopulations in the datasets being combined (Stuart 2019).

The applied methods, functions and their used parameters and thresholds are explained in the methods section.

2. Materials and Methods

The following chapter describes the approaches and exact steps considered to address the biological problems within the scope of this thesis. Experimental work has been performed by other members participating in the project. All the analysis that have been performed and detailed below, is my work, unless otherwise mentioned.

2.1 Samples

2.1.1 Eμ-TCL1 mouse samples

Wild type (WT) mice for adoptive transfer (AT) of TCL1 tumors were purchased from Charles River Laboratories (Sulzfeld, Germany). Eμ-TCL1 (TCL1) mice on C57BL/6 background were provided by Carlo M. Croce (The Ohio State University, Columbus, Ohio, USA). Adoptive transfer of TCL1 tumors was performed by enriching leukemic B cells from splenocytes of TCL1 mice using EasySep™ mouse Pan-B Cell Isolation Kit (Stemcell Technologies, Cologne, Germany) according to the manufacturer's protocol (Hanna, McClanahan et al. 2016). The CD5+ CD19+ content of purified cells was typically above 95%, as measured by flow cytometry. 2*107 leukemic TCL1 splenocytes were transplanted by intraperitoneal (I.P.) injection into 6-10 weeks old C57BL/6 WT females for TCL1 AT experiments. All animal experiments were carried out according to governmental and institutional guidelines and permitted by the local authorities (Regierungspräsidium Karlsruhe, permit numbers: G-36/14 and G-98/16). These experiments were performed by Dr. Selcen Özturk.

2.1.2 Human samples for CyTOF and scRNA analysis

Primary lymph nodes (LN), peripheral blood (PB) and bone marrow (BM) samples were obtained from CLL patients after informed consent in accordance with the guidelines of the Hospital Clinics Ethics Committees (University Hospital Clinic Barcelona and University of Heidelberg) and the Declaration of Helsinki. All primary CLL tumors in this study were

diagnosed according to the World Health Organization (WHO) classification criteria (Arber, Orazi et al. 2016). Detailed information about the CLL patient samples and cell counts per sample is provided in Supplementary Table 2, along with clinical information like age, IGHV status, treatment and gender. Non-malignant reactive lymph node samples (rLNs) were used as controls. Collaborators at University Clinic Barcelona who provided the samples: Dolors Colomer and Elías Campo. Collaborators at University Clinic Heidelberg who provided the samples: Sascha Dietrich and Tobias Roider.

2.2 Data Generation

2.2.1 Whole exome sequencing (WES)

Library preparation for targeted sequencing of CD19+ B cells from spleen of Eμ-TCL1 mice was performed using SureSelectXT Mouse All Exon kit from Agilent. The samples were subsequently sequenced on HiSeq 4000 platform using 100bp paired-end reads with 4 samples per lane according to the manufacturer's instructions at the DKFZ Genomics and Proteomics Core Facility.

2.2.2 Mouse immunoglobulin repertoire sequencing

RNA quality was assessed with the Agilent Bioanalyzer and 300-500ng were used for RACE PCR as previously established (Turchaninova 2016, Afzal 2019). Briefly, cDNA was synthesized using primers annealing to immunoglobulin heavy or light chain constant regions and a barcoded template-switching primer. This was followed by AMPure bead purification. Next, two consecutive exponential PCRs were performed using 2 µl of the single-stranded cDNA or the first amplification product, respectively. Libraries were purified with AMPure beads and pool size selection was performed on agarose gels prior to 400+100bp paired-end sequencing in the Illumina MiSeq platform. This protocol was performed by collaborators at DKFZ (Dr. Saira Afzal, Dr. Irene Gil-Farina).

2.2.3 RNA-sequencing

RNA libraries were prepared from CD19+ B cells from spleen Eμ-TCL1 mouse tumors using Illumina TruSeq Stranded mRNA protocol. The samples were subsequently sequenced on HiSeq 4000 platform using 125bp paired-end reads according to the manufacturer's instructions at the DKFZ Genomics and Proteomics Core Facility. RNA isolation and experiments prior to that was performed by Haniyeh Yazdanparast.

2.2.4 Mass cytometry (CyTOF) staining and acquisition of T cells

Following staining, filtering and counting of CD19- cells, samples were analyzed at a flow rate of 0.030ml per minute using the Helios mass cytometer (CyTOF) (Fluidigm) of the National Cytometry Platform (NCP) at the Luxembourg Institute of Health (LIH) in Luxembourg. After acquisition, initial data processing and quality control were performed. NCP and FCS (flow cytometry standard) files were normalized with the HELIOS instrument acquisition software (version 6.7.1014) by using EQ beads as standard. Patient samples for CyTOF were processed in collaboration with Luxembourg Institute of Health (Marina Wierz, Etienne Moussay, Jérôme Paggetti).

2.2.5 Mass cytometry (CyTOF) panel and metal labeling of antibodies

A custom 43-marker panel focusing on T cell phenotyping and including both surface and intracellular markers was designed by Laura Llao Cid and Martina Seiffert and measurements performed in collaboration with Luxembourg Institute of Health (Marina Wierz, Etienne Moussay and Jérôme Paggetti). Combinations of markers and heavy metal isotopes are detailed in Supplementary Table 3. For most of the markers, heavy metal-conjugated antibodies were commercially available and purchased from Fluidigm. For the other markers (*) heavy metal labeling was performed using the Maxpar® X8 Multimetal Labeling kit (Fluidigm, ref 201300) according to the manufacturer's instructions.

2.2.6 Generating single cell transcriptome of $CD3^+$ T cells from CLL TME

Single cell transcriptomes were generated for CLL patient LN as well as Eμ-TCL1 mouse tumors using the Chromium Single Cell Immune Profiling Solution Reagent Kit (ChromiumNext GEMSingle Cell V(D)JReagent Kits v1.1, 10X genomics) following the manufacture's protocol. Briefly, 10.000 cells per sample were loaded into the Single Cell Chip for separation into nanoliter-scale Gel Beads-in-emulsion (GEMs), with the aim to retrieve 5,000 cells in the end (50% output). GEM generation was followed by reverse transcription to obtain full-length cDNA from poly-adenylated mRNA. GEMs were subsequently lysed and pooled cDNA was PCR amplified. 5' Gene Expression library were prepared next. For targeted TCR enrichment, PCR amplification was performed using specific TCR primers. Libraries were prepared as recommended by 10X genomics (13). Samples for scRNA-seq were processed by Laura Llao Cid using the facilities at Single Cell Open lab (Jan-Philipp Mallm, Katharina Bauer, Michelle Liberio, Karsten Rippe) at DKFZ, Heidelberg. Sequencing of the libraries was accomplished on HiSeq 4000 machine (Illumnia) using 50bp single reads or a NovaSeq 6000 Paired-End (28+94 bp) at DKFZ Genomics and Proteomics Core Facility.

2.3 Steps involved in assessing clonal evolution dynamics of BCR and SNV-defined clonotypes in Eμ-TCL1 mouse tumors

The major steps involved in this part of the thesis are described as a flowchart in figure 2.1 and are detailed individually.

- Alignment

Raw DNA sequencing reads (fastq) were aligned using an in house roddy framework that essentially used Burrows-Wheeler Aligner with default settings (bwa-mem v0.7.8) for alignment to the mouse reference assembly (UCSC mm10). The framework used biobambam2 (v0.0.148) for sorting, marking duplicates and merging temporary alignment files to output aligned whole exome sequencing (WES) bam files.

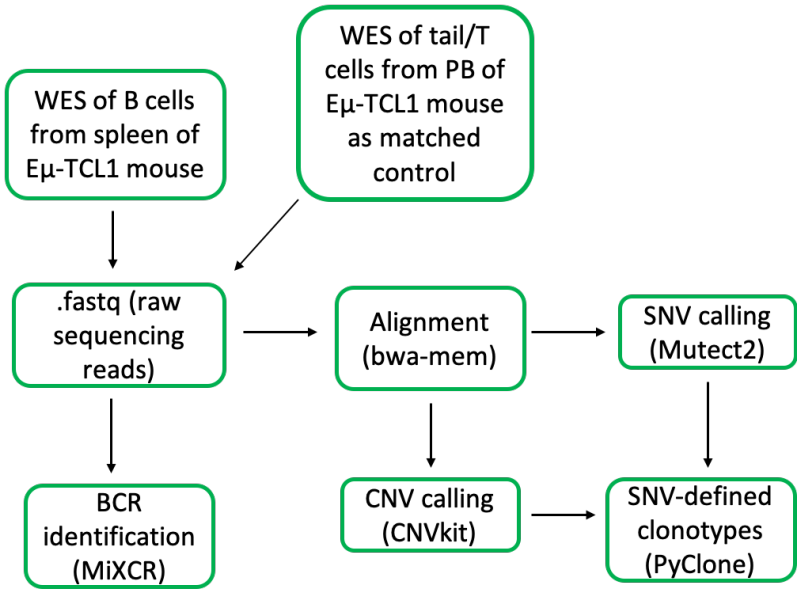


Figure 2.1: Steps used for assessing clonal heterogeneity in Eμ-TCL1 mouse tumors.

- Calling Single Nucleotide Variants (SNVs)

SNVs were called from WES of mouse tumor and control samples using Mutect2 from GATK (v4.0.2.0) (Benjamin, Sato et al. 2019). Mutect2 calls somatic short variants (SNVs) and insertion and deletion (indel) variants using local assembly of haplotypes. Paired tumor-matched normal setting of Mutect2 was utilized. The cohort included four primary Eμ-TCL1 mouse tumors and their four serially transplanted secondary tumors. Information about mouse tumor samples is detailed in Supplementary Table 1. Additionally, all available mouse normal/control samples (n = 8; 3 primary tumor matched T cells from spleen and 5 primary tumor matched tails including 4 tails from separate Eμ-TCL1 mice as additional controls) were used to construct a PON (panel of normals) wherein mutations were called first on each normal sample as if it was a tumor sample (tumor-only mode). Resulting germline mutations from all normal/controls were then combined (CreateSomaticPanelOfNormals) and supplied during paired Mutect2 mutation calling. A PON can also be created using a set of unrelated normal/controls sequenced using similar technology and preparation method as the tumor samples under study. This aids in removing recurrent technical artifacts to ultimately minimize false positive SNVs. Other optional parameters used with Mutect2 were base quality score threshold of 25 (--base-quality-score-threshold) and dbSNP v142 for mm10 as a germline resource (--germline-resource) that helped to eliminate germline variants common in the population in general that were missed during mutation calling. Finally, pileup

summaries (GetPileupSummaries) were estimated for each tumor bam file. Pileup summaries infer read support for a set number of known variants and this was used to calculate fraction of probable contaminants (CalculateContamination) in each sample. These contaminants were then filtered out from the Mutect2 paired tumor-matched normal mutation calls to generate a final variant call format (.vcf) file.

Variants were then annotated with ANNOVAR (v2017Jun1) and all the downstream analysis was performed with variants annotated by Mutect2 as 'PASS' and by ANNOVAR as 'somatic' (Wang, Li et al. 2010).

Mutations were also called and annotated for Eμ-TCL1 tumor samples from publicly available dataset processed at Salzburg (SRP150049) following the procedure described above but with additional normals from Salzburg (n=7) used to prepare the PON.

- Calling Copy Number Variations (CNVs)

CNVkit (v0.9.7.dev0) with default parameters was used for detecting copy number variants and alterations from the available WES data of Eμ-TCL1 mouse tumors (Talevich, Shain et al. 2016). The reference used while calling was built with all TCL1 mouse normal samples and a target .gtf file (downloaded from Agilent for SureSelectXT) supplied to infer CNV calls only from the genomic regions covered during sequencing. To get tumor specific focal aberrations, tumor CN (copy number) state calls were intersected with normal CN state calls. CNVkit maximizes copy number information that can be gathered from targeted sequencing data as it takes advantage of both high-resolution target reads and low resolution off target reads to determine genome wide copy number changes.

- Identification of B Cell Receptor Rearrangements

WES raw data in the form of FASTQ files was used to identify B cell receptors by quantitating V(D)J clonotypes. MiXCR (v3.0.8) was used to map and assemble V, D and J gene segments from immunoglobulin region to report clonotypes for each tumor sample (Bolotin, Poslavsky et al. 2015). The tool was run with default parameters: starting material - 'DNA', adapters - 'present' and receptor type - 'IGH'. The position of V and J primers was mentioned to be 5' and 3' end respectively. The output file reported the frequency, number of supporting reads,

CDR3 (complementarity region 3) amino acid sequence, CDR3 nucleotide sequence, the detected V, D and J genes for each assembled clonotype. If a clonotype was supported by more than 5 reads it was considered a true hit.

- Calculation of change in cellular prevalence of mutations from primary to serially transplanted tumors

To infer if identified SNVs were contributing to clonal changes in mouse tumors, allele frequency changes of tumor specific and shared (between primary and adoptive transfer tumor pairs) mutations were calculated. It was also important to include sufficient read evidence for the called mutation. VAF (variant allele frequency) cutoff of minimum 10% and depth at least five reads support at any one time point (primary: time point 1 or transfer: time point 2) was considered to filter variants for the analysis that follows. To check if the absence of an SNV at any one time point is purely biological (i.e. contributes to clonal evolution of tumor) and not technical (possibly due to sequencing coverage irregularities at off target regions); it was decided to check the "probability of existence of an SNV" at the calculate coverage (as below by SAMtools mpileup) at any one time point, given it is called at the other time point with a specific success rate (AF of the called variant). This formed a case of a binomial experiment represented in R (v3.5.0) as:

dbinom(n,size,prob)

Where prob = VAF at time point 1
n = number of successes needed out of
size = maximum attempts possible (coverage at time point 2)

Samtools mpileup (v1.9) was used to calculate coverage (depth) of all called SNVs in their respective primary-transfer tumor pairs. The calculated coverage is placed at "size" for dbinom for time point two when the VAF (prob) is from time point one. Success (n) was defined as the probability of existence of an SNV with at least two reads at one time point given it has been called at the other time point with 10% VAF and at least five reads.

The final probability was inferred in Rv3.5 using:

Probability <- 1-((dbinom(0, depth, VAF) + (dbinom(1, depth, VAF))))

To strictly avoid any technically artifactual SNVs, 1% error was allowed considering a 99% confidence interval for each SNV position. SNVs from here that passed adjusted p-value (Bonferroni correction (Armstrong 2014)) threshold of <0.05 were selected and used for plotting clonal changes between tumor pairs. Python tool PyClone (v0.13.0) was used for this (Roth, Khattra et al. 2014). PyClone uses a Bayesian statistical method and outputs putative clonal population clusters of grouped input SNVs. It also uses copy number state information to check for allelic imbalances in each sample and estimating accurate cellular prevalence of point mutation clusters defining clonal shifts in the tumor. Copy number states for the tumors was used as calculated during CNV calling as described above.

2.4 Steps involved in studying genomic and transcriptomic changes inflicted in Eμ-TCL1 mouse tumors on ibrutinib treatment

The TCL1 mouse model that manifests disease resembling aggressive form of CLL was used to study the effects of ibrutinib treatment (which is already a promising therapy in CLL patients). Intravenous adoptive transfer of $2*10^7$ splenocytes from a TCL1 mouse into a 6 weeks old BL/6 mouse was performed 2 weeks before ibrutinib treatment was started. Subsequently ibrutinib in an amount of 25mg/kg/day was mixed in the drinking water of the mice. Mice were dissected at time points of 3-, 5- and 8- weeks from the start of injection (adoptive transfer) which translates respectively to 1-, 3- and 6-weeks post treatment (ibrutinib) start. Splenocytes were isolated and flow sorted for CLL (CD5+ CD19+) cells on the same day for all three time points. DNA and RNA were isolated from frozen cell lysates. Experiments involving mouse work and cell sorting were performed by Haniyeh Yazdanparast. Raw WES data was processed and analyzed by me as discussed earlier in the methods section.

- Alignment, QC and counting for RNA sequencing raw reads

FASTQ files with raw RNA-sequencing data were aligned using STAR aligner (v2.5) with mm10 mouse genome assembly and its associated genome annotation from Gencode (v14) retrieved

from UCSC table browser (Karolchik, Hinrichs et al. 2009, Dobin, Davis et al. 2013, Frankish, Diekhans et al. 2019). Quality control metrics for the aligned .bam files were generated with the command line tool RNA-SeQC(v1.1.8) (DeLuca, Levin et al. 2012). Transcript level feature counting was performed using the function 'featureCounts' from the command line functionality of subread package (v1.5.3) in the paired end, strand specific mode (reverse stranded) using the same genome annotation file (.gtf) from Gencode as mentioned above. Output from this was a numerical matrix of read counts, with samples as columns and transcripts as rows.

- RNA-seq data exploration, transformation, normalization and differential expression

Data exploration and downstream analysis was performed using DESeq2 (v1.28) workflow on R 3.5 (Love, Huber et al. 2014). Feature counts output was converted to log counts per million (LCPM) and normalized using the TMM method by Robinson and Oshlack (Robinson 2010). For exploring the transformed data and checking inherent patterns conferred by the specific treatment groups in the data, unsupervised clustering using principal component analysis (PCA) and sample-sample correlation analysis using the expression of all and 1000 most variable genes across all the samples respectively was performed.

The aim of the study was to delineate factors responsible for ibrutinib resistance in TCL1 mice by comparing gene expression profiles of ibrutinib treated and untreated (vehicle) samples.

For this purpose, the limma bioconductor package (v3.38.3) was used to design a complex linear model, that incorporated an empirical bayesian method to compare the various treatment groups under study and identify genes putatively differentially expressed in ibrutinib resistance tumor groups as compared to others (Ritchie, Phipson et al. 2015). The following contrast was built:

IbResistance=(Ib.late-Ve.late)-(Ib.early-Ve.early)

Ib.late: ibrutinib treated at late time point

Ve.late: vehicle treated at late time point

Ib.early: ibrutinib treated at early time point

Ve.early: vehicle treated at early time point

This model eliminated effects of transcriptional changes characteristic of proliferation (in vehicle groups) and ibrutinib treatment (in ibrutinib treated groups) and resulted into a gene list with a transcription profile manifesting changes specific to ibrutinib resistance. The output included associated $\log_2$ fold change, p-values and adjusted p-values (Benjamini and Hochberg 1995) for the differentially expressed genes in ibrutinib resistance group. Significantly (adjusted p-value < 0.05) differentially expressed ibrutinib resistant genes were then displayed as a heatmap.

2.5 ScRNA sequencing and CyTOF analysis (mass cytometry) workflow

The workflow for single cell analysis is shown in figure 2.2 and described below in detail**.**

2.5.1 Steps involved in scRNA sequencing analysis

- Alignment for single cell transcriptome analysis

Raw transcriptome sequencing data (.fastq) was aligned to specie specific reference genomes and reads were counted using the analysis pipeline Cell Ranger from 10x genomics (cell ranger count). The input used either mm10 reference genome assembly for mouse or hg38 reference assembly for human samples and raw sequencing files. The pipeline uses the STAR aligner. Output from this step was a sparse count matrix, where gene names were rows and cell barcodes were columns.

In addition, results from this step also included quality control metrics for each sample and reported the number of cells and unique transcripts sequenced, identified and usable for further analysis. The count matrix generated was used for further downstream transcriptome analysis. This was a memory intensive step and needed at least 250GB memory and 16 cores.

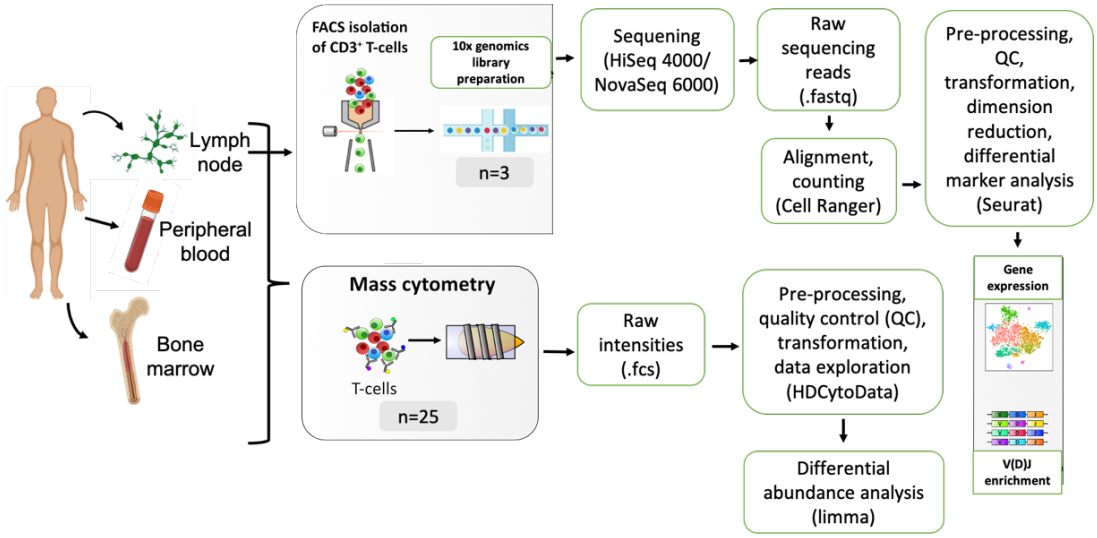


Figure 2.2: Single cell analysis workflow adopted in this thesis. Figure adapted from work of Laura Llao Cid, but modified to include details of bioinformatics analysis.

- Data preprocessing, normalization and scaling, dimension reduction, clustering and cell type identification using Seurat workflow

Seurat v3 toolkit in R3.5 was used for analyzing single cell transcriptomics data (Butler, Hoffman et al. 2018, Stuart 2019). The workflow includes the following steps:

1. QC and data filtration

Count matrix generated by the Cell Ranger pipeline was read into Rv3.5 and converted to an S3 type Seurat object. Cells were then filtered based upon QC metrics, library size and their mitochondrial content. QC metrics include number of detected genes in each cell. It was made sure that the cells being used for analysis are not low-quality cells or empty droplets that have low genes expressed. For this reason, genes expressed in less than two cells and the cells expressing less than 200 genes (cutoff of > 300, changed the downstream results) were removed. Cells that had aberrantly high number of genes (eg: double than the median genes expressed by all cells) and could be potential cell doublets and were removed. Cells expressing greater than 3000 genes were removed, as the median was reached at around 1500 genes per cell. After these cut offs the final count matrix had approximately 3k cells (columns) and 15k-17k genes (rows).

Another important factor to be considered at this step was the percentage of mitochondrial genes expressed in each cell. Detection of high number of reads mapping to the mitochondrial genome could be a sign of low-quality cells or cells that have undergone apoptosis. Hence cells with greater than 10% mitochondrial reads were removed. The median number of mitochondrial reads was at 5% for all the cells.

Therefore, downstream analysis steps were performed only on single cells with enough detected genes and minimal mitochondrial content.

2. Normalization and data transformation

The next step of data transformation was performed using the Seurat function SCTransform(). This function calls sctransform::vst. This function performs data normalization, identification of 3000 (by default) highly variable genes (to identify different populations in the data) and data scaling to remove technical bias (i.e. associate all genes with equal weights so that highly expressed genes do not dominate in downstream analysis).

SCTransform() function is based upon 'regularized negative binomial regression' to keep biological heterogeneity in the data intact while removing technical variation by using cellular sequencing depth (that can vary between cells in single cell sequencing, even within the same cell type) as a covariate in a generalized linear model (Hafemeister and Satija 2019). This statistical model is applied to each gene, in contrast to other approaches that apply normalization techniques to genes pooled either by same cell type or similar library sizes.

3. Linear dimensional reduction and selecting significant principal components

Highly variable features/genes identified within the SCTransform() function were then used to perform principal component analysis (PCA) to visualize the data in low-dimension, using the Seurat function RunPCA(). PCA is a statistical technique used to interpret meaningful differences within large data sets by reducing dimensionality of the data while preserving maximum uncorrelated variability in the form of principal components (Ian T. Jolliffe 2016).

The next step calculated PCA scores where each PC (principal component) represented a metafeature - essentially combining correlated information across a set of genes. In the present data it was observed that in most samples the first 10-15 PCs were significant, and the elbow appeared at PC 16 or 17. Hence approximately the first 15-18 PCs were used for clustering.

4. Clustering for subpopulation identification

To identify different subpopulations, present in the dataset, a K-nearest neighbor (KNN) graph was constructed based on the Euclidean distances in the PCA space of the principal components selected in the previous step. The weight edges between the cells were refined depending upon the shared overlap between their neighboring cells, that is using Jaccard similarity. FindNeighbors() function performs this step.

This was followed by hierarchical clustering by means of Louvain algorithm. The Louvain method iteratively detects and merges communities into a single node while maximizing the modularity score for each community. In other words, it recursively compares density of node connections within a community to random connections and increases the modularity score for more connected nodes to eventually form a condensed node. Clustering was implemented using FindClusters() function. The resolution parameter for this function was set to 0.6 selected from the optimal range of 0.4-1.2 recommended for datasets of around 3k cells. This sets the granularity for downstream clustering. A greater granularity increases the number of clusters. For most of the samples a granularity of 0.6 and approximately 15 PCs resulted into phenotypically distinct subpopulations.

5. Non-linear dimensional reduction for visualization

To visualize and explore these data with an aim to place similar cells together in low-dimensional space t-SNE (t-distributed stochastic neighbor embedding) or UMAP (uniform manifold approximation and projection) was used. As an input the same number of PCs was used as selected in the previous steps for clustering, i.e. approx. 15. Using these techniques, similar cells within the graph-based network described in the last step, co-localized in low dimension plot.

6. Identifying differentially expressed marker genes for each cluster and identifying cell types

For this important step Seurat helps identify marker genes that are representative for each identified cell population. FindMarkers() function (to identify markers different between two specific clusters or cell sets) and FindAllMarkers() function (to identify markers expressed in one cluster in comparison to all other clusters) used the non-parametric Wilcoxon rank sum test. The resulting data frame reports the average $\log_2$ fold change between the two groups, p-value of significance, adjusted p-value (95% confidence interval based on Bonferroni correction), percentage of cells in the first cluster where the gene is detected (called as pct1), and, percentage of cells in the other cluster where the gene is detected (called as pct2). Top expressed markers in each subpopulation were then used to annotate its characteristic phenotype.

2.5.2 Identification of T cell receptor (TCR) rearrangements per cell

Cell Ranger pipeline from 10X genomics was used to align and count V(D)J rearrangements per cell (cellranger vdj). The specifics for the pipeline were the same as described above for scRNA seq alignment. This pipeline estimated total clonotypes expressed in each sample (S1, S2 and S3). A clonotype was considered as a rearrangement with an α chain and/or β chain of the TCR. If two cells had the same chain, it was said to have the same clonotype. Output from this step reported: number of cells of a particular clonotype (clonotype frequency), cell barcodes, VDJ genes used, and whether the rearrangement was productive. This information was then mapped to the metadata of scRNA profiles of the 3 samples within the Seurat workflow with the common cell barcode ids in the two datasets (scRNA and scTCR). Cells

within the Seurat clustering could then be marked/colored based upon their clonotype information e.g. clonotype frequency (all clonotypes expressed in more than 2 cells).

2.5.3 CyTOF (mass cytometry) data analysis steps

To interrogate CyTOF marker intensities for patterns, subpopulations and differential abundance in tumor lymph nodes (LN) v/s control lymph nodes (LN), the software suite HDCytoData was employed (Nowicka, Krieg et al. 2017). The workflow used R (v3.6) and Bioconductor (v1.9) based packages. Each sample specific .fcs file from mass cytometer had isotope names in the columns and cells in the rows.

1. Quality control and pre-processing

   All .fcs files were merged into a flowSet object using the flowCore package (Hahne, LeMeur et al. 2009). This step by default transformed intensities and removed cells with extreme positive values for all samples. In addition, a metadata file was provided to the function that had information on patient id, sample id and the condition. Also, the panel of markers used, and their respective isotope information was provided. To transform the varying range of marker intensities, arcsinh transformation with a cofactor of 5 (default) was used (Bruggner, Bodenmiller et al. 2014). Wherever, the data was used for visualization it was transformed further to scale all the expression values in between 0 and 1. The data from three files (flowSet object, panel information and metadata) was then stored in an object of class “SingleCellExperiment” (SCE). The data was then checked for cell counts and library size across all samples. Following this MDS (multi-dimension scaling) plot was used to assess similarities and potential technical batch effects in the data. The MDS plot used median arcsinh transformed marker expression of all the markers (n= 43) listed in the panel file, across all the cells in each sample.

2. Clustering and cell population identification

   FlowSOM and ConcensusClusterPlus metaclustering approaches were used to identify cell clusters, as they were previously known to be the best approaches for CyTOF data in terms of speed and reliability (Wilkerson and Hayes 2010, Van Gassen, Callebaut et al.

2015, Weber and Robinson 2016). These methods were implemented in the workflow within a wrapper function cluster() of class CATALYST. All cells from all samples were used for clustering. Clustering was based upon arcsinh-transformed expression of 32 biological markers (Supplementary Table 3). Since the approach was based upon over-clustering, maxK (number of clusters the cell populations are allowed to form) was set as 15 to identify as many CD4+ subpopulations of relevantly different biological phenotypes as possible. K=15, resulted in phenotypically distinguishable subpopulations (k > 15 resulted in over fitting; k < 15 resulted in under fitting). For visualization purposes 1000 random cells from each sample were represented on a t-SNE plot. This step was repeated with a different seed for t-SNE and different number of random cells (but at least 500, i.e. approximately half the number of cells in the sample with lowest number of cells) from each sample, but 15 phenotypically distinguishable subpopulations were retained. Using only 1000 cells per sample reduced computation time and resources needed to display all cells from all samples. Identified subpopulations were then annotated based upon their unique marker expression into one of the several T cells subtypes (annotation was performed by Laura Llao Cid).

3. Differential abundance analysis

To identify subpopulations that were enriched in CLL LN (n=25) as compared to control LN (n=13), a limma model was constructed (Ritchie, Phipson et al. 2015). This model normalized for the variation in library size across all samples. Limma uses voom to calculate observation-level weights from variance in the data. Also, effects of 'treatment' and 'gender' were added as covariates to check their influence on differential expression of subpopulations. Input for limma analysis was the proportion of cells contributed by each LN sample to each subpopulation. The input data matrix had samples as columns and subpopulations as rows. Proportions of cells per sample in each subpopulation was extracted from the daframe object created using CATALYST in step 2. Limma used empirical Bayesian method to calculate differential subpopulation abundance in CLL LN v/s control LN. Significantly (adjusted p-value < 0.05) different subpopulations were then displayed in the form of a heatmap along with their adjusted p-value (Banjamini and Hochberg method).

3. Results

BCR clonal dynamics in Eμ-TCL1 CLL mouse model

The Eμ-TCL1 mouse is a preclinical tool for investigative studies of CLL in lab. It manifests an aggressive form of CLL-like disease and offers a reliable method to monitor CLL progression. Mouse tumors can be serially transplanted and are then called adoptively transferred (AT) tumors[1]. After transfer they develop into a more aggressive CLL in less time. However, the long latency period in the primary mouse suggests that even though TCL1 over expression does predispose the animal to leukemia, additional genetic and clonal pressures are needed for CLL development. Therefore, to understand the course of CLL evolution in the Eμ-TCL1 mouse model, I analysed data of somatic variations, frequency of mono- or oligoclonal BCRs and their dynamics over serially transplanted tumors. The following observations instate the heterogeneity identified between mouse tumors with respect to number and types of BCRs, patterns of evolving BCR and SNV-define clonotypes. Observations with respect to the copy number profiles of these samples offer a novel avenue, which could add to existing knowledge about development of CLL in the Eμ-TCL1 mouse.

[1] Experiments, serial transplantation of mouse was carried out by Dr. Selcen Özturk, RACE-PCR was performed by collaborators from DKFZ (Dr. Saira Afzal, Dr. Irene Gil-Farina), ibrutinib drug treatment study i*n-vivo* and isolation of DNA, RNA was performed by Haniyeh Yazdanparast.

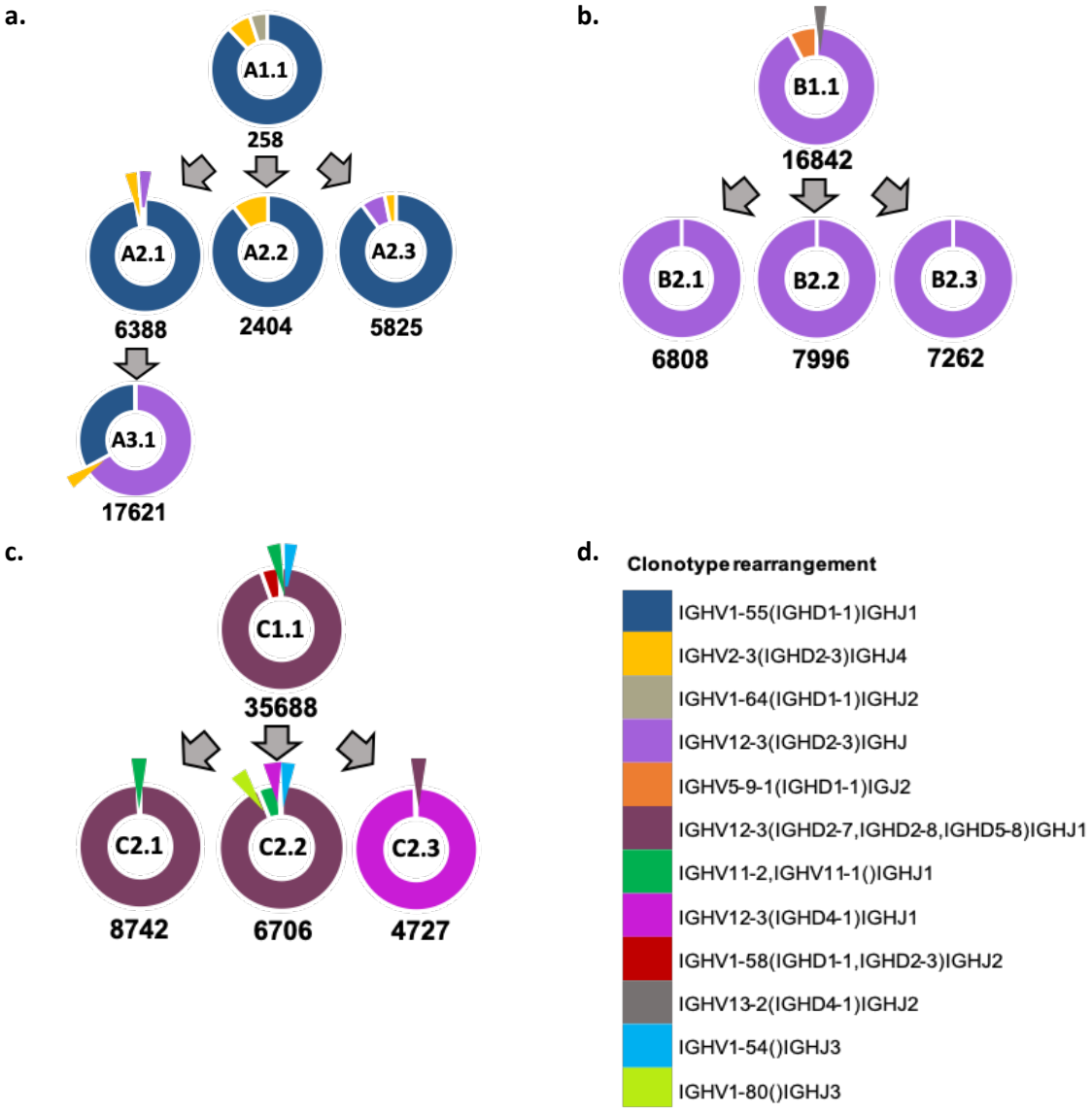


Figure 3.1: Eμ-TCL1 mouse clonotypes as sequenced by RACE-PCR (collaborators Dr. Saira Afzal and Dr. Irene Gil-Farina), and identified using MiXCR. Only 3 out of 13 tumors are monoclonal. Clonotypes with identical V(D)J rearrangement are highlighted with same color. Total number of reads in each sample is indicated below the donut plot. Transfer number of the tumor is shown as, Example T3.1 is 3rd round of transplantation starting from the primary tumor.

3.1 10 out of 13 EμTCL1 mouse tumors have oligoclonal B cell receptors (BCRs)

BCRs of tumor B cells from the CLL mouse model Eμ-TCL1 were sequenced using RACE-PCR (details in methods section, performed by experimental collaborators: Dr. Saira Afzal, Dr. Irene Gil-Farina).

These samples were procured at both parallel and subsequent tumor transplantations. MiGEC (Shugay, Britanova et al. 2014) and MiXCR (Bolotin, Poslavsky et al. 2015) were then applied by myself to process the data and identify the underlying V(D)J rearrangements of *Ighv* genes from B cells of each sample. Clonotypes supported by more than 10 reads were considered for analysis. It is known that these tumors have not undergone somatic hypermutation and are hence of the unmutated-*Ighv* CLL type.

Three primary tumors along with their transferred samples were analyzed (figure 3.1). V(D)J genes of the color coded clonotypes are described in the legend (figure 3.1 d). Number of total reads identified for each sample are described below each donut plot.

Two kinds of clonotype evolution patterns were identified in Eμ-TCL1 mouse tumors:

1. A new clonotype emerged as a major clone in subsequent transplantations: Tumors A3.1 and C2.3 showed this pattern, where the new major clonotype was previously undetected in the primary tumor. Such a pattern could be attributed to changing tumor microenvironment after transplantation, acquired new mutations, outgrowth of a previously small but aggressive subclone or heterogenous CLL course wherein more than one selected autoantigen seems to be driving the disease. Oligoclonality is observed in 10 of the 13 tumors depicted in figure 3.1. This is in contrast to patient CLL, which is mostly monoclonal (5-24% of total CLL cases (Darwiche, Gubler et al. 2018)).

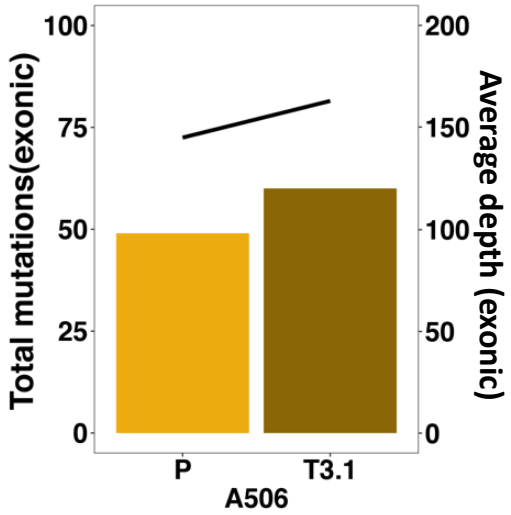
2. Monoclonal disease persists in transplanted tumors, B2.1, B2.2 and B2.3 (figure 3.1 b). Restricted BCRs are linked to severe disease and aggressive course in CLL patients (Yan, Albesiano et al. 2006, Sarkar, Liu et al. 2016).

Major clonotype in A3.1 with *Ighv12-3* gene is identical to the major clonotype in B1.1 and its subsequent transplantations (B2.1, B2.2, B2.3). *Ighv1-55*, *Ighv11-2*, *Ighv12-3* detected in 5 out of 12 clonotypes in 13 samples have previously been reported to be present as stereotyped BCRs in the CLL TCL1 mouse model. It has been reported that TCL1 Tg (transgenic mice) BCRs using - *Ighv11* and *Ighv12* genes are cross reactive with phosphatidylcholine (PtC). One of the ways of CLL progression is accelerated by preferential expansion of these BCRs to the autoantigen PtC (Chen 2010). Hence, CLL in at least six (A3.1, C2.3, B1.1, B2.1, B2.2, B2.3) mice can be attributed to chronic stimulation of BCRs by autoantigens. But PtC negative mice have also been known to show a BCR involving *Ighv11*. In addition, reports also suggest that combinations of preferentially selected light chains and virus specific heavy chains gives B cells the ability to recognize broad range of autoantigens that could contribute to leukemia progression (Jimenez de Oya, De Giovanni et al. 2017). There is, however, no information about viral infection in the analysed mouse tumors.

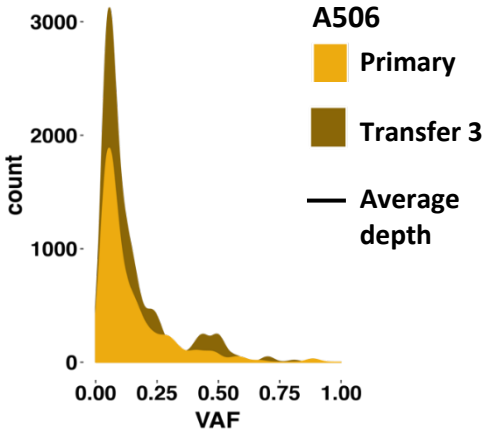
3.2 Mutation load increases with subsequent tumor transplantations while low allele frequency mutations persist

In-house whole exome sequencing (WES) was performed for B cells from spleens of 8 Eμ-TCL1 mice manifesting CLL-like symptoms (Dr. Selcen Özturk and GPCF at DKFZ). Out of these eight, four were primary tumors and 4 secondary tumors transplanted from them. For each of the primary cases a matched control was also sequenced, which was either T cells from spleen or tail of the mice. More information about the 8 samples is presented in Supplementary Table 1. I used Mutect2 to identify somatic mutations (Benjamin, Sato et al. 2019). In addition, publicly available CLL WES data (SRP150049) for three primary Eμ-TCL1 mouse tumors and their subsequent transplantations was included for analysis (Zaborsky, Gassner et al. 2019). Figure 3.2 gives an overview of somatic mutation trends in in-house CLL cohort (figure 3.2 a, b, c, d) and public dataset (figure 3.2 e, f, g). Since the public dataset was processed in Salzburg, those samples are denoted by sample number followed by (S).

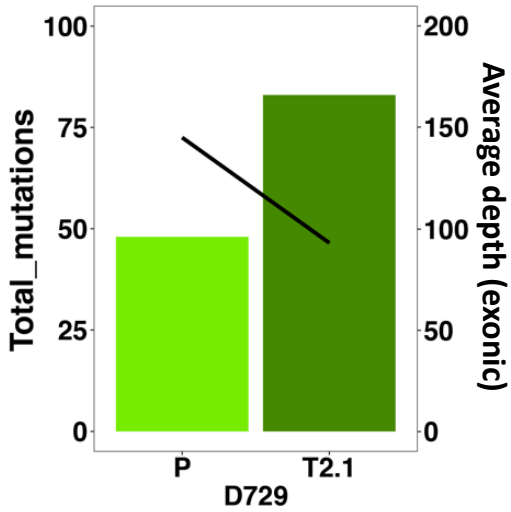
a. i.



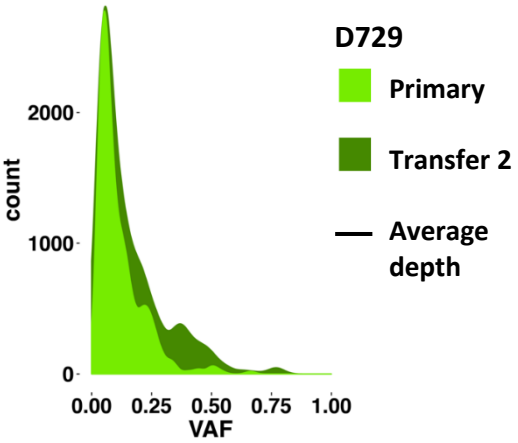
a. ii



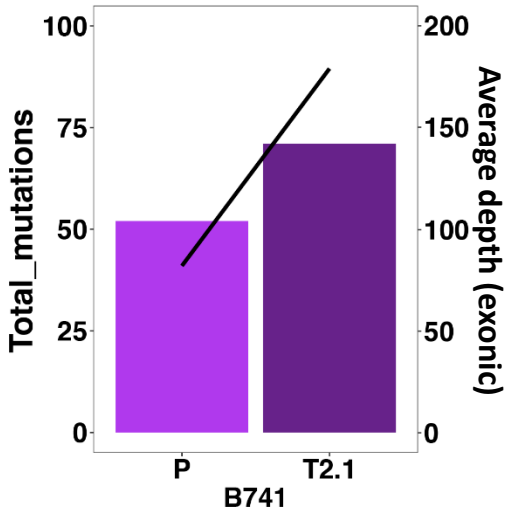
b. i.



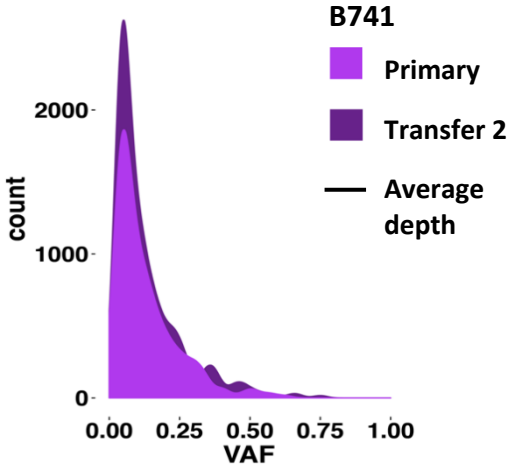
b. ii



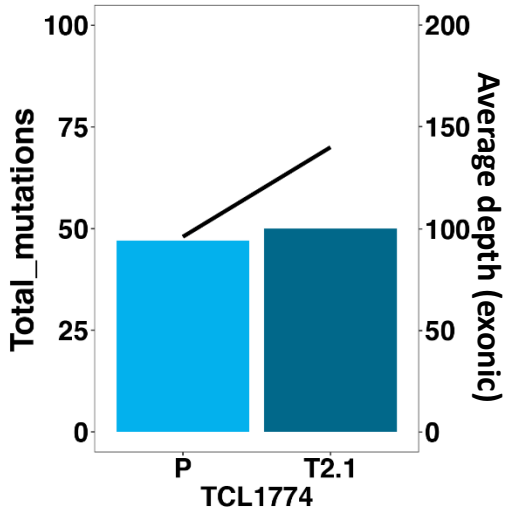
c. i.



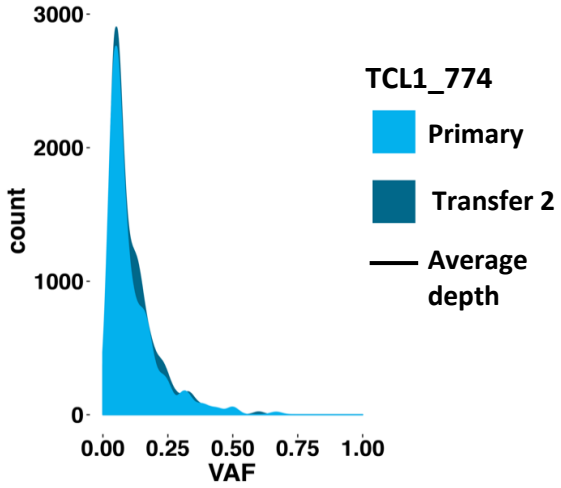
c. ii.



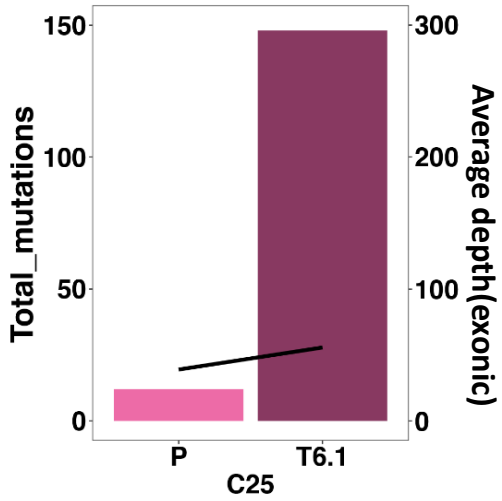
d. i.



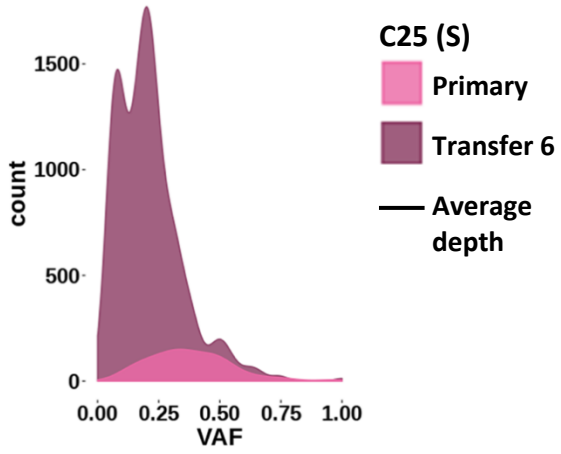
d. ii.



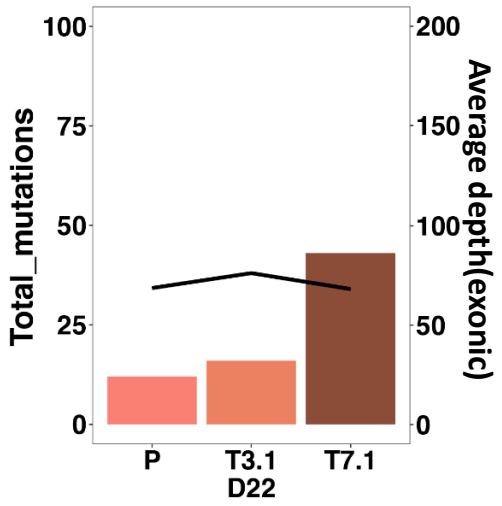
e. i.



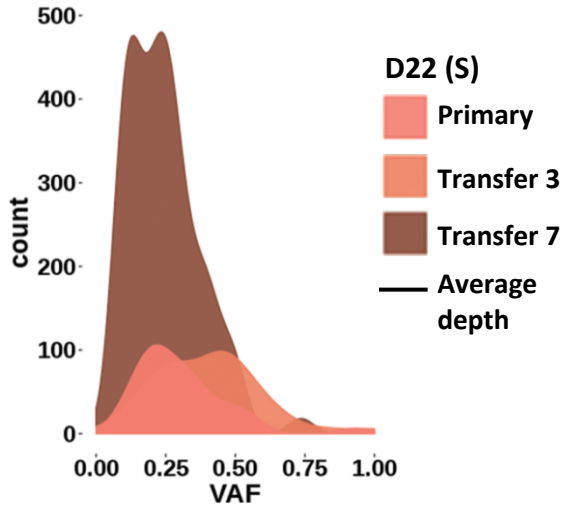
e. ii.



f. i.



f. ii.



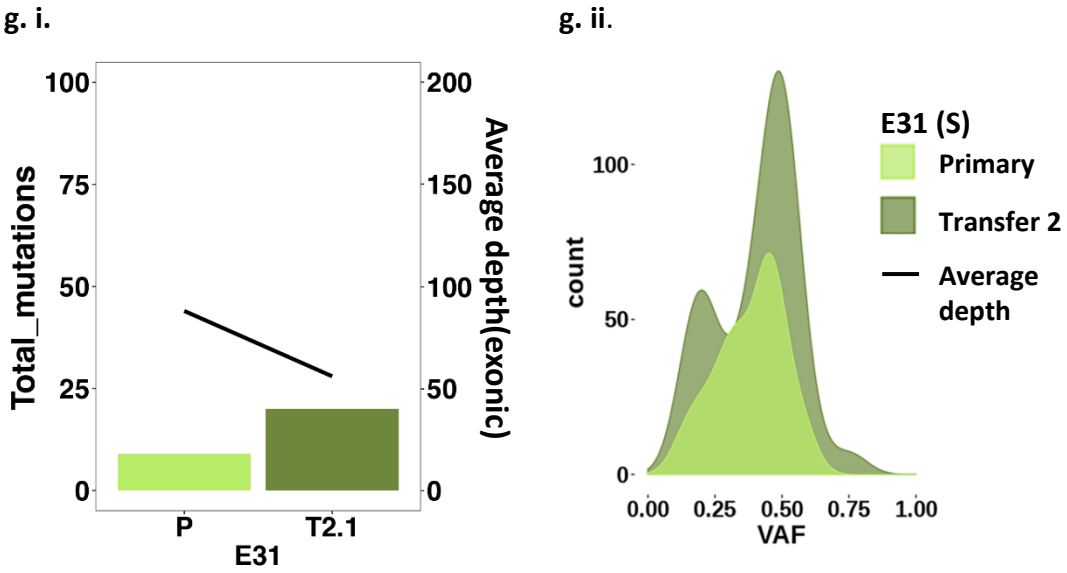


Figure 3.2: Column on the left shows an increase in number of variants with subsequent tumor transplantations in 7 TCL1 mouse tumors (refer to the left axis). It is also shown that this increase in number of variants is not necessarily correlated with increase in average depth at the region (refer to the right axis). Column on the right shows variant allele frequency (VAF) distributions all mutations in all primary and subsequent tumors transfers. Transfer number of the tumor is shown as, Example T3.1 is 3rd transplantation and T7.1 is the seventh round of transplantation starting from the primary tumor. All tumors have data from 2 time points/transplantations except D22 (S) which has data from 3 time points. Salzburg cohort samples are indicated with (S) alongside their sample names.

The following was concluded from single nucleotide variation (SNV) profiles of the two cohorts:

1. Left column of figure 3.2 depicts number of mutations (left axis) and average depth (right axis) at the target (exonic) region for Heidelberg and Salzburg primary and transplanted tumors. Number of mutations in the target region increased with each transfer in all 7 tumors. The fact that increasing mutation numbers with adoptive transfer did not necessarily correlate with increase in average depth (for example in b. i., f. i. and g. i.) at exonic regions, indicated growing mutation load (purely biological and not technical) with tumor evolution.

2. Right column of figure 3.2 shows variant allele frequency (VAF) distribution of all identified SNVs in primary and secondary tumors. In the Heidelberg samples (a. ii., b. ii., c. ii., d. ii.), the allele frequency peaked between 5% - 10%. However, allele frequency of mutations from Salzburg samples (e. ii., f. ii.) peaked at 25% and for (g. ii.) peak at 50%.

Low allele frequency mutations also occur in CLL patients (Guieze and Wu 2015). Identification of mutations with low VAFs could imply existence of several clones and subclones that would eventually evolve variedly when subjected to intrinsic and extrinsic factors like the tumor microenvironment and therapy pressure respectively.

Therefore, I next inspected the clonal evolution dynamics of these serially transplanted tumors with respect to somatic mutation as well as their BCRs.

3.3 CLL clonal evolution dynamics in Eμ-TCL1 mice exhibit three kinds of patterns

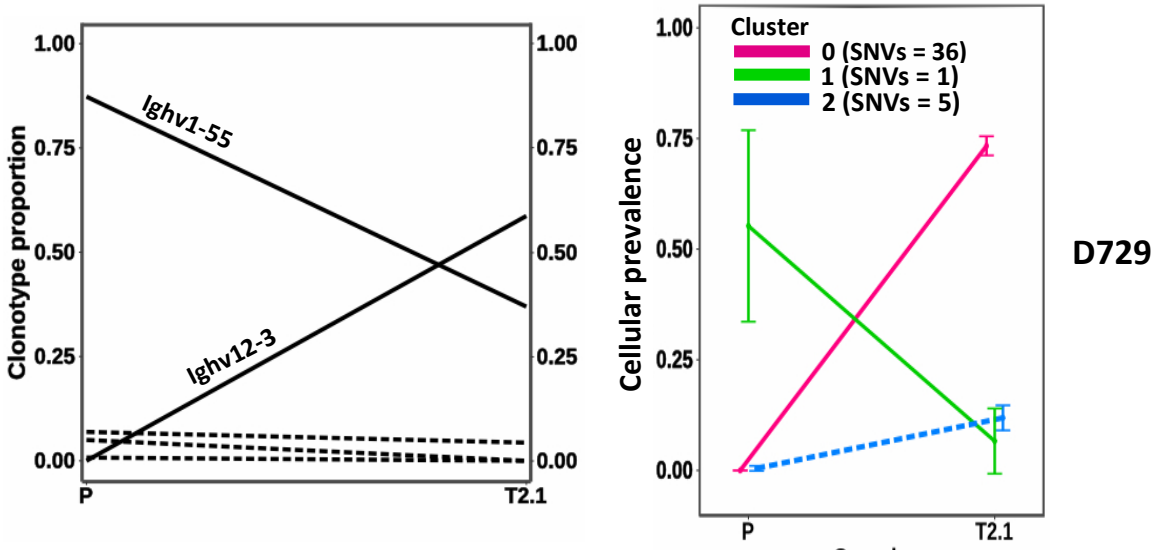
Chronic lymphocytic leukemias are transformations of mature and differentiated B cells. Transformation to malignant B cells can be attributed to chronic stimulation of the BCR (by persistent or intermittent exposure to an antigen) as well as constitutive and acquired mutations. However, it is not clear which of these events provides a selective advantage for the mature B cell to persist, relentlessly proliferate and respond differently to signals from the microenvironment.

The cohort of mouse tumors used for this analysis was the same as shown in section 3.2, i.e. 4 sets of primary and transplanted Eμ- TCL1 mouse tumors from Heidelberg (in-house cohort), and 3 sets of publicly downloaded tumors (Salzburg cohort). FASTQ files from WES data were used as input for MiXCR to infer V(D)J clonotypes, and Mutect2 was used to identify somatic variants (Bolotin, Poslavsky et al. 2015, Benjamin, Sato et al. 2019).

Filtered variants (only somatic SNVs), their allele frequencies and copy number states as estimated using CNVkit, were subsequently used as input for PyClone that evaluated clusters of putative somatic clones and changes in the fractions of their cellular prevalence from primary to transplanted tumor (Roth, Khattra et al. 2014, Talevich, Shain et al. 2016). Steps for both the analyses are detailed in the methods section along with the thresholds and criteria used.

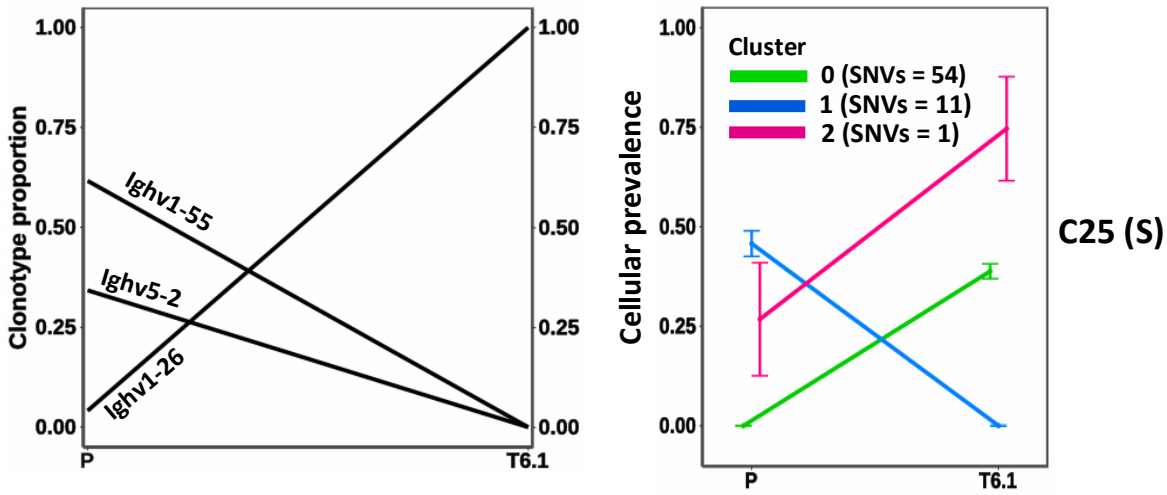
BCR clonotypes **SNV-defined clones**

a.



D729

b.



C25 (S)

Tracking the patterns of BCR evolution and changes in somatic mutation cellular prevalence, three different patterns of CLL progression emerged (figure 3.3):

1. Displacement of one clonotype by a novel one identified by BCR clonotype change and the same change in the SNV-defined subclones. This pattern is indicative of an expanding new major clone that could be driven by acquired novel somatic SNVs in the secondary transplantations. Figure 3.3 a, b, and c (samples D729, C25 (S), E31 (S)) represent this pattern. These samples showed change from one stereotyped TCL1 mouse BCR in the primary tumor to another stereotyped TCL1 mouse BCR in the transplanted tumor. E.g.: from *Ighv1-55* in T1.1 to *Ighv12-3* in T2.1 (figure 3.3 a).

2. Stable BCR clonotype but ongoing SNV-defined subclone change indicative of a mutating tumor clone that might be unstable. Figure 3.3 g (sample A506) is an example of such a process.

3. Stable clonotype proportions and SNV-defined subclones indicative of a stable disease course without novel somatic SNVs. Figure 3.3 d, e and f (samples B741, TCL1_774, D22 (S)) show this pattern. In this case, the same stereotyped BCR was consistently expressed in both primary and transplanted tumors.

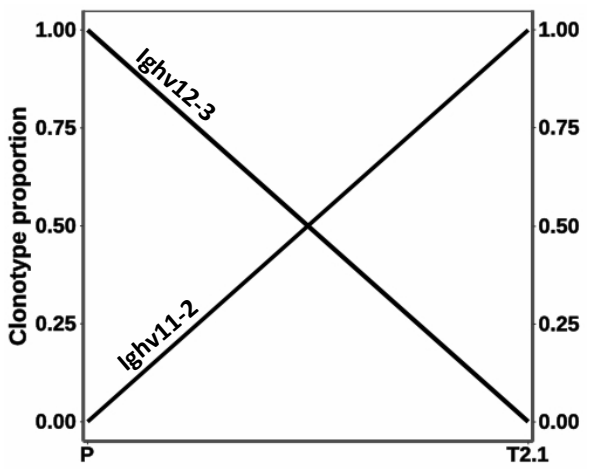
From these observations, the course of CLL evolution in mice can be attributed to either a BCR clonotype change from primary to transplanted tumor which is an indicative of displacement of one tumor clone by another (potentially a more stable one); or ongoing somatic mutations with stable BCR clonotype proportions demonstrating ongoing evolution on the genetic level.

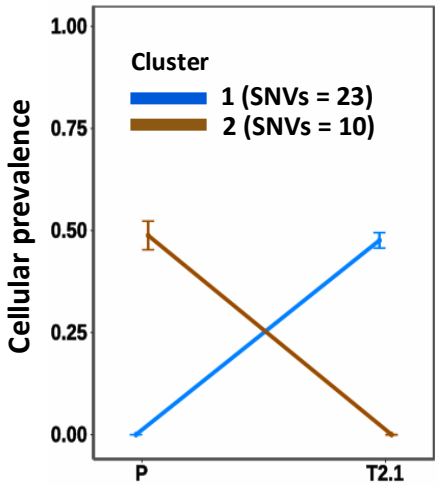
Patterns described in points 1, 2 and 3 are similar to ones previously associated with CLL pathogenesis. There have been reports that support CLL evolution by selection of somatic variants that can induce enhanced B cell receptor signaling by stronger binding affinities between the BCR and autoantigens (Domenech, Gomez-Lopez et al. 2012).

BCR clonotypes

SNV-defined clones

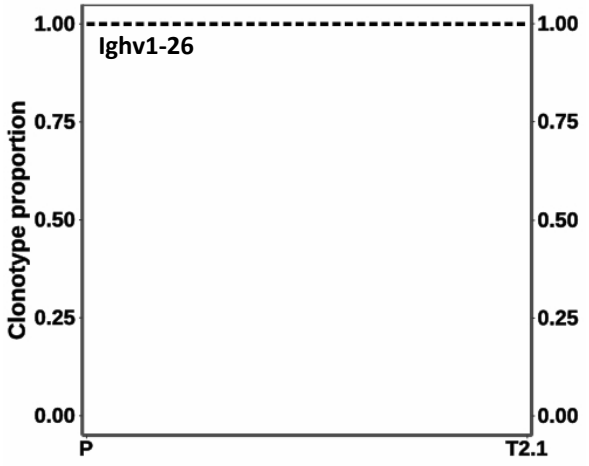
c.

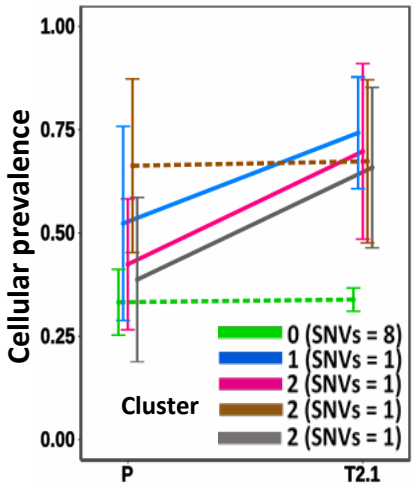




E31 (S)

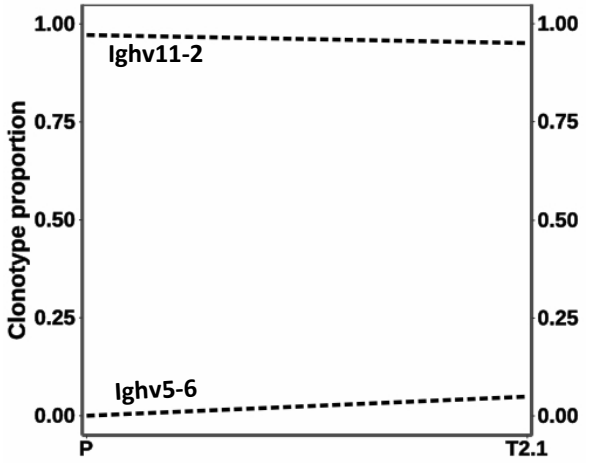
d.

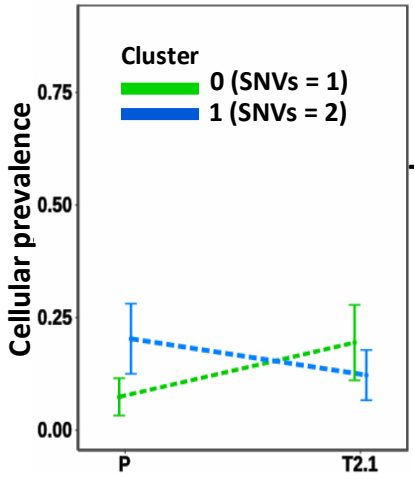




B741

e.





TCL1_774

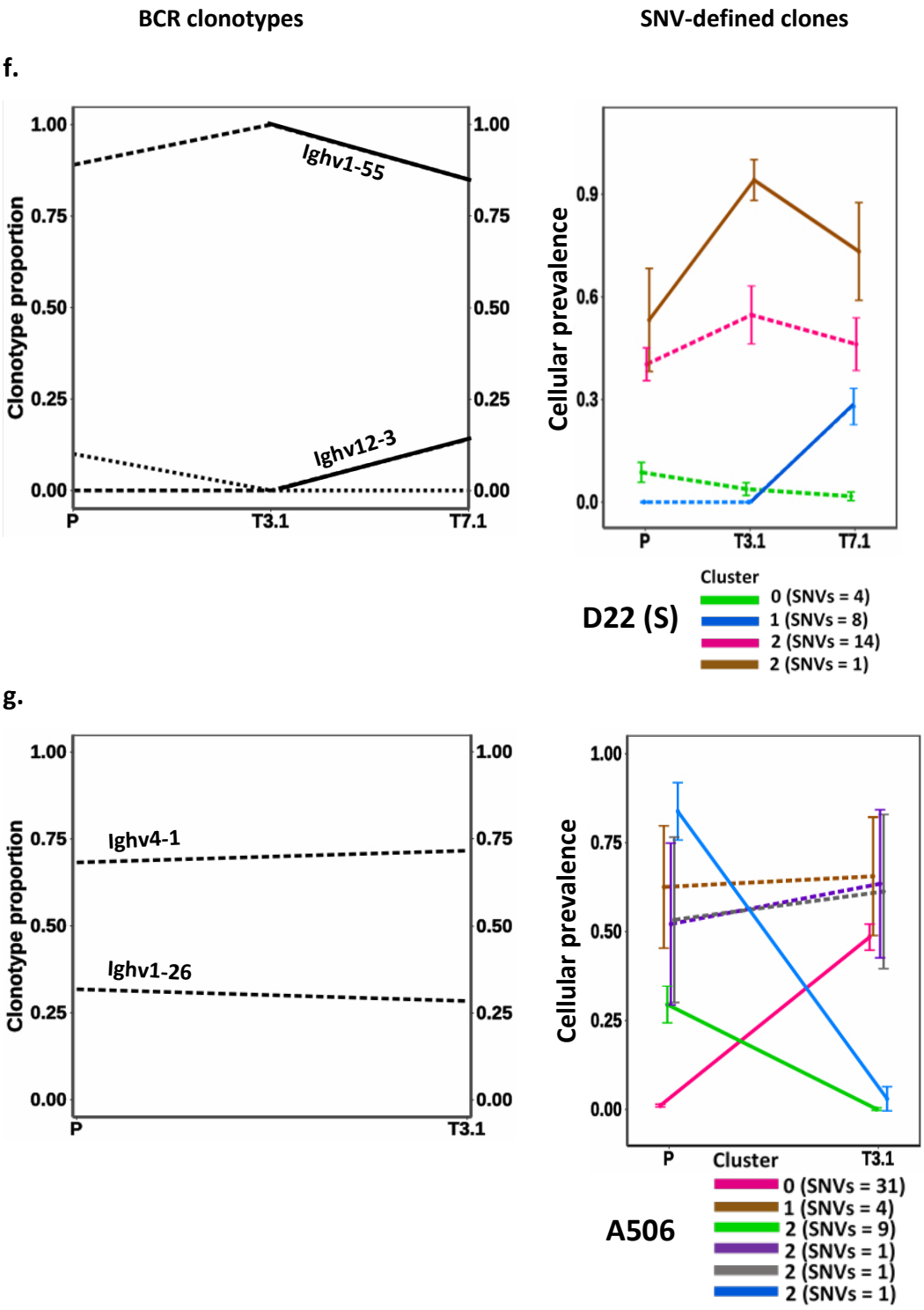


Figure 3.3: (a-g) Clonal evolution attributed to somatic mutations (right) and presence of one or more BCRs (left). BCRs and cellular prevalence that change by less than 10% between primary and transfers are marked by a dotted line. Pattern 1: a, b, and c (samples D729, C25 (S), E31 (S)); pattern 2: g (sample A506; pattern 3: d, e and f (samples B741, TCL1_774, D22 (S)). Transfer number of the tumor is shown as, Example T3.1 is 3rd transplantation and T7.1 is the seventh round of transplantation starting from the primary tumor. All tumors have data from 2 time points/transplantations except D22 (S) which has data from 3 time points. Salzburg cohort samples are indicated with (S) alongside their sample names.

3.4 Trisomy 15 corresponding to *Myc* over expression might be essential contributors to CLL pathogenesis

A peculiar observation within copy number analysis was the frequent amplification of chromosome 15 in six out of 8 Heidelberg tumors and 10 out of 11 Salzburg cases (including the ones not used for analysis by myself), as also pointed out in their paper (Zaborsky, Gassner et al. 2019). On checking the genes lying in the amplified chromosome region 15, *Myc* oncogene was identified. It has been claimed in literature that TCL1-tg mice (a similar mouse model that instead develops a disease similar to human T cell prolymphocytic leukemia) exhibits trisomy 15 and over expression of *Myc*, which are essential contributors for malignant transformation (Shen 2006). It was therefore proposed to validate *in-vitro Myc* amplification in the in-house Eμ-TCL1 mouse tumors, and comment if it's over expression was often a driver in CLL pathogenesis along with TCL1 over expression. Validation of this follow up experiment would genetically make this model very different from the CLL genetics in patient settings, raising speculations about its use in studying CLL.

3.5 Effects of ibrutinib treatment on clonality of Eμ-TCL1 mouse tumors

Eμ-TCL1 mice were analyzed for their transcription profiles, at time points 1-, 3- and 6-weeks post ibrutinib treatment start. Splenocytes were isolated and flow sorted for CLL (CD5+ CD19+) cells on the same day for all three time points (mouse work and experimentation performed by Haniyeh Yazdanparast). DNA was isolated and the whole exome was sequenced (HiSeq 4000 platform) to investigate genetic changes as a result of ibrutinib treatment and resistance. To identify effects of ibrutinib resistance on the clonality of Eμ-TCL1 tumors, ibrutinib late (6 weeks, n=3) and ibrutinib early (1 week, n=4) tumors were compared to vehicle late (6 weeks, n=4) and vehicle early (3 weeks, n=4) tumors. Tumors from ibrutinib late time point showed potential signs of resistance after which ibrutinib treatment had no effect, and the tumors started to grow again.

Ibrutinib treatment may impact CLL development of the tumors at three genomic levels: Copy Number Variations (CNVs), dynamics of the B cell receptor (BCR) and somatic single

nucleotide variations (SNVs). Below, I detail my observations from analysis of CNVs, BCR dynamics and SNVs in 15 tumors included in this study.

Copy number profiles of ibrutinib treated mouse tumors

CNVs were analysed from WES data using CNVkit tool as described previously in this thesis. No focal changes were observed in ibrutinib treated tumors at the late time point as compared to ibrutinib treated tumors at the early time point. These tumors, however consistently showed amplification of chromosome 15 (trisomy 15), same as the primary Eμ-TCL1 mouse from which tumors for this study were serially transplanted. Presence of trisomy 15 in Eμ-TCL1 mouse tumors has been detailed in section 3.4 of this thesis as a potential genetic predisposition for leukemia in this mouse model.

Impact of ibrutinib treatment on dynamics of BCR rearrangements

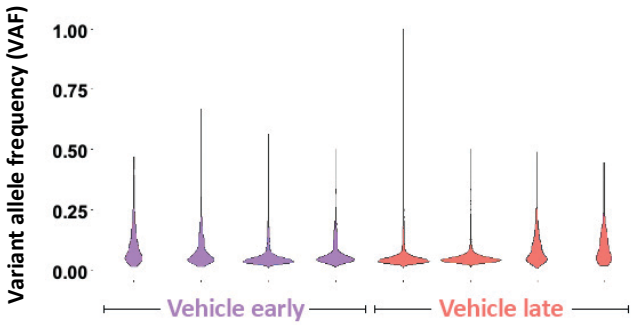
Also, on probing into type of B cell receptor rearrangement conferred on these 15 tumors at different time points of ibrutinib and vehicle treatments, I identified the same BCR clonotype in all 15 tumors. The identified rearrangement included the *Ighv11-2* gene. This gene is a part of a known stereotype mouse BCRs (as cited before in this thesis), and expands in response to autoantigens, adding to CLL pathogenesis by means of chronic stimulation of the BCR. Importantly, the primary tumor from which other tumors used for this analysis were serially transplanted and treated with ibrutinib, was also found to be monoclonal and harboured the same *Ighv11-2*.

Identification of somatic SNVs from ibrutinib treated tumors

I then investigated the SNV landscape of these tumors. Somatic mutations were called using Mutect2. Variant allele frequency (VAF) distribution of identified mutations in vehicle treated (early, n=4; late, n=4) and ibrutinib treated (early, n=4; late, n=3) groups are shown in figure 3.4 a and b respectively. No difference was observed between treated and untreated samples with respect to the VAF distribution, as allele frequencies of identified variants across all samples peaked at less than 5%. This is indicative of presence of several SNV-defined subclones in the tumor. The aim here was however, to identify mutations specific to ibrutinib late time point, that is, mutations causing the tumors to become resistant. For this, mutations identified from tumors at vehicle late, vehicle early and ibrutinib early time points were

intersected out and only mutations identified from tumors at ibrutinib resistant time point were considered

a.



b.

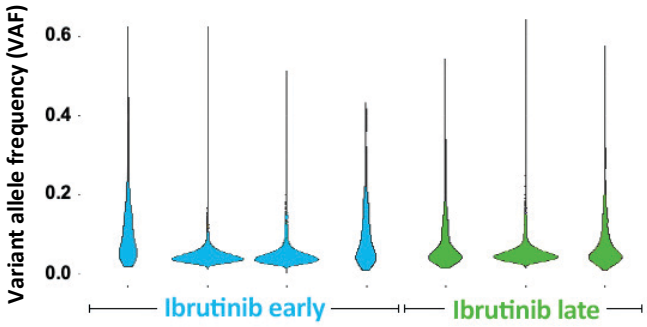


Figure 3.4: (a) Variant allele frequencies (VAFs) of identified mutations from vehicle treated tumors at early and late time points. (b) Variant allele frequencies (VAFs) of identified mutations from ibrutinib treated tumors at early and late time points.

Interestingly, none of the ibrutinib late time point tumors manifested mutations in the known *Btk* and *Plcg2* genes, as reported in ibrutinib resistant patient cases. Also, none of the identified mutations were linked to Bcr signalling pathway.

From the above observations, it was concluded that the development of ibrutinib resistance within just 6 weeks of treatment in Eμ-TCL1 mice, is a relatively short period for identifiable genomic changes (CNVs/SNVs) to occur as a response to selection pressure from the drug. It was therefore hypothesized that ibrutinib resistance in the mouse model is visible by changes in transcriptional profile of the tumor rather than the genetic profile. Thus, I next examined

the transcriptional changes in ibrutinib treated v/s untreated tumors at late and early time points.

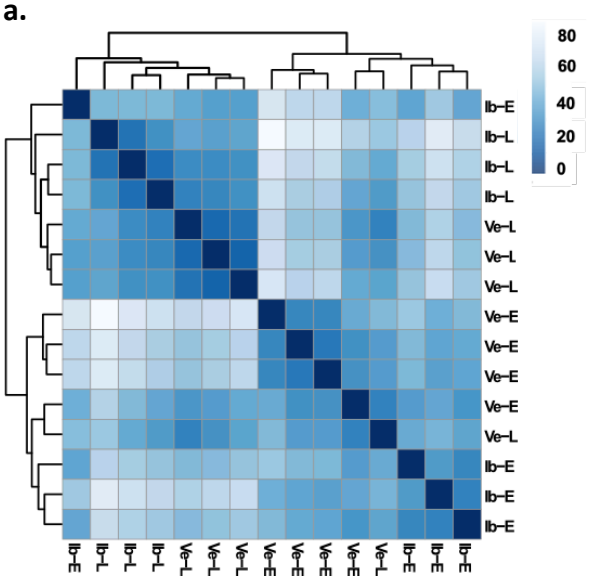
3.6 Transcriptional profile of ibrutinib resistant Eμ-TCL1 mouse tumors

To identify transcriptional changes attributed to ibrutinib resistance in Eμ-TCL1 mouse tumors, RNA sequencing of CD19+ B cells from the spleens of ibrutinib and vehicle treated tumors at early and late time points was performed (by Haniyeh Yazdanparast, described in methods section).

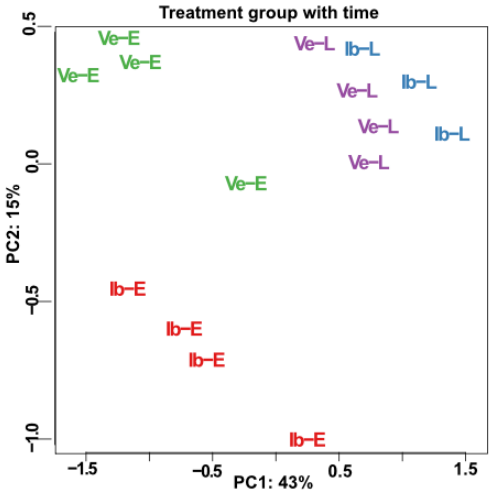
Hierarchical clustering of sample wise normalized and log transformed gene counts obtained by pre-processing raw RNA sequences of the 15 samples is shown in figure 3.5a. Clustering was based upon a distance matrix calculated using the thousand most variable genes across all samples (pairwise Euclidean distance and complete linkage). It can be seen from the plot that most of the variability in the samples is conferred by the sampling time point, as the late treated samples (ibrutinib and vehicle) clearly cluster together and separate from the early time point samples (except for one sample in each late and early time points that show otherwise). It can also be identified that vehicle early samples have the most distinct profile as compared to the late treated group.

Overall differences between ibrutinib and vehicle treated tumor groups (biological) and presence of any potential batch effects (technical) are shown in the form of a 2D principal component plot depicting the first (PC1) and the second (PC2) principal components (figure 3.5 b). Expression across all genes was used as input for principal component analysis (PCA). PC1 separates the treatment groups based on sampling time (early and late) and indicated heterogeneity within the same groups (technical variability of 43%, e.g.: ibrutinib early (Ib-E)). PC2 clearly separates ibrutinib early treatment group from the rest three groups (variability: 15%).

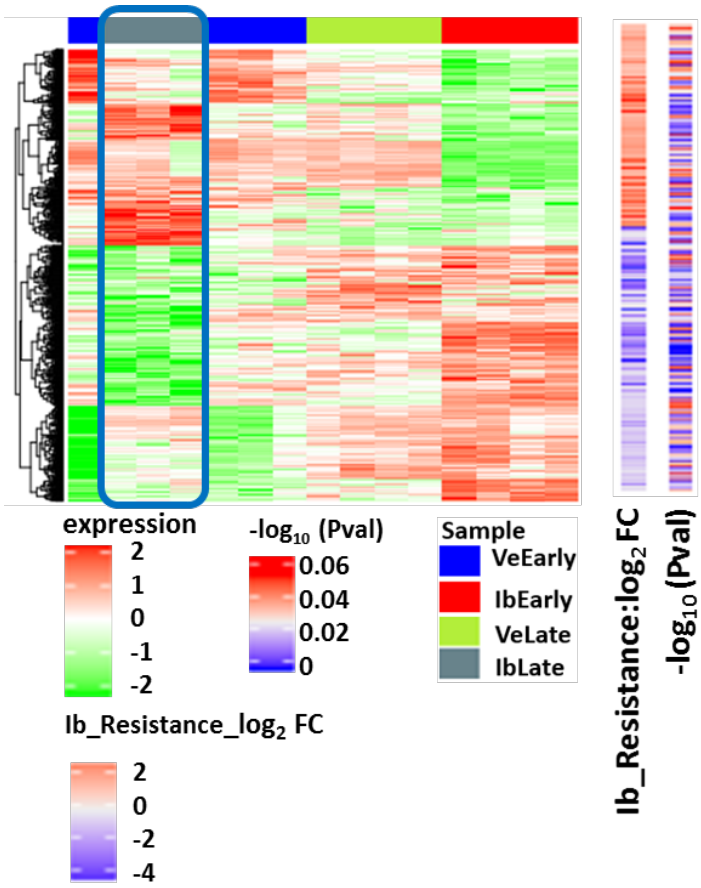
a.


80
60
40
20
0
Ib–E
Ib–L
Ib–L
Ib–L
Ve–L
Ve–L
Ve–L
Ve–E
Ve–E
Ve–E
Ve–E
Ve–L
Ib–E
Ib–E
Ib–E

b.


Treatment group with time
PC2: 15%
PC1: 43%
0.5
0.0
−0.5
−1.0
−1.5
−0.5
0.5
1.5
Ve–E
Ve–E
Ve–E
Ve–L
Ib–L
Ve–L
Ib–L
Ve–L
Ib–L
Ve–L
Ve–E
Ib–E
Ib–E
Ib–E
Ib–E

c.


expression
2
1
0
-1
-2
-log10 (Pval)
0.06
0.04
0.02
0
Sample
VeEarly
IbEarly
VeLate
IbLate
Ib_Resistance_log2 FC
2
0
-2
-4
Ib_Resistance:log2 FC
-log10 (Pval)

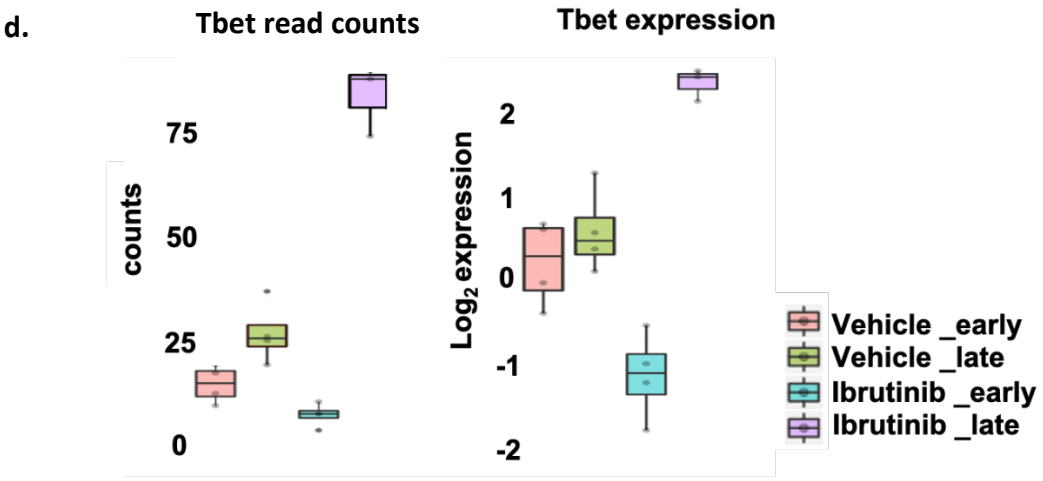


Figure 3.5: (a) Unsupervised hierarchical clustering of the 1000 most variable genes across all samples (distance: Euclidean, linkage: Complete). (b) PCA plot showing sample clustering using expression of all genes. (c) Differentially expressed genes (DEG) in ibrutinib resistance phenotype as compared to the three other groups. Ibrutinib resistance phenotype is boxed in blue in the heatmap (distance: Euclidean, linkage: complete). Log_2 fold change with significance of DEG in ibrutinib resistant phenotype are shown as vertical bars alongside the heatmap. (d) Tbet counts (left) and expression on log_2 scale (right) at vehicle and ibrutinib treatment time points 'early' and 'late'. Ib-E/IbEarly: ibrutinib early, Ve-E/VeEarly: vehicle early, Ib-L/IbLate: ibrutinib late, Ve-L/VeLate: vehicle late.

Separate clustering of Ib-E tumors could be attributed to their being the only ones manifesting the effects of ibrutinib treatment and, hence showing a different biological phenotype and transcription state.

Also, it is not unexpected that the vehicle late (Ve-L) treated tumor group lies in between vehicle early (Ve-E) and ibrutinib late (Ib-L), as the phenotype of vehicle late tumors was both of increased cell proliferation (similar to ibrutinib late) and vehicle treatment (similar to vehicle early).

To identify genes that specifically confer ibrutinib resistance in Eμ-TCL1 mouse tumors, a limma model ((Ritchie, Phipson et al. 2015), R/Bioconductor software package) was constructed that eliminated effects of proliferation (expected to be enhanced in the late treatment time points: Ve-L and Ib-L) and treatment (as expected in the ibrutinib early time point: Ib-E) (explained in detail in methods section). 803 genes were differentially regulated in ibrutinib resistance group as compared to other three groups at adjusted p-value < 0.05 (Benjamini and Hochberg method). Unsupervised hierarchical clustering of the expression of

803 genes in the four groups (vehicle early, vehicle late, ibrutinib early and ibrutinib late) is shown in figure 3.5 c (Euclidean distance and complete linkage). Out of these, 456 genes were downregulated and 347 were upregulated in ibrutinib late treatment group as compared to others. Transcription profile of the ibrutinib late group (marked with a blue box) was distinct from that of the other groups, indicating pronounced transcriptional changes driving ibrutinib resistance in the mouse tumors.

One of the top upregulated genes in the ibrutinib resistance phenotype, *Tbet*/*Tbx21* was chosen for follow up studies. Its increased expression (> 2 fold on $\log_2$ scale) in ibrutinib late group as compared to others is shown in figure 3.5 d (left: read counts, right: $\log_2$ scale). *Tbet* has previously been reported to enhance survival of B cells. It possibly controls chronic inflammation in autoimmune diseases and in ankylosing spondylitis with a potential therapeutic role (Weigmann and Neurath 2002, Barnett, Staupe et al. 2016, Vecellio, Cohen et al. 2018). Studies of the tumor microenvironment showed its involvement in tumor immune surveillance in lung cancers (Reppert, Boross et al. 2011). Observations from this analysis made it possible to hypothesize that Tbet might represent a novel target to control ibrutinib resistance in the Eμ-TCL1 mouse model. Follow-up experiments are currently being performed (by Dr. Lavinia Arseni) to validate the role of *Tbet* in ibrutinib resistance of mouse tumors.

CyTOF single cell analysis

CLL perpetuation as well as progression has been directly linked with a supportive tumor microenvironment (TME). The CLL (TME) is composed of both stromal and immune cells. With such deep infiltration of immune cells in the tumor, reinvigoration of the immune system to stimulate anti-tumor activity is a promising mode of therapy for CLL. Anti-PD-1/PD-L1 antibody inhibitors are increasingly becoming therapeutically useful for establishing continuous immune surveillance in several malignancies (Hodi F.S. 2010, Topalian S.L. 2012, Xiaomo Wu 2019).

Using information about surface marker intensities (relative abundances) from CyTOF (mass cytometry) analysis, I identified T cell subpopulations in the tumor microenvironment (TME) of CLL at the single cell level. 43 surface markers focusing on T cell phenotyping were measured[2]. The following subsections describe these subpopulations characterized by the intensities (expression) of one or more surface markers in detail. I also investigated differential abundance of these subpopulations across CLL lymph node (LN) samples from 23 patients, out of which peripheral blood (PB) and bone marrow (BM) was also available from 8 and 3 patients respectively. CLL T cell subpopulations were also compared in between tissues and to control LNs (n=13). Raw data was obtained in the form of .fcs files which were then processed for quality and downstream analysis to infer the observations detailed next. Both CD4+ and CD8+ cell types were assessed using CyTOF. However, the scope of this thesis is limited to observations on the CD4+ T cell subtype.

[2] CyTOF set up and experiment was performed in collaboration with Luxembourg Institute of Health (Marina Wierz, Etienne Moussay, Jérôme Paggetti). Laura Llao Cid and Martina Seiffert from DKFZ designed the panel. After clustering by myself, annotation of the identified subpopulations was performed by Laura Llao Cid.

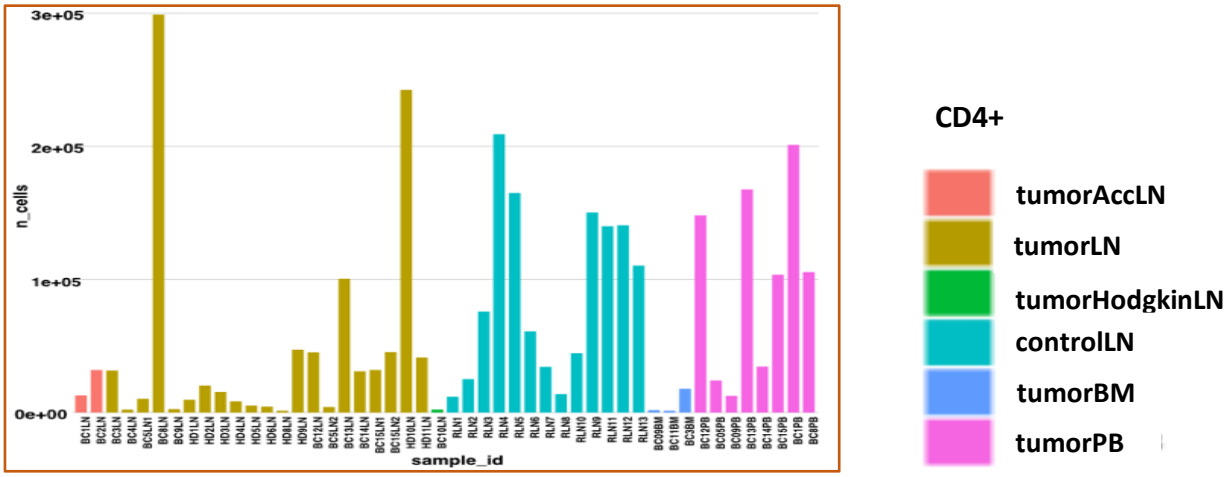
3.7 Quality control diagnosis of CyTOF samples

The total number of cells profiled across 48 samples was highly variable as displayed in figure 3.7 a for CD4+ sorted sample subsets. Samples included in the analysis had at least 1000 cells. To identify any peculiar similarities between patients and potential technical outlier samples in an unsupervised manner, the samples were plotted on an MDS plot (Multi-Dimensional Scaling; distance measure: pairwise Euclidean, linkage: complete). The input for this plot were arcsinh-transformed median marker expression values for 32 surface markers (used to measure biological phenotypes) and 11 technical markers (used to identify in technical biases) mapped to 48 data points (number of samples=48). E.g.: 191Ir DNA1 and 193Ir DNA2 markers are expressed by single cells only, and this helps filter out potential doublets. Details of the samples used for CyTOF analysis is presented in Supplementary Table 2. CyTOF markers are detailed in Supplementary Table 3. Figure 3.6 b is an MDS representation for CD4+ samples.

Expectedly, the one sample of Hodgkin lymphoma (in green) was different from other CLL samples, due to it being a completely different disease. Tumor PB (pink) mixed with control LN (cyan). Certain control LN (cyan) (RLN3, 6, 8 and 12) mixed with tumor LN (brown). It will be interesting to further investigate whether these samples that mix together also have similar phenotypic properties. No technical bias was identified from this step.

After initial quality checks, it was decided to identify CD4+ and CD8+ cell subpopulations, and CD4+ subpopulations are discussed in this thesis.

a.



b.

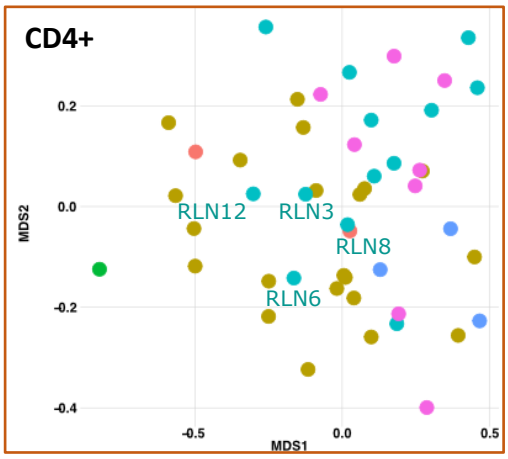


Figure 3.6: Sampling across 6 conditions (tumorLN, controlLN, tumorPB, tumorBM, tumorAccLN, tumorHodgkinLN) (LN=lymph node, PB=peripheral blood, BM=bone marrow). (a) histogram of sample counts for CD4+ samples. (b) MDS plot clustering (distance measure: pairwise Euclidean, linkage: complete, 48 mapping points) using median marker expression of 43 markers (32 biological + 11 technical across all cells in each sample of CD4 cell type).

3.8 CD4+ subpopulations and their known and potential contribution to the tumor microenvironment

All cells from all samples CD4+ T cell type were clustered by FlowSOM method (Van Gassen, Callebaut et al. 2015) as a functionality of the HDCytoData workflow (Nowicka, Krieg et al. 2017) in Rv3.6. This included a stepwise process of building a self-organizing map (SOM), and meta-clustering of these SOM codes. Even though it was made sure that the clustering is performed using all the cells from each sample, for purpose of plotting and visualization, 1000 random cells from each sample were selected and visualized using t-SNE dimension reduction

technique. Hence, every sample had equal representation despite differences in library sizes. The FlowSOM approach identified 15 CD4+ subpopulations as shown in figure 3.7 a. These populations were first annotated broadly into naïve, T central-memory (Tcm), T effector-memory (Tem), and T effector (Tef) based upon expression of CCR7, CD45RA, and CD45RO surface markers as described in (Golubovskaya and Wu 2016). Annotation was performed by Laura Llao Cid. To further classify these broad groups, subpopulations were then named based upon the unique expression of a surface marker, e.g. CD4+ Tem1 and Tem2 were both effector memory but characteristically expressed Ki67,CD38 and CD39 respectively.

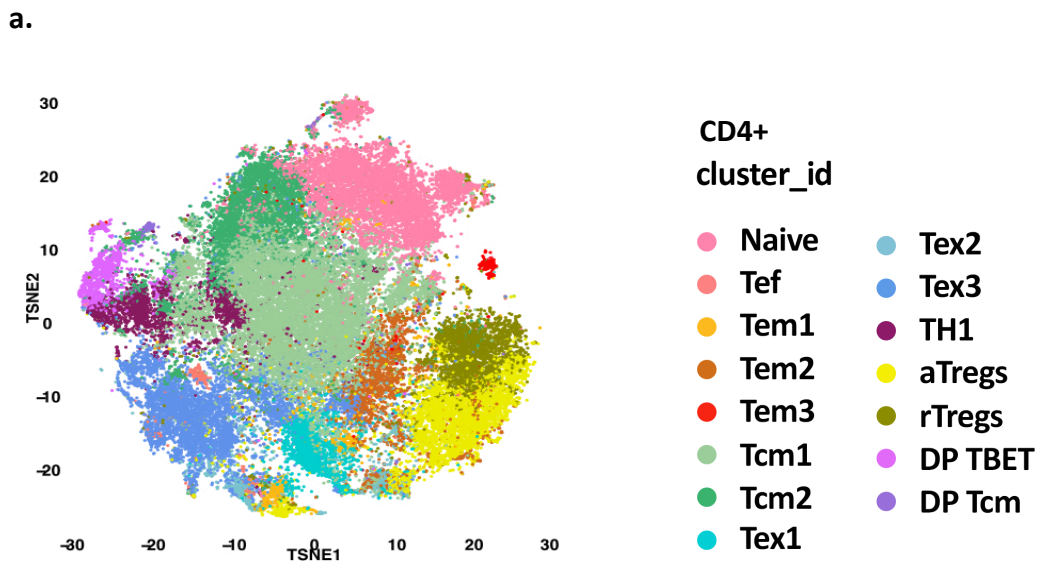


Figure 3.7: (a) 15 CD4+ subpopulations visualization by t-SNE plot. One naïve, one Tef (T effector), 3 Tem (T effector memory), 2 Tcm (T central memory), 3 Tex (T exhausted), one Th1 (T helper 1), two Tregs (aTregs: activated regulatory T cells, rTregs: resting regulatory T cells), and two DP (double positive) subpopulations were identified. Extended annotation based upon specific marker expression is as follows: Tem1=Tem1 Ki67+ CD38+, Tem2=Tem2 CD39+, Tem3=Tem3 ICOShi (high), Tcm2=Tcm2 CD39+, Tex1=Tex1 PD1hi (high), Tex2=Tex2 CD39+ Ki67+, Tex3=Tex3 EOMES+ PD1+, TH1=TH1 KLRG+ TBET+, DP TBET=DP Tem TBET+.

Heatmap of transformed and normalized intensities of 32 surface markers that identify 15 CD4+ subpopulations is shown in figure 3.7 b (distance measure: Pearson, linkage: complete). The 15 subpopulations represented row wise in the heatmap can be grouped broadly into 5 recognizable biologically and/or CLL relevant phenotypes represented column wise (in colored boxes from right to left). They are:

1. Set of *activation* and *naïve cell markers (red box)*. The former decides the fate of various cell types in the TME, e.g. CD7 and IL7RA regulate activation of CD4+ T cells. Cell populations negative for CD7 are identified in pathological conditions. Expectedly this group is separated from the exhausted subpopulations (negative for CD7). TCF1, regulating T cell development and response to infection is also clustered in this group. Naïve cell markers include CCD7 and CD27. These subpopulations can be speculated to hold the capacity to divide and interact with tumor neoantigens.

2. *Markers associated with regulatory functions of T cells (black box):* CD25 (expressed by Tregs), CD39 (expressed by Tregs and activated CD4+ and CD8+ T cells, and has a role in promoting an immunosuppressive environment in association with CD73), HELIOS (activation marker and expressed on a subset of Tregs), and FOXP3 (a Treg marker and high in patients with CLL). CD39+ CD73- T cell population has been found to be abundant in CLL patients and signifies inflammation (Raczkowski, Rissiek et al. 2018).

3. OX40 and CXCR5 (blue box) are involved in *differentiation process of Tfh cells* (Qin, Waseem et al. 2018). T follicular helper (Tfh) subsets are present in increased frequencies in advanced stage CLL patients, but their role in CLL is still unknown.

4. *Tumor inhibiting and CLL specific molecules (orange box)*: CTLA4 (highly expressed in CLL T cells, renders T cell proliferation and tumor inhibiting T cell functions)(Mittal, Chaturvedi et al. 2013), inhibitory molecules of exhaustion phenotype (PD-1 and TIGIT), transcription factor promoting exhaustion (TOX), CD38 (prognostic marker of CLL) (Matrai 2005), and FAS associated with poor survival in CLL (Groneberg, Pickartz et al. 2003). This group of markers is notably upregulated in exhaustion specific subpopulations and could be potential contributors of T cell inactivation.

b.

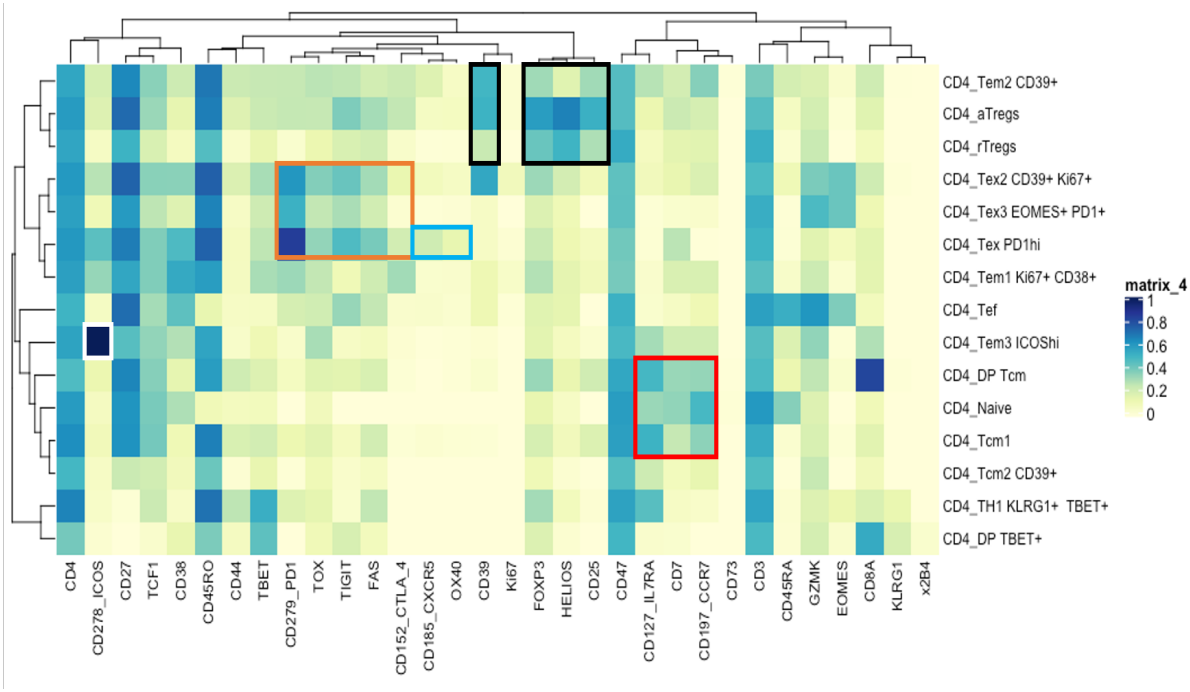


Figure 3.7: (b) Marker intensities (columns) across CD4+ subpopulations (rows). 5 biologically relevant subgroups are numbered from right to left and marked with different colored boxes are explained in the text (distance measure: Pearson, linkage: complete).

Identification of the above T cell phenotypes necessitates the study of how they interact and regulate the tumor niche by either promoting or supressing immune responsiveness in CLL. This is important because compromised immune response is one of the key reasons for treatment failure and disease relapse.

The next step therefore would be to associate the abundance of identified clusters (that have distributed expression of at least the 5 phenotypes described above) to CLL patient clinical information (age, gender, treatment etc). This could broadly inform us about observable shifts in T cell phenotypes across stages of disease development in CLL patients.

3.9 Associating CyTOF subpopulation abundances across samples to clinical information

Associating CD4+ subpopulation proportions in the CLL microenvironment to clinical parameters (age, IGHV status, treatment, gender and so on) across samples, can establish the clinical relevance of these populations in CLL pathogenesis. In addition, this enables identification of inter-patient heterogeneity with respect to their CLL microenvironment.

Abundances of identified CD4+ subpopulations per sample along with associated clinical parameters: age, gender, treatment and IGHV status were studied and are shown in figure 3.8. Subpopulation abundances (calculated as proportions) were then subjected to unsupervised hierarchical clustering (distance measure: Euclidean, linkage: complete), with the aim to group together patients manifesting similar microenvironment subpopulations. It should be noted that similar subpopulations have been represented with shades of the same color, e.g. all Tem subtypes have been shown with shades of yellow-brown, Tcm populations have been shown with shades of green.

Emerging patterns of samples with similar phenotypes are discussed below:

CD4+ cell type abundances

Based on CD4+ subpopulation abundance clustering, samples could be majorly grouped into 5 sets as numbered in figure 3.8 from top to bottom (1-5).

1. Tcm populations (in green) (Tcm1 and Tcm2 CD39+) occurred in all samples except for the ones that had Tem CD39+ T cell population (in brown) and increased aTregs (activated Tregs). With no naïve cell phenotype, these samples included a proportion of Tex3 EOMES+ PD1+ exhausted phenotype. These CLL LN samples showed mixed IGHV status and no gender bias; with most of them being untreated (6 out of 7 samples).

2. A group of 3 control LNs, 3 tumor LNs had naïve and Tcm subpopulations.

3. All control samples with central memory, naïve and Tex1 PD1hi phenotypes. Tex1 was specifically represented in the control group and could be attributed to acute or chronic infections in samples whose rLNs (reactive lymph nodes) were acquired.

4. Samples high in Tex3 EOMES+ PD1+ (darker blue) were all CLL patient lymph nodes, except for RLN6 and RLN10 (both control LNs), that in the MDS plot in section 3.7, also clustered with CLL samples. It could be suggested that they group together due to their similar exhausted phenotype.

5. Group 5.1 (2 tumor PB and 1 tumor BM) and group 5.2 (5 tumor PB and 1 tumor BM were identified. Major cell types contributing to these were Tcm1 and TH1 KLRG1+ TBET+. Group 5.1 in addition, also had some proportion of naïve subpopulation. These groups were almost devoid of exhausted T cells. Tumor PB and tumor BM were matched tumor samples from CLL patients from whom lymph nodes were acquired (example BC9LN, BC09BM, BC09PB), and the latter had a much higher exhausted phenotype (group 4) in comparison to PB and BM.

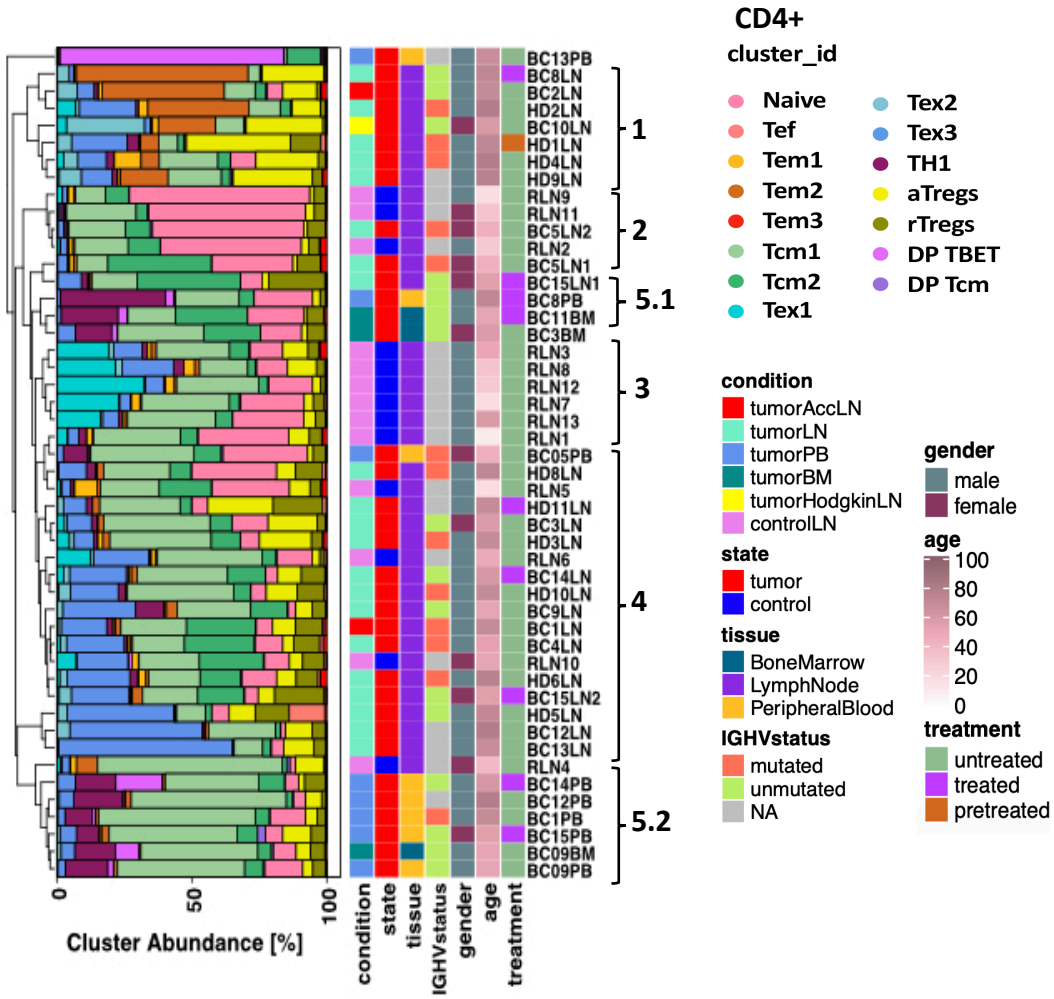


Figure 3.8: Hierarchical clustering of CD4+ subpopulation abundances across samples. The 5 groups marked with characteristic enrichment of different subpopulations are explained in the text. Cluster abundances are associated with condition, state, tissue, IGHV status, gender, age and treatment.

Extended annotation for cluster ids: Tem1=Tem1 Ki67+ CD38+, Tem2=Tem2 CD39+, Tem3=Tem3 ICOShi (high), Tcm2=Tcm2 CD39+, Tex1=Tex1 PD1hi (high), Tex2=Tex2 CD39+ Ki67+, Tex3=Tex3 EOMES+ PD1,TH1=TH1 KLRG+ TBET+, DP TBET=DP Tem TBET+.

It was observed that while tumor LNs from different patients had varied abundances of cell types (Tcm, Tregs, Tex3, Tem39), they were also different from T cell microenvironments of bone marrow and peripheral blood from the same CLL patient (expressing Tcm1 and TH1 KLRG1+ TBET+). This indicated inter as well as intra patient heterogeneity with respect to CLL TME.

Moreover, tumor BM and tumor PB cluster together, which might be due to contamination of bone marrow samples with peripheral blood during procurement. CD4+ clustering based on subpopulation abundances showed mixed patterns of corresponding IGHV status,

treatment and gender information. It was however noticeable that probands providing control LN samples had an age-range of 20-60 years, while age of the tumor patients was between 40-90 years.

The next section compares subpopulation proportions across samples after statistically correcting for differences in their library sizes, and tests whether the differences are influenced by age, IGHV status and treatment.

3.10 Impact of variation in patient's age on subpopulation abundances between tumor and control lymph nodes

As observed, the age of probands providing control LN samples was in the range of 20-60 years, while age of the tumor patients was between 40-90 years. It was therefore decided to evaluate whether age related differences were confounding phenotypic differences between the samples. Therefore, subpopulation abundances across samples were correlated with provider's age. 15 CD4+ subpopulations were hence evaluated for increasing/ decreasing/ constant abundances with age. Also, for a fair comparison a minimum and maximum age range cut off that included tumor patients and control persons within a similar age range were chosen. This criterion included 14 samples (5 controls LNs and 9 tumor LNs) within the age range of 40-60 years. Subpopulation proportion correlating significantly with age was that of DP Tcm ($R = -0.79$, p-value = 8.5e-4) showing decreasing proportion with age. All other subpopulation differences between tumor and control LNs could hence be attributed to effects of CLL in the TME of these samples. However, to robustly accept this claim, more samples should be included for the same analysis.

3.11 Differentially expressed CD4+ subpopulations in tumor v/s control lymph nodes

A limma model for cell counts from lymph node samples (tumor, n=23; control, n=13) was constructed to model subpopulation proportions dependent on tumor status, treatment and gender. This model normalized the variance introduced by unequal library size across all samples. Figure 3.9 shows z-score scaled proportions across samples in the form of a heatmap. Vertical bars alongside the heatmap show whether a subpopulation abundance was significantly different between tumor and control samples (green = significant, grey = not significant). The first bar is tumor LNs v/s control LNs (model 1 = without covariates). The second and third bars include significance values after categorical covariates treatment and gender respectively are added to model 1 separately. Multiple testing to identify differentially abundant subpopulations was performed using eBayes method and was corrected for significance by p-value adjustment by Benjamini and Hochberg method.

Proportions have also been shown for peripheral blood, bone marrow and Hodgkin malignancy lymph node samples, but they were not included in the limma model.

From the heatmap in figure 3.9 a, naïve subpopulation was significantly decreased (-1.76-fold on $\log_2$ scale, adj. p-value = 2.4e-2) in the CLL group. Significantly reduced (-2.86-fold on $\log_2$ scale, adj. p-value = 8.0e-04) Tex1 PD1hi (high) in tumor v/s control LN conveyed presence of exhausted cells in the control samples. This exhaustion could be attributed to acute and chronic infections other than CLL in the control group. Importantly, Tem1 Ki67+ CD38+ subpopulation was significantly decreased (-1.6-fold on $\log_2$ scale, adj. p-value = 1e-2) in tumor v/s control LNs. Strikingly, when treatment information was added to the model, the decrease attributed to CLL in naïve and Tem1 Ki67+ CD38+ subpopulations, was rescued. Convincingly, these subpopulations can be related to CLL pathogenesis in the present data. The covariate gender, however, had no impact. Box plots below the heatmap (figure 3.9 b) show the proportions of significantly different CD4+ T cell subpopulations in all tissues. The p-values are for differences between tumorLN v/s controlLN only.

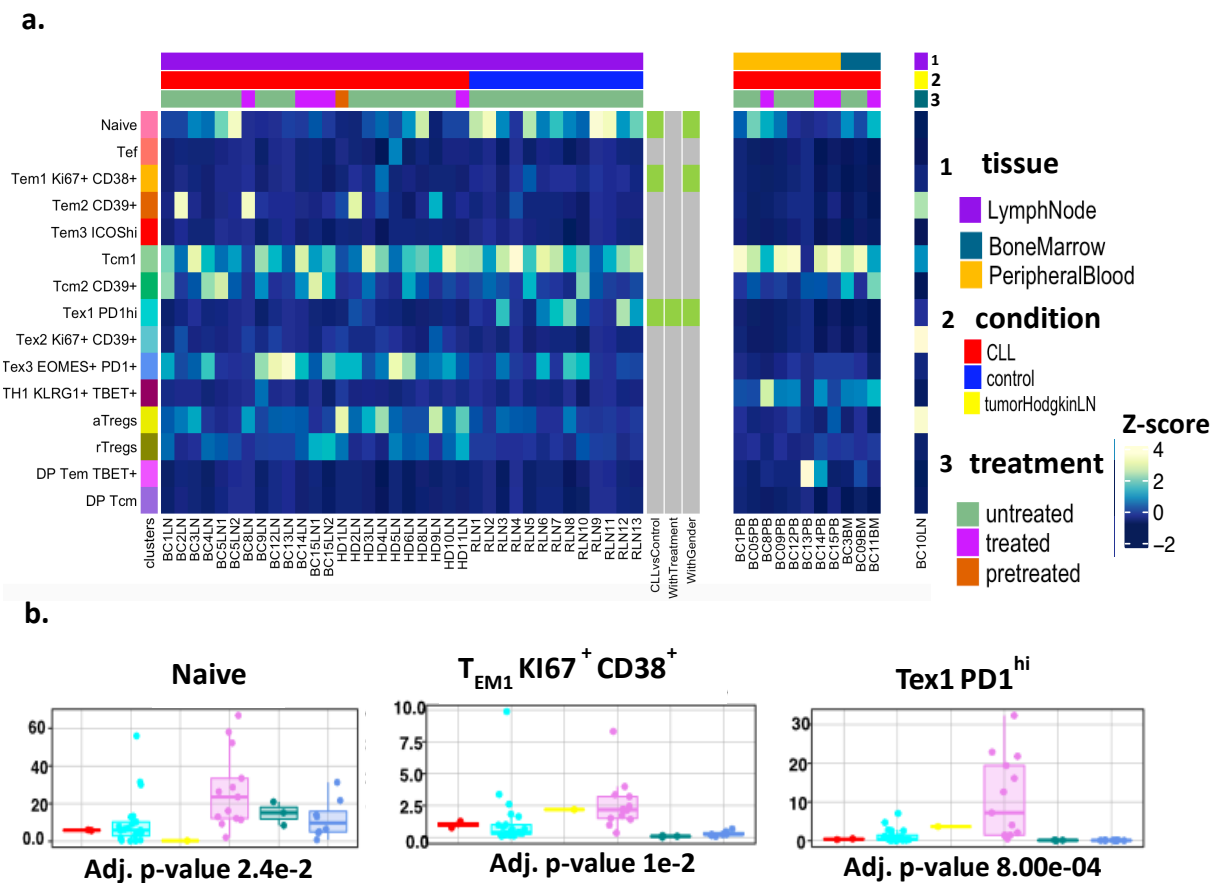


Figure 3.9: (a) Comparison of differentially abundant CD4+ T cell subpopulations in tumor LN v/s control LN setting using limma modelling. (b) Bar plots for significantly different populations are represented below the supervised heatmap (missing scale and legend). Adjusted (Adj. by Benjamini and Hochberg method) p-values mentioned below the box plot are significant values for differences between tumorLN and controlLN only.

ScRNA sequencing of the T cell compartment from CLL TME

After having investigated the single cell proteomic differences between T cell subpopulations, additional information about RNA profiles of T cell compartments on the single cell level was obtained. Single cell RNA sequencing (scRNA-seq) of CD3+ T cells (including CD4+ and CD8+ sub types) from CLL TME of lymph nodes of 3 CLL patients and spleen of 3 Eμ-TCL1 mice was performed to understand the interplay between heterogenous T cell populations and CLL progression at transcript level[3]. However, the scope of this thesis is limited to assessment of CD3+ CD4+ T cell subset.

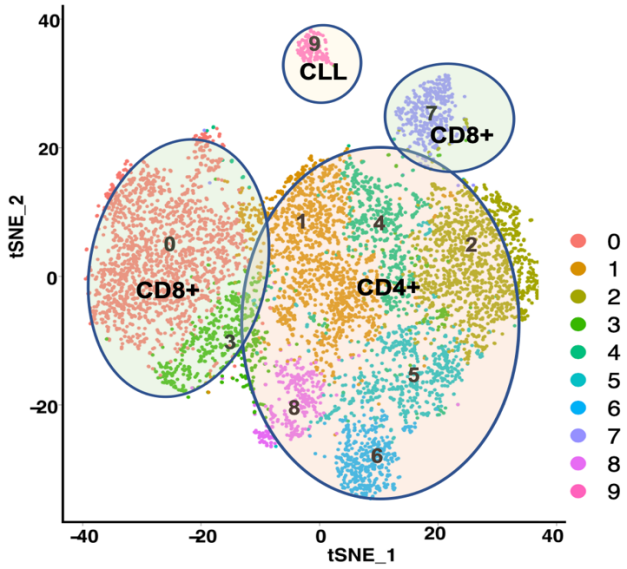
Importantly, single cell expression of T cell subsets was supported by paired T cell receptor (TCR) clonotype information. Defining the TCR repertoire from the TME provided an insight into diversity of T cell clones surrounding, reacting and expanding in response to CLL and/or additional infections in the patients

In the following sections I detail the transcriptional characterization of T cell subtypes in the CLL microenvironment from patient lymph node and compare it with that from the spleen of Eμ-TCL1 mice. I also characterize the TCR repertoire in patient samples and compare CLL T cell transcriptional profiles to that of T cells from breast cancer.

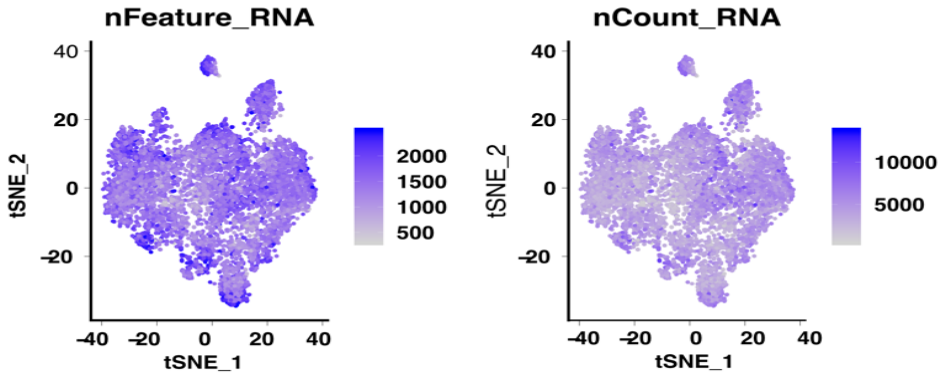
Observations from these results point towards similar and unique microenvironment subpopulations in CLL patients and the mouse model, consisting of similar and unique subpopulations as identified by proteomic (CyTOF) and transcriptomic profiles, and CLL lymph node specific subpopulations, if any, as compared to breast cancer lymph nodes (the only publicly available lymph node samples that used droplet-based library preparation, similar to in-house CLL data at the time of this analysis). The results also describe subpopulation of potential therapeutic value, that could be targeted to reinstate immune response against CLL.

[3] ScRNA processing of samples and library preparation were performed by Laura Llao Cid using facilities at Single Cell Open lab of DKFZ (Jan-Philipp Mallm, Katharina Bauer, Michelle Liberio, Karsten Rippe). Annotation of T cell subsets was performed by Laura Llao Cid, after I performed the clustering.

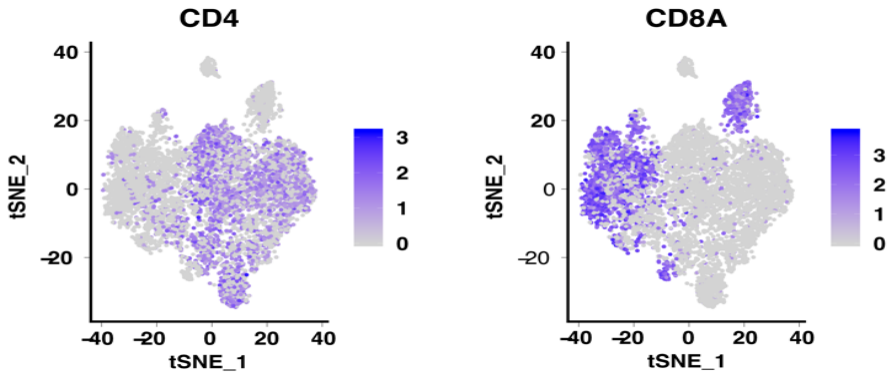
a.


CLL
CD8+
CD8+
CD4+
tSNE_1
tSNE_2
0
1
2
3
4
5
6
7
8
9

b.


nFeature_RNA
nCount_RNA
tSNE_1
tSNE_2
2000
1500
1000
500
10000
5000

c.


CD4
CD8A
tSNE_1
tSNE_2
3
2
1
0

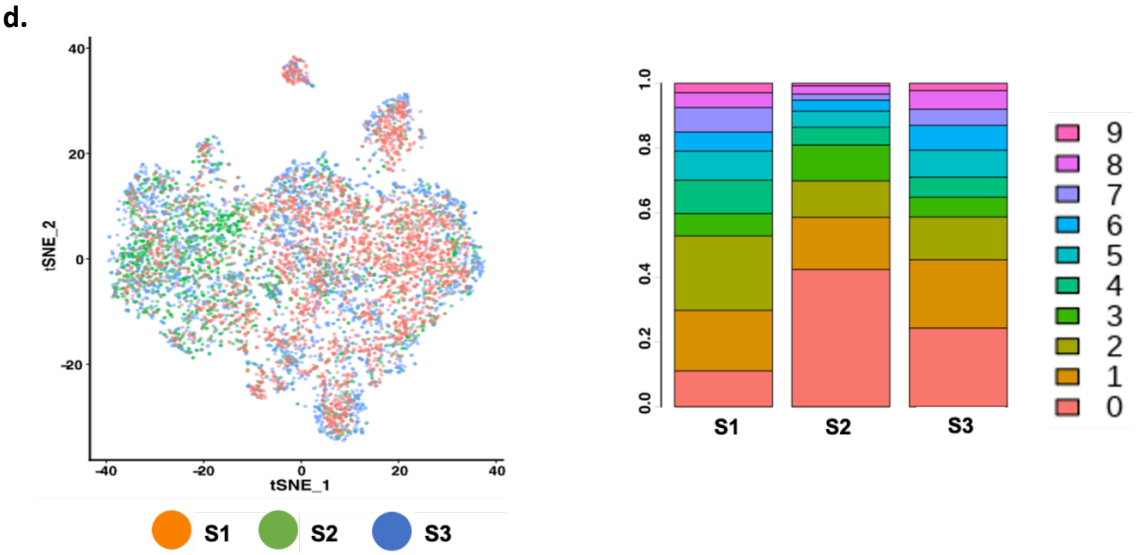


Figure 3.10: ScRNA sequencing of CD3+ T cells from 3 CLL patient lymph nodes.(a) t-SNE clustering to visualize 9 T cell and one CLL subpopulation. (b) Uniform distribution of number of genes (left plot) and number of transcripts (right plot) across all subpopulations. (c) Identification of CD4+ (left plot) and CD8+ (right plot) cell types by overlaying expression of CD4 and CD8A genes. (d) Distribution of cells from three samples (S1, S2 and S3) across all 10 populations on the t-SNE (left plot) and quantification of the same as stacked bar plots (right).

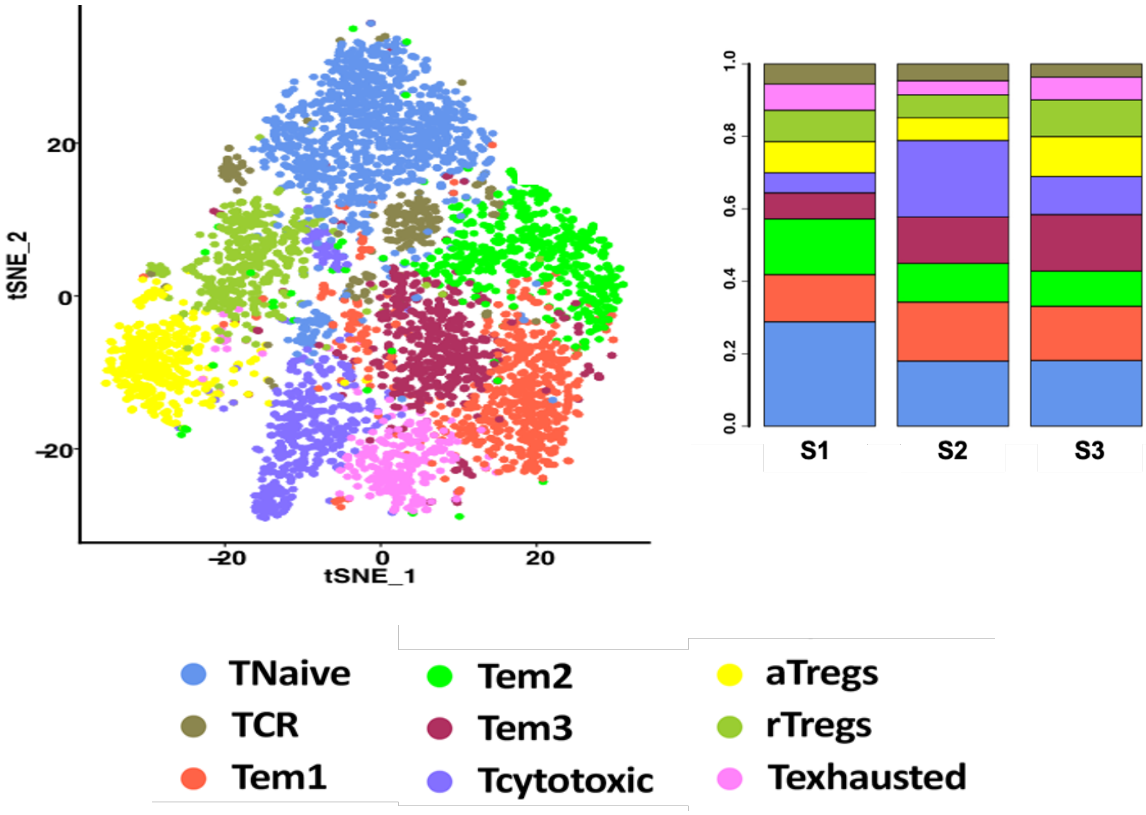
3.12 CD3+ T cell subpopulations identified using scRNA sequencing

Paired scRNA and targeted TCR sequencing for CD3+ T cells from three CLL patient lymph nodes was performed (experiments by Laura Llao Cid). Next, by graph based cluster analysis of CD3+ T cells followed by Louvain modularity optimization as a part of Seurat workflow, I observed 10 distinct subpopulations (Figure 3.10 a) (details of thresholds and data preprocessing steps in methods section). Figure 3.10 b shows uniform number of genes (nFeatures (left plot)) and transcripts (nCounts (right plot)) expressed in all cells across the clustering. Broadly, clusters 1, 2, 4, 5, 6 and partially cluster 8 were high in expression of *CD4* marker (figure 3.10 c (left plot)). Clusters 0, 3, 7 and partially cluster 8 showed high expression of *CD8A* marker (figure 3.10 (right plot)). Cluster 9 showed enrichment of CD19+ cells. These CLL B cells were spiked in to monitor uniformity in sequencing depth, and genes identified across all cell types and to make sure that the clustering is not dominated by technical differences.

Also, sample wise contribution to the proportion of each cluster (figure 3.10 d t-SNE (left) and bar plot (right)) showed proportional representation of all samples in all clusters except in cluster 0 (CD8+ subtype), where cells from sample 2 (S2) were enriched.

For further characterizing CD4+ and CD8+ T cell subsets, integrated CD3+ t-SNE was subset into two separate *CD4* and *CD8A* cell types using respective cluster classification described above. Only the CD4+ T cell subset will be discussed in this thesis.

a.



b.

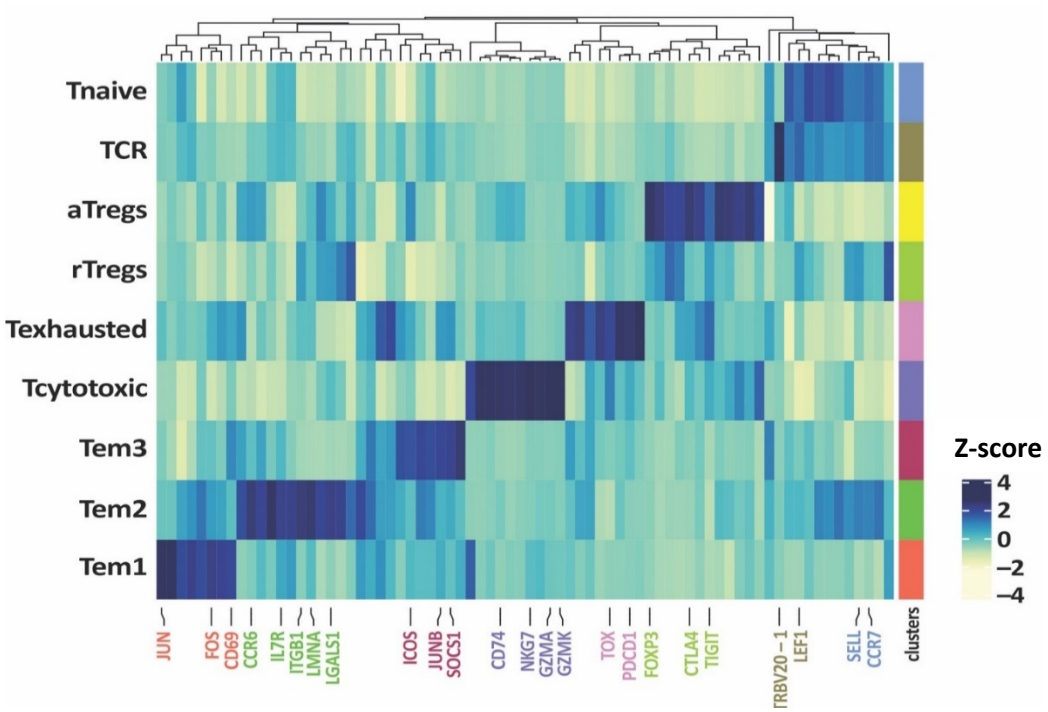


Figure 3.11: (a) 9 identified CD4+ T cell subpopulations (left), and their abundance in the three samples S1, S2 and S3 (right). (b) Subpopulations can be identified by the unique set of top upregulated genes they express, some of which are marked at the bottom of the heatmap (distance: Euclidean, linkage: complete).

3.13 Nine CD4+ T cell subpopulations were identified using characteristic expression of top marker genes

Graph based clustering and Louvain modularity optimization, as a part of Seurat workflow of CD4+ T cell subset identified 9 subpopulations (figure 3.11 a (left)), characterised by expression of top upregulated genes shown in figure 3.11 b. The heatmap depicts average expression of the top 10 highly upregulated genes in each subpopulation across all the cells in the subpopulation. Genes are clustered by column (distance=Euclidean, linkage=complete), and the expression values z-score normalized. The 9 subpopulations as in the heatmap are described below:

CD4+ naïve cell populations

Naïve and TCR phenotypes were enriched in *CCR7*, *LEF1* and *SELL*. TCR cluster in addition to the highly expressed naïve cell markers, characteristically expressed the T cell receptor gene

TRBV20-1. Also, the marker gene list of these clusters showed high expression of ribosomal protein genes encoding ribosomal proteins in the small subunit (Rps_) and large subunit (Rpl_) of ribosomal translation machinery. It has been previously shown that ribosomal subunits are downregulated in CD8+ exhausted T cells as compared to naïve effector and memory T cells, attributed to supressed translation (Wherry, Ha et al. 2007). Similar differences in expression of ribosomal protein genes between CD4+ effector and naïve subsets could be speculated from the present observation.

CD4+ regulatory T cells

Subpopulations annotated as aTregs (activated regulatory T cells) and rTregs (resting regulatory T cells) expressed *FOXP3*, *CTLA4*, *IL2RA*. These markers were expressed at least three times more in aTregs as compared to rTregs, for example expression of *FOXP3* in the former is 1.49-fold on $\log_2$ scale, as compared to the latter, where it is expressed at 0.4-fold on $\log_2$ scale. aTregs and rTregs were also previously described as phenotypically and functionally distinct subsets of FOXP3+ CD4+ T cells both in humans and in mouse (Xin Chen 2011).

CD4+ exhausted T cells

Coinhibitory molecules like *PDCD1*, *LAG3*, *TIGIT* and pro-inflammatory transcription factor *MAF* that drive T cell exhaustion and inhibit anticancer effector T cells were expressed in subpopulation annotated as T-exhausted (Verdeil 2016).

CD4+ cytotoxic T cells

The next subpopulation was highly upregulated in *GZMK* (2.3-fold on $\log_2$ scale) and *GZMA* (1.89-fold on $\log_2$ scale); and was downregulated in *LEF1* (naïve T cell marker) and *FOS*. This cluster was hence annotated as cytotoxic CD4+ T cell subset. Such cells have previously been observed to have roles in antiviral immune responses (Takeuchi and Saito 2017, Hashimoto, Kouno et al. 2019).

CD4+ effector-memory T cells

Last three subpopulations shown in the heatmap were those of antigen experienced effector memory CD4+ T cells. They expressed *JUN*, *FOS* and *CD69* activation marks (in Tem1 and Tem2). Tem3 showed upregulation of *ICOS*, which is also a costimulatory molecule expressed by activated CD4+ T cells (Mahajan, Cervera et al. 2007).

CD4+ T cells in the CLL TME, were found to exhibit naïve, regulatory, cytotoxic, effector-memory and exhausted phenotypes. These subpopulations were identified in all three samples (figure 3.11 a (right)).

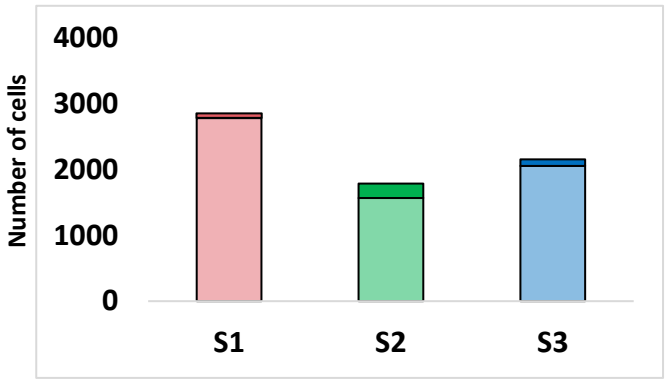
Adding T cell receptor information to singe cell transcription profiles of T cells

TCR information is used to track clonality and developmental diversity of T Cells (Marco De Simone et al 2018, Han et al 2014). In the premise of the present study, identification of an expanded and inactive TCR could be correlated to a potential response against a tumor neoantigen. The ultimate aim would be to develop a treatment approach by activating the T cells that are identified to expand against the tumor. Therefore, to identify T cell populations recognizing potential tumor antigens and expanding to external stimuli, paired single cell V(D)J sequencing was employed along with single cell RNA sequencing for all cells of the three patient lymph node samples (S1, S2, and S3) under study. 10X chromium protocol dictates that the V(D)J sequencing be done by 5' chemistry of single cell library preparation[4]. This is because, the V(D)J genes are closer to the 5' end of the TCR mRNA. The following section details identified TCR clonotypes, some of which are biologically relevant, for the combined human CD3+ T cell dataset (including both CD4+ and CD8+ cell types) unless otherwise specified. After the raw sequencing data is pre-processed by Cellranger (cellranger vdj) pipeline from 10X genomics (details in methods section), the following information can be retrieved for each cell from the output files produced:

1. The combination of V(D)J genes characteristic of each clonotype
2. Frequency of occurrence of each clonotype
3. Cell barcode associated with each clonotype
4. CDR3 amino acid and nucleotide sequence for the alpha and beta variable chain associated with each clonotype

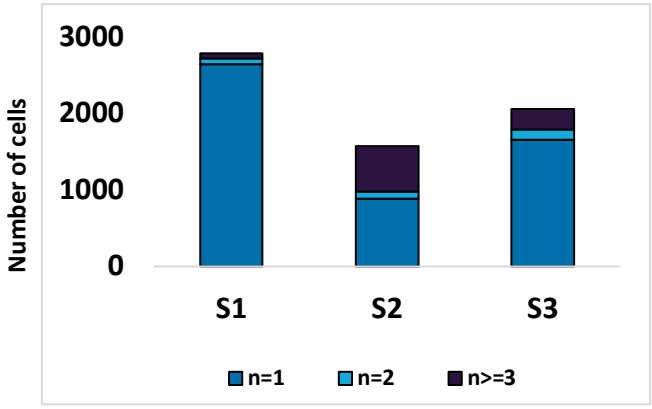
[4] TCR library preparation was performed by Laura Llao Cid at DKFZ Single Cell Open Lab (Jan-Philipp Mallm, Katharina Bauer, Michelle Liberio, Karsten Rippe). Annotating identified TCR clonotypes by VDJdb was performed by Laura Llao Cid and represented here by myself.

a.



	Cells annotated with clonotypes	Total
S1	2789	2865
S2	1572	1792
S3	2061	2161

b.



	n=1	n=2	n>=3
S1	2646	75	68
S2	889	91	592
S3	1658	131	272

c.

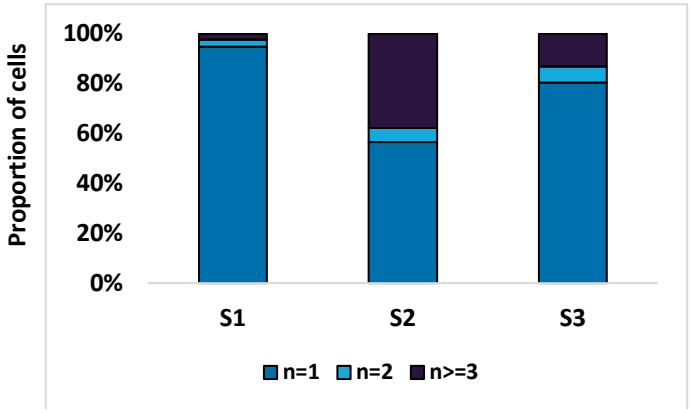
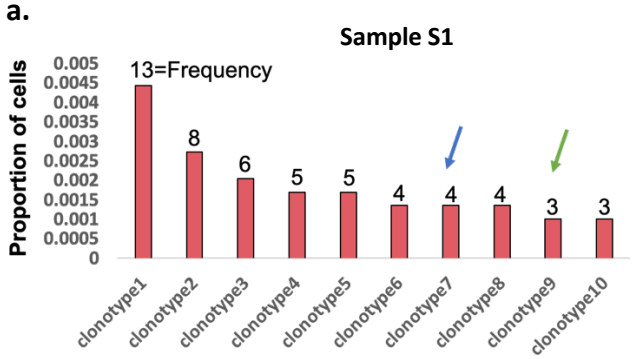


Figure 3.12: (a) Total number of cells (darker shade) and total number of cells annotated with clonotype information (lighter shade in the three samples S1, S2, S3). (b) Clonotype rearrangements mapping to only one cell (n=1), to two cells (n=2) and to three or more cells (n >=3), the same in proportions is depicted in (c).

3.14 Enrichment of clonally expanded CD3+ T cells revealed by TCR identification

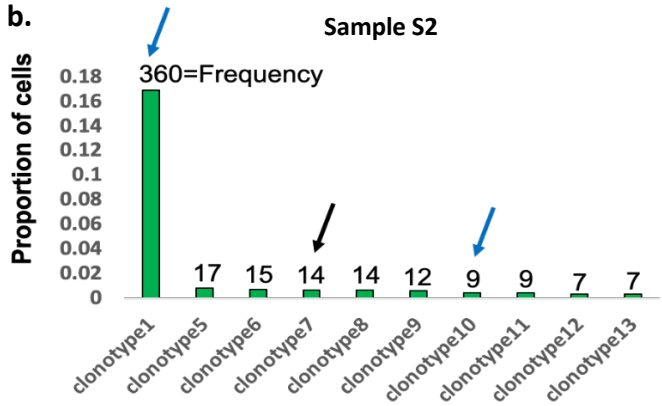
TCR information per cell was superimposed on the existing clustering of CD4+ and CD8+ T cells by common cell barcode information from both RNA-seq and V(D)J sequencing. Cells exhibiting the same V(D)J gene rearrangement were categorized under the same clonotype. While most cells presented at least one unique pair of α and β TCR chain alleles, a fraction of cells had non-unique allelic chain representations of single α, single β, two α and one β, or two β and one α. I defined a TCR as valid if it expressed one unique combination of paired α-β chains. In total, I detected TCR information with productive alpha and beta chains for 95% of CD4+ T cells and 90% of the CD8+ T cells (6422 out of 6820 CD3+ T cells from three samples). Figure 3.12 a, shows total number of cells (darker shade) and cells with clonotype information (lighter shade), separately for all three samples. Also, an expanded clonotype was defined as the one that in addition to having one unique combination of paired α-β chains had its chain combination shared at least between 3 cells suggesting a common cell of origin (Zheng, Zheng et al. 2017). According to these criteria I identified 106 CD4 + T cells and 826 CD8+ T cells with shared clonotypes in at least 3 cells and annotated them as expanded T cell clonotypes. Figure 3.12 b and c show the number and proportion respectively, of the cells with a unique clonotype (n=1), clonotypes shared between 2 cells (n=2), and clonotypes shared between three or more cells (n>=3, expanded clonotype). Sample 'S2' had the maximum proportion of expanded clonotypes. Following this, potential biologically interesting clonotypes were searched for from amongst all the identified clonotypes.

a.



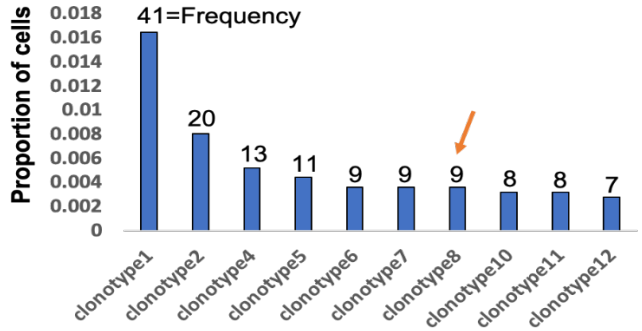
Clonotype	Frequency	CDR3_nucleotide sequence
clonotype1	13	TRA:CATPRTGANSKLTF;TRB:CSARGTGGGSDTQYF
clonotype2	8	TRA:CARIAAGNKLTF;TRB:CASSPDLLGAQHF
clonotype3	6	TRA:CAAHNTDKLIF;TRB:CATSDGVYEQYF
clonotype4	5	TRA:CAASDYKLSF;TRB:CSARDPWTSGRGYQETQYF
clonotype5	5	TRA:CAEMDSSYKLIF;TRB:CASSVGTRGGHEQFF
clonotype6	4	TRA:CAASTSNDMRF;TRB:CASSFDPSHEQYF
clonotype7	4	TRA:CIGRHLNNNAGNMLTF;TRB:CASSYSSGGYTEAFF
clonotype8	4	TRA:CAVNLYNNNDMRF;TRB:CASSLSTDTQYF
clonotype9	3	TRA:CAALRDDKIIF;TRB:CASSFGSYNEQFF
clonotype10	3	TRA:CAARTGANNLFF;TRB:CASRLMGNEKLFF

b.

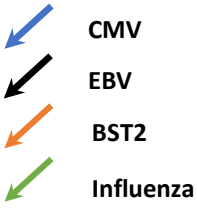


Clonotype	Frequency	CDR3_nucleotide sequence
clonotype1	360	TRA:CAASFSGTYKYIF;TRB:CASSPSPGGVEKLFF
clonotype5	17	TRA:CALSEVRASGTYKYIF;TRB:CASSPPGTSGNEQFF
clonotype6	15	TRA:CAEKDTGGFKTIF;TRB:CASSQGAGAGELFF
clonotype7	14	TRA:CADFSGGYNKLIF;TRB:CASSQSPGGTQYF
clonotype8	14	TRA:CALSEGDSGNTPLVF;TRB:CASSLTGGNQPQHF
clonotype9	12	TRA:CAVNLYSGGGADGLTF;TRB:CASSVVGSRYTF
clonotype10	9	TRA:CVVNNYGQNFVF;TRB:CASSLKGPLETQYF
clonotype11	9	TRA:CAGASNTGKLIF;TRB:CASSLAQGSGANVLTF
clonotype12	7	TRA:CAGGAIQGAQKLVF;TRB:CASSGTSGRADTQYF
clonotype13	7	TRA:CAVEPHRGSTLGRLYF;TRB:CASSSLGENIQYF

c.



Clonotype	Frequency	CDR3_nucleotide sequence
clonotype1	41	TRA:CAVSTGNQFYF;TRB:CASSQGPGQLHTF
clonotype2	20	TRA:CAVLLIGGSYIPTF;TRB:CASSPLGGNTEAFF
clonotype4	13	TRA:CAVSYQAGTALIF;TRB:CASEADTQYF
clonotype5	11	TRA:CAVKEGADGLTF;TRB:CASSPTGNPYEQYF
clonotype6	9	TRA:CAMSVARLMF;TRB:CASSPDKRGTEAFF
clonotype7	9	TRA:CATCLYGQNFVF;TRB:CASSPPGQVFNQPQHF
clonotype8	9	TRA:CAVSLSNFGNEKLTF;TRB:CASSYGDRGSNQPQHF
clonotype10	8	TRA:CAMSKGREYNFNKFYF;TRB:CASSIQGNEQFF
clonotype11	8	TRA:CAVRIYTGTASKLTF;TRB:CASRLGTGGHTEAFF
clonotype12	7	TRA:CAASPKGTYKYIF;TRB:CASSPAPDRYQQYF



Clinical information

S1: CMV+ in 2010

S2: CMV+, EBV status NA

S3: CMV+ EBV+

Figure 3.13: 10 most frequent clonotypes, their abundances and CDR3 amino acid sequence for (a) S1, (b) S2 and (c) S3. If a clonotype is mapped in VDJdb to a previously known antigen then it is accordingly marked with colored arrows, blue: cytomegalovirus (CMV), black: Epstein-barr virus (EBV), orange: BST2 (tetherin), green: influenza.

3.15 VDJdb identifies biologically interesting clonotypes

VDJ database (VDJdb) is a collection of T cell receptor (TCR) sequences whose antigen specificity is known (Shugay, Bagaev et al. 2018). The CDR3 (complementarity region 3) amino acid sequences of the clonotypes identified, were queried against this online database by Laura Llao Cid. CDR3 amino acid sequence is highly diverse and cells having the same CDR3 amino acid sequence share a clonotype. This results of this analysis identified biologically relevant VDJ rearrangements, whose antigen specificity has previously been documented. I represent these observations in figure 3.13, separately for the three samples (a) S1, (b) S2 and (c) S3; alongside a tabulated information about the clonotype nucleotide sequence for each case.

The frequency of the most abundant clone varies across the three samples. It was seen that S1 clonotype 7 and 9 resembled TCRs that recognize CMV (cytomegalovirus) and Influenza respectively (figure 3.13 a). In S2 the clonotype with highest frequency (S2 clonotype 1) was also known to recognize CMV (as per VDJdb, figure 3.13 b). This was in compliance with the clinical information of these two patients who previously tested positive for CMV. Interestingly, the CMV recognizing CDR3 sequence varied between the patients (S1 and S2) and also within one patient (S2). This might point towards multiple T cell epitopes recognizing the same antigen. S2 clonotype 7 showed the presence of EBV (Epstein-Barr virus) recognizing TCR (figure 3.13 b). However, there was no information of the EBV status of this patient to validate this. S3 clonotype 8 was identified to recognize BST2 (protein tetherin or CD317, figure 3.13 c) which is found to be over expressed in B cells in CLL (Gong 2015). Although this clonotype occurred in only 9 cells, it can still be hypothesized that these T cells were activated against CLL neoantigens. Further investigation including experimental verification is needed to prove this. Sample S3 was positive for CMV and EBV infections. However, TCRs detecting these infections were not identified at least in the top 10 most expanded clonotypes.

Next, CD4 and CD8 cell type specific clonotypes were visualized on their respective cell type clusterings to identify the subpopulations where they overlap. For the scope of this thesis only CD4 cell type specific TCRs are discussed. The above mentioned BST2 specific clonotype overlapped with CD8 cell type and is not shown in this thesis.

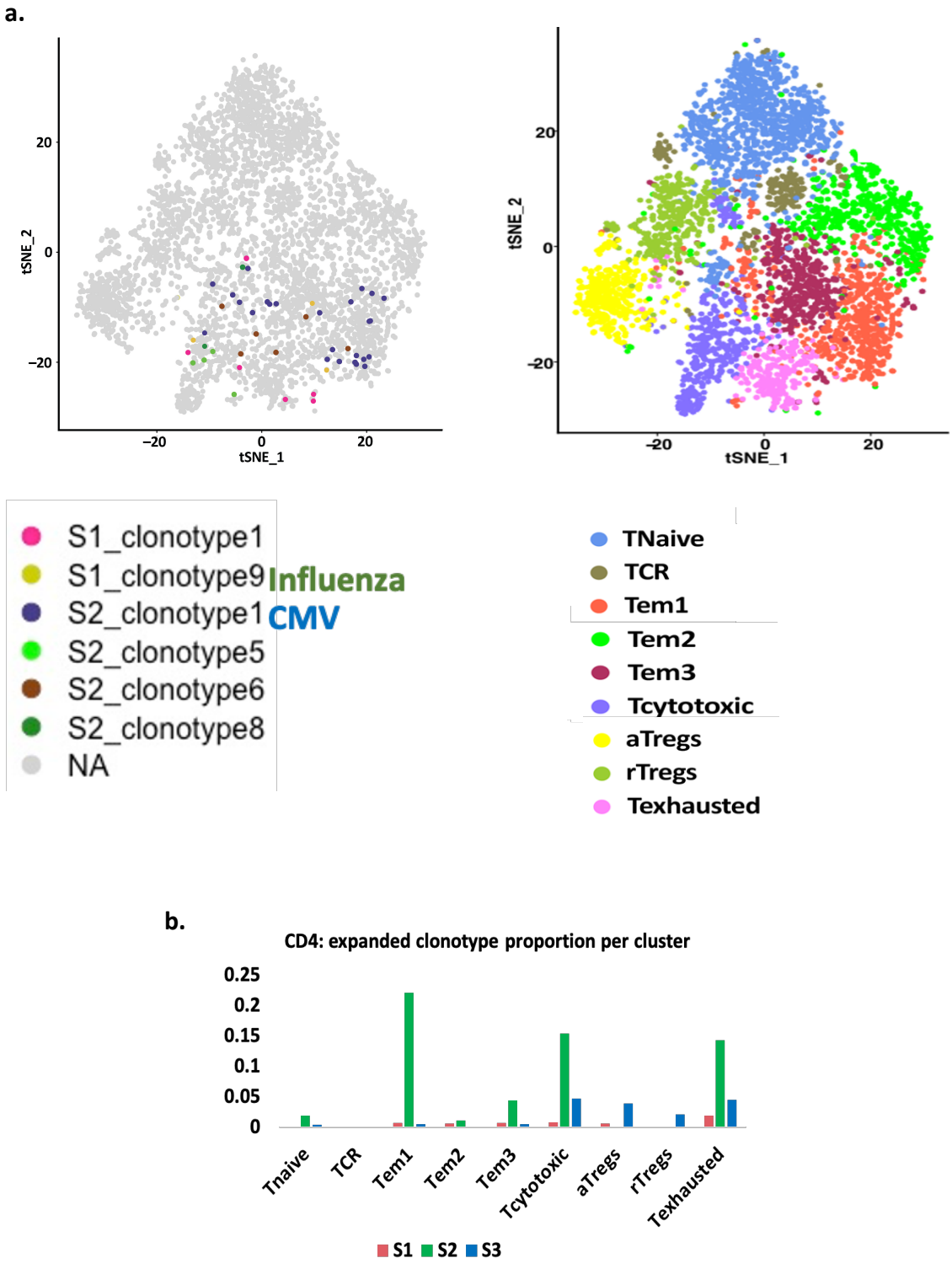


Figure 3.14: (a) t-SNE map of the CD4+ cell expression data highlighting top expanded clonotypes. (b) proportions of all expanded clonotypes per CD4+ subpopulation.

3.16 Exhausted and effector memory cell populations show highest proportion of expanded clonotypes

Clonotype information for the expanded clones (T cells) was mapped to the scRNA-seq profiles of individual cells of all three samples using the common cell barcode and visualized on the CD4+ t-SNE clustering previously obtained for scRNA sequencing. For purpose of clear visualization only the top 10 most abundantly expanded clonotypes from each sample, also discussed in figure 3.13, were displayed for CD4+ cell type as shown in figure 3.14 (a).

Figure 3.14 a for CD4 cell type shows S1_clonotype 1 to be concentrated in T-exhausted subpopulation, whereas the clonotype responding to CMV (S2_clonotype 1) got mapped to phenotypes Tem1 and Tem3. S1 clonotype 9 mapped to Tem1 and T-cytotoxic. Figure 3.14 b quantifies the proportion of expanded clonotypes (i.e. not only the top 10 but all clonotypes present in 3 or more cells) in all subpopulations of CD4+ cell type, sample wise. T-exhausted is the only subpopulation that consistently shows higher proportion of expanded clonotypes in all three samples. Other than this, Tem1 had higher proportion of expanded clonotypes from S2 and T-cytotoxic had major contribution from S2 and S3.

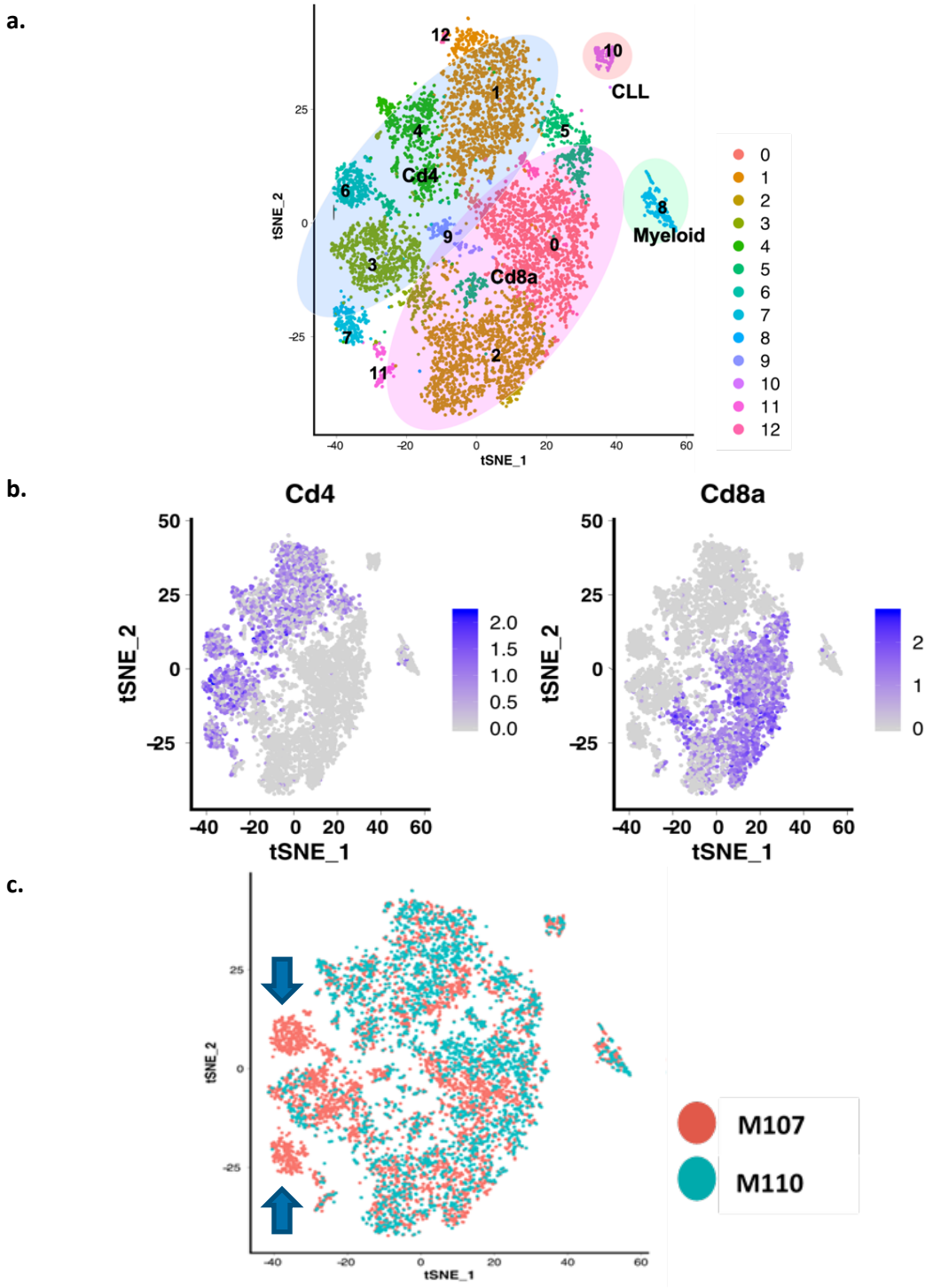


Figure 3.15: (a) t-SNE expression map of 12 Cd3+ T cell subpopulations identified from the spleen of Eμ-TCL1 mice. Major subtypes: Cd4, Cd8a have been marked inside a colored bubble. A myeloid cell population and CLL B cell population was also identified. (b) Clustering was dominated by expression of Cd4 and Cd8a markers. (c) 2 subpopulations markers with arrows were unique to mouse M107.

3.17 Cd3+ T cell subpopulations identified from spleens of 2 Eμ-TCL1 AT mice

Combined clustering of scRNA expression profiles of two Cd3+ T cell mouse samples was performed applying Seurat workflow and clustering techniques described in the methods section. A total of 13 clusters emerged including 4 Cd4+ clusters, 2 Cd8+ clusters, 1 cluster of myeloid cells, 1 cluster with Cd19+ B cells (used as a spike) (figure 3.15 a). 5 clusters had mixed Cd4+ and Cd8+ cell types. The clustering was expectedly dominated by differences in Cd4+ and Cd8+ cell types (figure 3.15 b). Two Cd4+ sub populations specific to sample 'M107' were observed (figure 3.15 c).

To identify subtype intrinsic subpopulations, Cd4+ and Cd8+ T cells were separated. Clusters, which had mixed populations of the two major cell types, were separated into their respective phenotypes based upon pairwise distances between counts (individual gene counts each from cells of c3,c5,c9,c11 and c12) and centroids (averages of same genes from rest of the clusters) of the most variable genes between CD4+ and Cd8+ specific clusters. The third mouse sample, for which there is information only about Cd8+ T cells, was added later to the Cd8 only clustering, after the two samples mentioned above have been sub-setted into Cd4 and Cd8 types. For the scope of this thesis only the Cd4+ cell type is discussed further.

3.18 Eμ-TCL1 mouse Cd4+ T cells manifest naïve, regulatory and exhausted T cell subpopulations similar to those identified in human CD4+ T cells

A total of 12 Cd4+ subpopulations were identified in T cells from spleens of two Eμ-TCL1 AT mice (figure 3.16 a). Figure 3.16 b represents the expression of these markers in each mouse subpopulation in detail in the form of violin plots (right column). Next to their expression in the mouse, subpopulations are also expressing the same markers as identified in CD4+ CLL patient clustering (left column). Comparing subpopulations in mouse and human in parallel gives an overview of similar and unique CD4+ T cell subpopulations in both species.

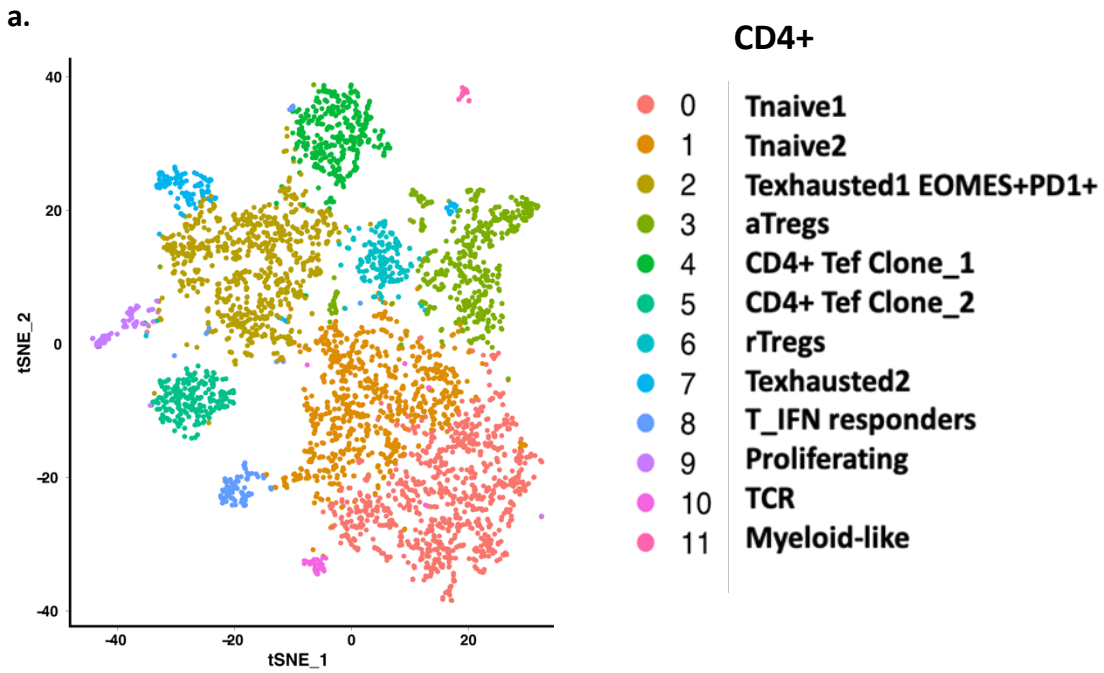
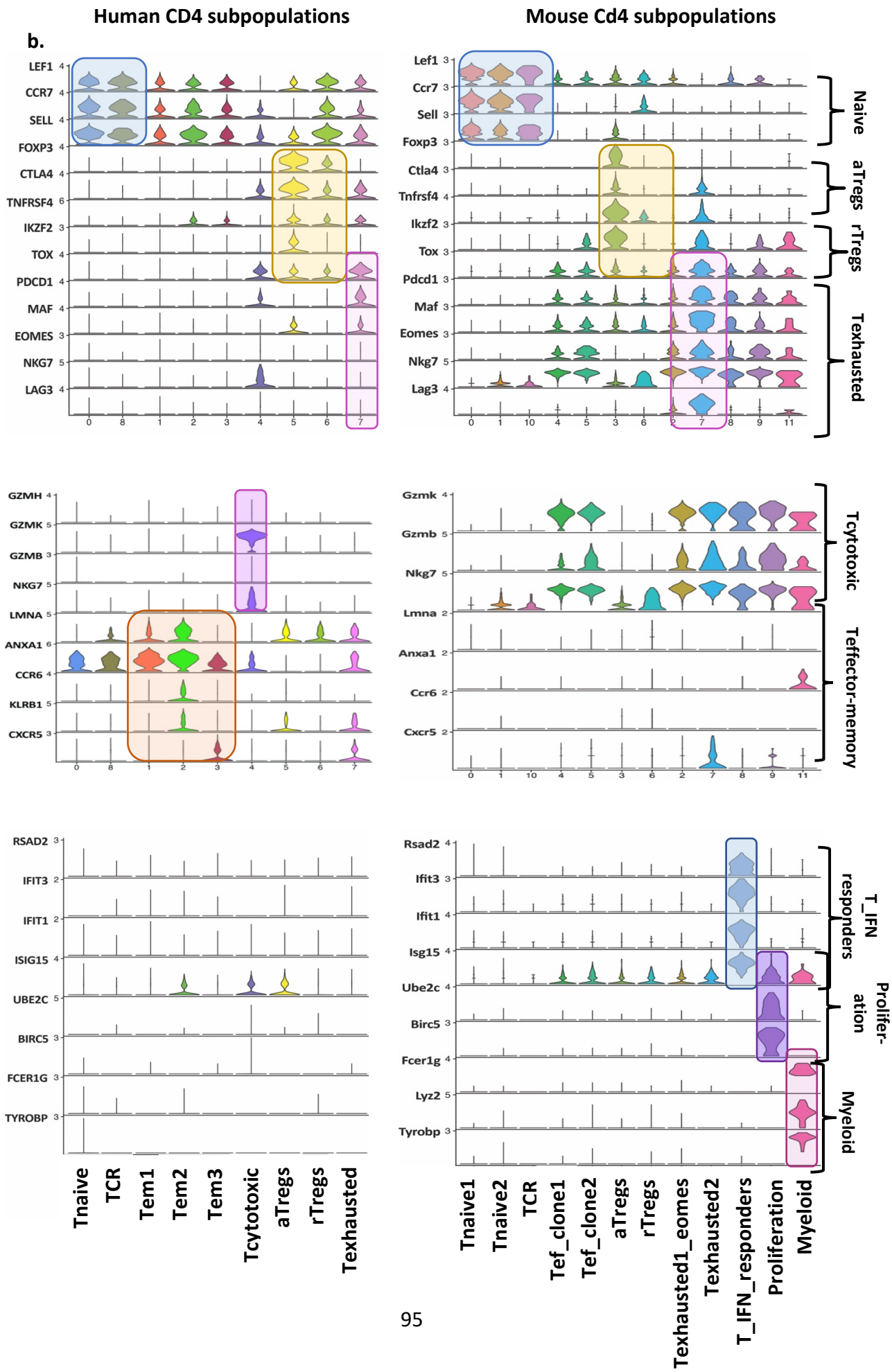


Figure 3.16: (a) 11 identified Cd4+ T cell subopulations from the spleens of Eμ-TCL1 mice. (b) Comparison of human (left column) and mouse (right column) Cd4+ T cell subpopulations.


Human CD4 subpopulations
Mouse Cd4 subpopulations
b.
LEF1
CCR7
SELL
FOXP3
CTLA4
TNFRSF4
IKZF2
TOX
PDCD1
MAF
EOMES
NKG7
LAG3
Lef1
Ccr7
Sell
Foxp3
Ctla4
Tnfrsf4
Ikzf2
Tox
Pdcd1
Maf
Eomes
Nkg7
Lag3
Naive
aTregs
rTregs
Texhausted
GZMH
GZMK
GZMB
NKG7
LMNA
ANXA1
CCR6
KLRB1
CXCR5
Gzmk
Gzmb
Nkg7
Lmna
Anxa1
Ccr6
Cxcr5
Cytotoxic
Teffector-memory
RSAD2
IFIT3
IFIT1
ISIG15
UBE2C
BIRC5
FCER1G
TYROBP
Rsad2
Ifit3
Ifit1
Isg15
Ube2c
Birc5
Fcer1g
Lyz2
Tyrobp
T_IFN responders
Prolifer-ation
Myeloid
Tnaive
TCR
Tem1
Tem2
Tem3
Tcytotoxic
aTregs
rTregs
Texhausted
Tnaive1
Tnaive2
TCR
Tef_clone1
Tef_clone2
aTregs
rTregs
Texhausted1_eomes
Texhausted2
T_IFN_responders
Proliferation
Myeloid

Naïve, regulatory and exhausted CD4+ T cells are identified by the same markers in Eμ-TCL1 mouse and CLL patient lymph nodes

As seen in figure 3.16 b (right column), Cd4+ mouse clustering showed three subpopulations enriched in naïve cell markers *Sell*, *Lef1* and *Ccr7* (T-naive1, T-naive2, TCR). This was similar to patient samples (left column), where two naïve cell specific subpopulations were identified as well (T-naive, TCR).

A similar trend was observed for regulatory T cells (Tregs); with two phenotypes (aTregs and rTregs) being identified in both mouse and human samples. *Foxp3*, *Ctla4*, *Ikzf2* and *Tnfrsf4* marker genes were upregulated in Tregs in both mouse and human samples.

Existence of aTregs and rTregs has previously been reported in aging studies of mice (Elyahu 2019). Tregs are specific to the CD4+ T cell type. Previous reports show increasing numbers of Tregs CLL patients and the Eμ-TCL1 mouse. They have been known to be high in progressive CLL (Giannopoulos K 2008; D'Arena G 2011; Biancotto A 2012).

Two subpopulations with increased expression of inhibitory receptors *Pdcd1*, *Lag3*; transcription factors *Tox*, *Maf* and *Eomes*; were identified in mouse clustering (T-exhausted1_eomes and T-exhausted2). One exhausted phenotype subpopulation was observed in patients (T-exhausted).

Human unique CD4+ specific T-cytotoxic and T-effector-memory subpopulations

T-cytotoxic represented one unique phenotype in the human clustering, however, in mouse T-cytotoxic human markers (*GZMK*, *GZMB*, *NKG7*) showed expression in all populations except naïve, Tregs and TCR groups. T-cytotoxic markers therefore can be said to represent antigen experienced activated populations in mouse.

Similarly, the three distinct T-effector-memory (Tem1, Tem2, Tem3) subpopulations in human clustering represented by *ANXA1*, *CCR6*, *KLRB1* and *CXCR5* were not expressed in the mouse subtype.

Mouse CD4+ specific IFN responders and proliferation subpopulations

Three subpopulations uniquely present in mouse TME were identified. The first subpopulation was enriched in *Rsad2*, *Ifit3*, *Ifit2* and *Isg15*. These genes are related to interferon gamma signalling and cytokine signalling in the immune system. Interferon-gamma-expressing CD4+ and CD8+ T cells are often in increasing numbers in CLL patient's PBMCs (peripheral blood mononuclear cells) as previously reported (Zaki 2000).

There were a few cells (number of cells = 84), that expressed *Ube2c*, *Birc5* and *Mki67*, associated with proliferation phenotype.

Another population identified specifically in mouse was that of myeloid cells. These cells express markers like *Fcer1g*, *Lyz2* and *Tyrobp*. We recently in our group elucidated that myeloid cells in the tumor microenvironment contribute to the pathogenicity of CLL in patients as well as disease progression in Eμ-TCL1 mouse model (S.Hanna 2019, Hanna, Yazdanparast et al. 2020). However, inclusion of myeloid cells in this data is a contamination from cell sorting, and only the T cell compartment is important for the scope of this thesis.
IFN responders, proliferating and myeloid cells were not identified in CLL patient lymph node samples in the present study.

Classification and comparison of human and mouse Cd4+ T cell subpopulations showed marked similarities in expression of the markers recognizing naïve, Tregs and T-exhausted phenotypes. However, there was no specific T-effector-memory (Tem) mouse population, as was seen in the human CLL samples. Also, several mouse subpopulations expressed T-cytotoxic markers *GZMB*, *NGK7*, upregulated specifically in only one human subpopulation. These might, however, indicate activated subpopulations in mice. Lastly, IFN responders were not observed in the human CD4+ subpopulation.

Comparison of CLL T cell compartment with that of Breast Cancer

After comparing scRNA-seq profiles of CLL patients and those from the Eμ-TCL1 mouse model and having discussed the similar and unique subpopulations present in both the species, I also investigated differences between transcriptional profiles of T cell subsets from CLL and breast cancer patients.

ScRNA-seq of breast cancer (BC) TME data was chosen because of their inclusion of patient matched lymph nodes (Azizi, Carr et al. 2018). Comparing similar tissue from different cancers (CLL and BC) helps remove tissue specific variation and identify CLL specific subpopulations. Also, libraries for BC data were prepared using similar droplet-based methods like the in-house CLL data. The data was available with GEO ids: GSE114727, GSE114725, and GSE114725. These analyses paved way to define CLL specific subpopulations, absent in breast cancer TME. For the comparison, I used both CD4+ and CD8+ CLL subpopulations to overlay on the breast cancer data.

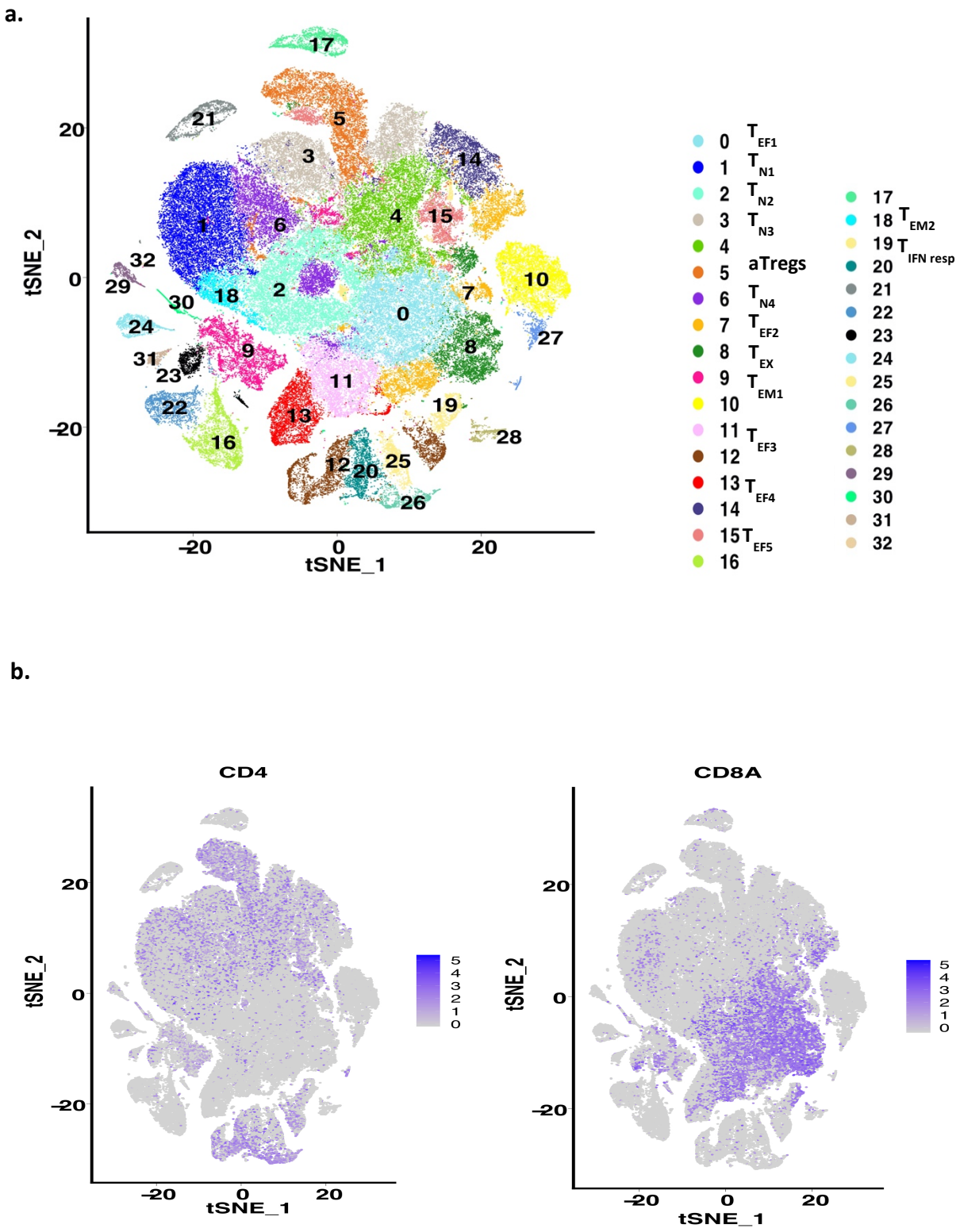


Figure 3.17: (a) Integrated clustering using canonical correlation analysis (CCA) for breast cancer (BC) and CLL cohort identified 33 clusters, the annotated T cell clusters (by Laura Llao Cid) are marked in the legend (b) CD4+ and CD8+ markers expression in the integrated clustering for broadly recognizing the phenotypes.

3.19 CLL specific subpopulations observed after integrating CLL lymph node data with publicly available breast cancer data

Reference based Canonical correlation analysis (CCA), an extended functionality of Seurat workflow was used to integrate the in-house scRNA seq LN CLL samples to breast cancer (BC) public dataset. The BC dataset, being the larger one, was used as a reference. The workflow transformed both the datasets using SCTransform function (detailed in methods) (Hafemeister and Satija 2019). After normalization, highly variable features were identified (default = 3000), in each dataset. Mutual nearest neighbours /anchors were then scored between BC and CLL dataset. Anchors were scored based upon the how many similar anchors it was surrounded by. High scoring anchors were hence limited to the same biological phenotype. CLL query dataset was then integrated upon the BC reference dataset with highest scoring anchors in two-dimensional space. After this step, rest of the Seurat workflow remained the same as applied previously for other clusterings in this thesis (detailed in methods section). Seurat based CCA was applied because of its proven better performance that other available methods (Stuart, Butler et al. 2019).

Since the breast cancer cohort and CLL cohort were both prepared with drop-let based library preparation methods, strong technical variation due to libraries was not expected.

After integration, 33 clusters were identified based upon the expression of their top regulated marker genes (figure 3.17 a). Five subpopulations of T-effector phenotype (T_{EF1-5}), four of T-naive (T_{N1-4}), two T-effector-memory (T_{EM1-2}), and one subpopulation each of T-exhausted (T_{EX}) and that of aTregs (activated regulatory T cells) was identified.

Functional annotation of the clusters based on marker gene expression was performed by Laura Llao Cid. The subpopulations not annotated were phenotypes other than T cells from the breast cancer TME. Integrated clustering was then looked at for the expression of the two major cell types of interest: CD4 and CD8A (figure 3.17 b left and right respectively). The two markers separated their subtype specific subpopulations.

The integrated t-SNE clustering was then faceted by tissues (breast/ lymph node/ peripheral blood) and the state (breast cancer/ CLL/ normal breast) to check the distribution of the same in 33 expression clusters, and is shown in figure 3.17 c. Cells from BC TME were spread in all the clusters. Subpopulations from naïve (clusters 1, 3, 6), effector (clusters 7, 15), Tex (cluster 8), Tem1 (cluster 9) and others (non-T cells) were absent from the microenvironment of matched normal breast. These were speculated as breast cancer specific subpopulations in the TME.

As compared to CLL lymph nodes (CLL_CD4 and CLL_CD8A) cells, breast lymph nodes (LymphNode) showed enrichment of naïve subpopulations (clusters 1, 3 and 6). Microenvironment of peripheral blood from breast cancer patients was enriched in naïve and effector phenotypes.

CLL_CD4 and CLL_CD8A cell types overlapped with the naïve (clusters 1, 6), Tregs (cluster 5), effector subsets (clusters 13, 11, 7, 0) and effector-memory subset (cluster 18). Most interestingly, CLL_CD4 and CLL_CD8A had in common cluster 9 (Tem1), which was seen to be uniquely present in CLL lymph nodes. Next, to identify the phenotype of CLL cells that make up cluster 9, I tracked CD4+ and CD8+ CLL cells in the integrated breast cancer and CLL clustering.

c.

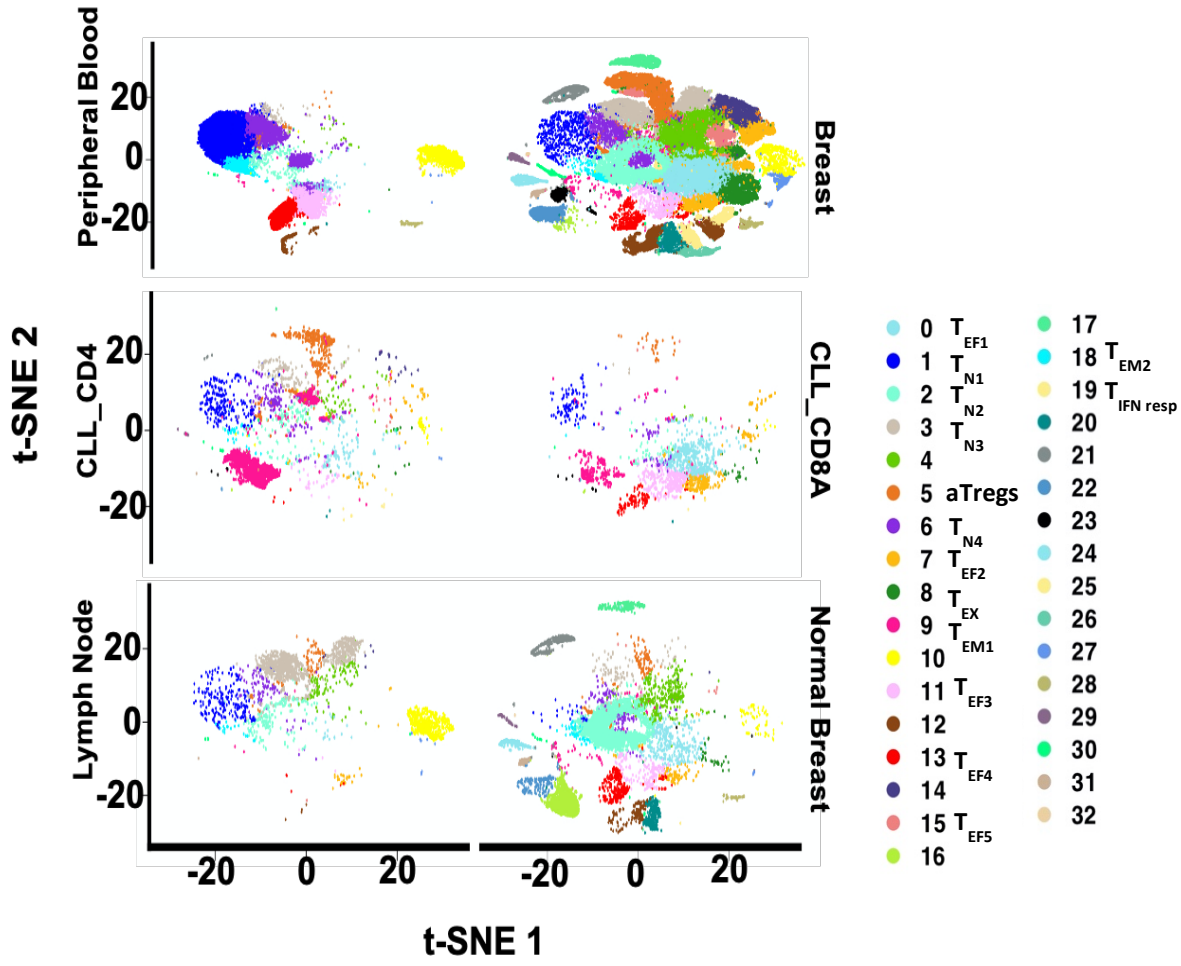


Figure 3.17: (c) 33 expression clusters faceted by tissue (lymph node/ peripheral blood/ breast) and state (breast cancer/ CLL/ normal breast). CLL specific cluster 9 is in pink identified in both CLL_CD4 and CLL_CD8A but absent from BC cohort.

T_{EF} = Teffector, T_N = Tnaive, aTregs = activated Tregs, T_{EX} = Texhausted, T_{EM} = Teffector-memory.

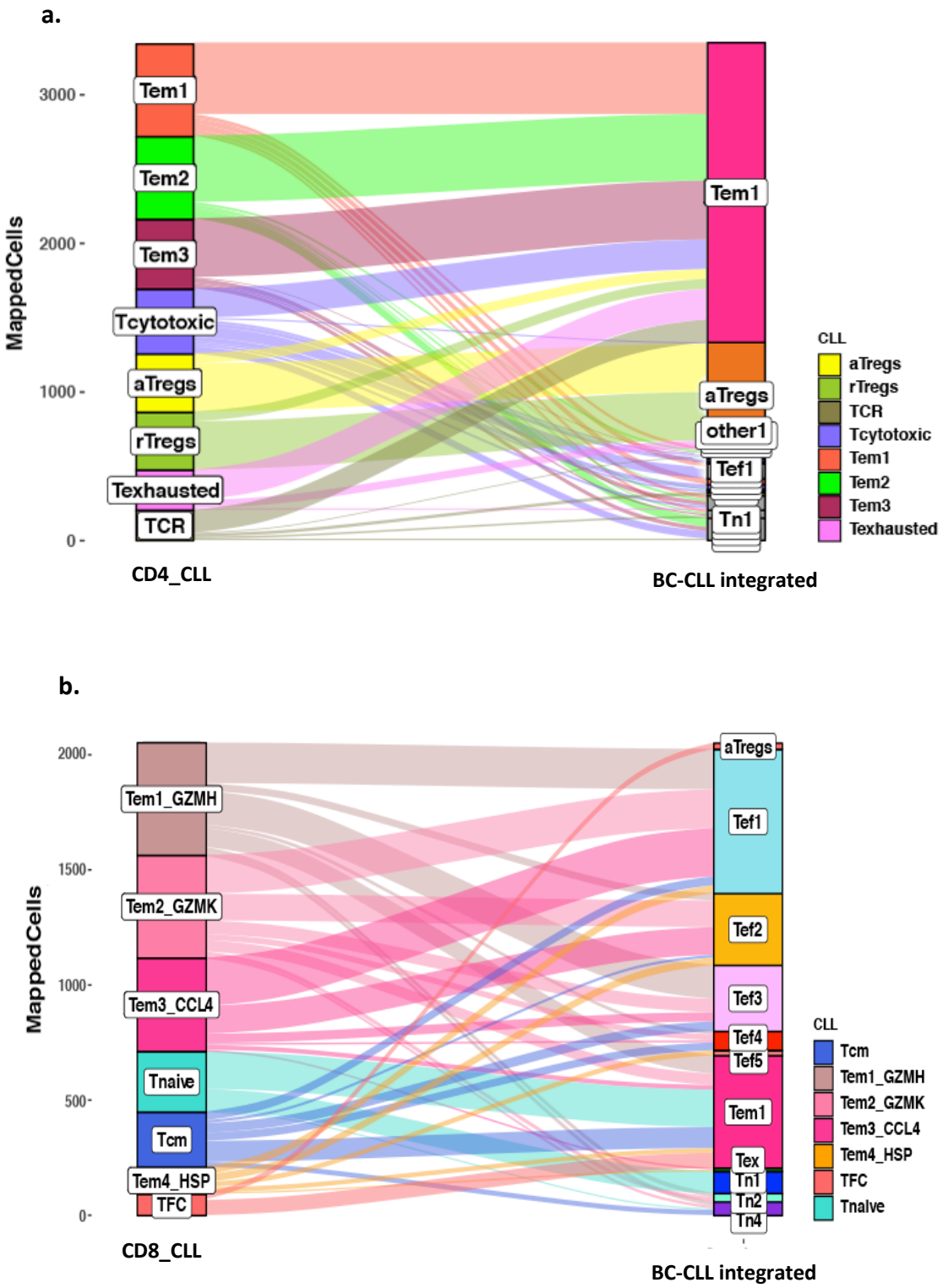


Figure 3.18: (a) Tracking of cells from CD4 CLL specific subpopulations into the identified BC-CLL (Breast Cancer- CLL) integrated subpopulations (b) Tracking of cells from CD8 CLL specific subpopulations into the identified BC-CLL integrated subpopulations. Tem=Teffector-memory, TCR=T cell receptor, TFC=T follicular cytotoxic, Tcm=Tcentral memory, Tregs=regulatory T cells, Tef=Teffector, Tn=Tnaive, Others: subpopulations other than T cells

3.20 CLL specific cluster 9 (Tem1) is composed of a mixture of cells from both CD4+ and CD8+ cell types

CLL specific cells were tracked for CD4+ and CD8+ CLL subpopulations annotated earlier for the CLL cohort. This is shown in figure 3.18 a (CD4_CLL) and b (CD8_CLL).

The following observations were made:

1. Naïve cells from both CLL CD4+ and CD8+ were mixed into naïve subpopulation in the BC-CLL clustering (Tn1/Tn2/Tn4).
2. Activated Tregs (aTregs) from CD4_CLL cell type were mapped to aTregs (cluster 5) from integrated clustering (figure 3.18 a).
3. Tem1 GZMH and Tem2 GZMK populations from CLL CD8+ cell type, mixed into T_{EF1}, T_{EF2}, T_{EF3}, T_{EF4} and T_{EF5} of the BC-CLL clustering (figure 3.18 b).
4. CLL CD8+ Tem4 HSP cells mapped to T_{EF1} and T_{EF2} from the BC-CLL subpopulations.
5. CLL CD8+ T-follicular cytotoxic (TFC) cells lie in cluster 5 (aTregs) and cluster 9 (CLL specific) of the integrated clustering.
6. Cells from T-effector-memory populations of both CLL CD4+ and CD8+ cell types mixed together in cluster 9 (Tem1) of the integrated clustering.

In general, similar phenotypic T cell subpopulations from both datasets fall together, with the exception of CLL specific cells in Tem1 (cluster 9) of the integrated breast cancer and CLL clustering. From figure 3.18 shows overlap of CD4_CLL and CD8_CLL subpopulations with Tem1 in the BC-CLL (Breast cancer-CLL) clustering. From here, it could be speculated that these CLL specific cells, that formed a separate subpopulation in the integrated clustering, were effector cells responding to specific CLL neoantigens. However, further analysis is needed to remark any further on this cluster.

4. Discussion

Even though more options than ever, with tremendous therapeutic efficacy like ibrutinib are available for chronic lymphocytic leukemia (CLL), the disease remains incurable. The success of CLL treatment, right now is limited to increased progression free survival and avoidance of a full-blown relapse.

Treatment outcome depends mainly upon patient age at diagnosis, existing comorbidities, IGHV mutation status as well as manifestation of prognostic mutations (del17p, del13q, *NOTCH1*, *SF3B1*) (Gaidano 2017). One of the factors that contributes to differential response against therapies across patients and complications in formulation of a curative medicament is tumor heterogeneity exhibited during CLL progression and pathogenesis. I have presented in my thesis clonal heterogeneity in the Eμ-TCL1 mouse model of CLL and heterogeneity with respect to CD4+ T cells in the TME of CLL patients and mouse tumors.

CLL heterogeneity as contributed by CNVs, B cell receptor rearrangements and somatic variations in Eμ-TCL1 mouse tumors

Investigation of copy number variations (CNVs) in Eμ-TCL1 primary and transplanted tumors identified gain of chromosome 15 in 6 out of 8 in-house samples, as well as all samples of the public dataset included in the analysis (7 out of total 11). This trisomy 15 included the "*Myc*" oncogene. Previously a similar observation has been reported in the context of malignant transformation in the TCL1-tg mouse model (Shen 2006). This is a similar mouse model that develops a disease similar to human T cell prolymphocytic leukemia, in contrast to Eμ-TCL1 that develops B cell leukemia. Experiments are presently being performed to validate effects of trisomy 15 and amplified *Myc* expression on CLL pathogenesis in Eμ-TCL1 mouse tumors. A positive correlation between *Myc* overexpression and CLL progression would make the genetic predispositions of this mouse model very different to the ones in CLL patients.

The other two factors investigated in this thesis and contributing to CLL clonal heterogeneity in Eμ-TCL1 mice included evolution of B cell receptor (BCR) clonality and constitutive as well as acquired somatic variations. Exploring evolutionary dynamics contributing to tumor heterogeneity in the mouse validated the suitability of the mouse model as a preclinical tool for CLL.

V(D)J gene rearrangements of *Ighv* genes, also called BCR clonotypes, were sequenced and analyzed in serially transplanted mouse tumors. BCR rearrangements in 10 out of 13 (77%) Eμ-TCL1 mouse tumors were oligoclonal. Mouse tumors have also previously been described to be oligoclonal (Lascano 2013). On the other hand, the disease is mostly monoclonal, and multiple IGHV rearrangements only occur inm5-24% of total CLL patient cases (Darwiche, Gubler et al. 2018). Restricted BCRs using the variable genes *IGHV3-21* and *IGHV1-69* are linked to poor prognosis in patients, and *IGHV4-34* and *IGHV2-30* are pronounced in patients with indolent course (Slupsky 2014). Even in the mouse model, overlapping BCRs between tumor samples were observed. *Ighv12-3* occurred in all three independent mouse tumors: A1.1, B1.1, and C1.1 (figure 3.1), either in subsequent transfers (A1.1, C1.1) or as a single clonotype in B1.1 and its transfers. Proportion of oligoclonal BCRs was dynamic over serial transplantations, i.e. as the disease progressed.

Identified stereotyped BCR rearrangements with variable genes *Ighv11-2*, *12-3* and *1-55*, were previously reported in mice to expand in response to chronic stimulation by autoantigens like PtC and others; causing persistent inflammation and playing a vital role in CLL pathogenesis (Hayakawa, Formica et al. 2016, Jimenez de Oya, De Giovanni et al. 2017). From these observations it can be suggested that in this mouse model, *Tcl1* overexpression acts only as a predisposing factor for pre-malignant transformation, and that an interplay by autoantigens is another likely factor that adds to leukemic progression. Further research on BCR evolution during adoptive transfer should be performed based on an independent cohort of 13 Eμ-TCL1 mouse primary tumors that were sequenced for their BCRs and were identified to be oligoclonal (not shown in this thesis). Three of these tumors had either no or late engraftments or grew very slowly (as observed by the mouse experimentalist). These tumors should be studied for clonal evolution by secondary transfers in the future, to strengthen the claims made in this thesis.

Another important contributor to CLL pathogenesis in Eμ-TCL1 mouse are genomic aberrations. Like in human CLL, non-overlapping low-allele frequency (< 5%) SNVs (Somatic Nucleotide Variations) were identified in primary and serially transplanted mouse tumors using WES (Guieze and Wu 2015). Cellular prevalence using variant allele frequencies (VAFs) was calculated over serial transplantations of tumors. This study was first of its kind that investigated tumor heterogeneity due to patterns of BCR dynamics in association with SNV-defined subclones in serially transplanted Eμ-TCL1 mouse tumors (n=7, 4 in-house, 3 publicly available). Three patterns of evolving SNV-defined and BCR clonotypes emerged as the disease aggressed from primary to secondary tumor. In the first pattern, both BCR clonotypes as well as somatic variants were displaced by novel ones after serial transplantation of the tumors. In the second pattern BCR clonotype remained constant but new SNV defined subclones emerged, and in the last one the same BCR clonotype and somatic variants perpetuated at primary and transplanted time points. This revealed a variation in CLL tumor evolution with a potent impact on CLL progression, pathogenesis and treatment response. The dynamic usage of BCR clonotypes as well as SNV- defined subclones could be indicative of varied selection pressures defining the strength of immune responses during the course of CLL (Darwiche, Gubler et al. 2018). From these observations the course of CLL in Eμ-TCL1 mouse seems as heterogenous as in patients. Moreover, the oligoclonality of BCRs in mice, reactivity with several autoantigens, and 'Myc' overexpression (as suggested in this thesis) could be attributed to development of aggressive CLL in the mouse within 5-6 months, which in humans occurs at a median age of 70 years.

Effects of ibrutinib treatment on the BCR dynamics, somatic mutation landscape and transcription profile of Eμ-TCL1 mouse tumors

Ibrutinib is used as a first line monotherapy to treat CLL patients. Often around 20% of patients relapse due to acquired resistance through mutations in the *BTK* or *PLCG2* genes. Therefore, a monoclonal BCR tumor was serially transplanted (by Haniyeh Yazdanparast) to investigate genetic and transcriptomic changes inflicted upon ibrutinib treatment and eventual resistance in Eμ-TCL1 mouse tumors as well. Tumors were subjected to WES- and RNA-sequencing at 6 weeks after start of treatment when they were showing first signs of treatment resistance and uncontrolled growth. No apparent genomic changes were observed

as a result of ibrutinib treatment. It was hypothesized that 6 weeks was a relatively short period for the genetic make-up of the tumors to change as a response to ibrutinib treatment. Also, the same monoclonal BCR was identified before and after treating the tumors.

Next, a limma model was constructed that removed effects of tumor proliferation and ibrutinib treatment effects, and instead only identified genes whose expression profiles were modulated by impact of potential ibrutinib resistance. Pronounced transcriptional changes were observed in ibrutinib resistant tumors as compared to vehicle treated ones. From the top upregulated genes identified to be putatively involved ibrutinib resistance, *Tbet* gene was chosen for mechanistic studies as a potential therapeutic target to counteract ibrutinib resistance (as a follow up study by Dr. Lavinia Arseni). *Tbet* has been implicated as a potential biomarker for pathogenic T cells (Ji, Sosa et al. 2011). It has also been known to enhance survival of B cells and play a role in B cell mediated immune responses (Barnett, Staupe et al. 2016). However, its functional role in clinical setting with respect to B cells remains unexplored.

Even though the tumors stayed monoclonal over the course of this treatment, to conclude that ibrutinib treatment induces no selection pressure on BCR dynamics is premature. Suggestively, the treatment study should be carried out on Eμ-TCL1 mouse tumors that are primarily oligoclonal. Only then a fair conclusion can be drawn about the impacts of ibrutinib treatment on BCR dynamics of Eμ-TCL1 mouse tumors.

Heterogeneity in CLL and the Eμ-TCL1 mouse attributed to tumor microenvironment (TME)

Chronic lymphocytic leukemia is known to have a pro-tumor microenvironment (Wiestner 2017). The aim here was to first gather insights into differences in CD3+ T cell subpopulation abundances across CLL patients, and relate that to heterogeneity in the CLL TME T cell compartment in different tissues. Secondly, differential proportions of T cell subpopulations in CLL in comparison to control samples were investigated for their potential contribution to CLL pathogenesis. These subpopulations could be followed as potent therapeutic targets to reinvigorate otherwise supressed immune responses in CLL.

Intensities of 32 T cell surface markers were measured using a CyTOF single cell procedure for T cells from the TME of 23 lymph node (LN) CLL samples, matched peripheral blood (PB) and bone marrow (BM) samples from 8 and 3 patients respectively. This study was the first of its kind to study the TME of CLL LN samples at the single cell level.

Interestingly, it was observed that the exhausted (Tex3) phenotype of CD4+ T cells was prominent in CLL lymph nodes as compared to PB and BM samples from the same patients. PB and BM samples on the other hand had enrichment of central-memory (Tcm) and TH1 KLRB1 TBET+ cells (figure 3.8). Two previous gene expression microarray studies profiled B cells from matched LN, BM and PB samples from CLL patients. LN-resident malignant B cells manifested activated BCR, NF-κB signalling and showed an increased proliferation rate (Mittal, Chaturvedi et al. 2014). Furthermore, the fraction of newly divided B cells was found to be highest in LNs as compared to PB or BM (Wiestner 2017). My finding of enriched exhaustion in LNs is in line with observations from the microarray studies that are indicative of a pronounced disease in LN as compared to BM and PB. Also, the proteomic profile of BM and PB T cells was similar as these samples clustered together and separated from the LN samples in the CD4+ cell type. This raises concerns about appropriateness of BM and PB samples for investigating CLL linked exhaustion.

A limma model that normalized differences in library size across samples was constructed to identify differentially expressed T cell subpopulations between CLL (n=23) and control (n=13) LNs. CD4+ naïve, Tex1 PD1hi and Tem1 Ki67+ CD38+ were significantly less frequent in tumor LNs v/s control LNs (figure 3.9). Intriguingly, T cells with higher expression of CD38 are known to promote immune response, and Ki67 is a proliferation marker (for T cells in this case) (Soares, Govender et al. 2010, Konen, Fradette et al. 2019, Santegoets, Duurland et al. 2019). Observation of reduced expression of these cells in CLL tumor LN complements a compromised immune response in CLL. When treatment information is added to the model, the decrease attributed to CLL in naïve and Tem1 Ki67+ CD38+ subpopulations, is rescued. However, it should be noted that increase in CD38 on T cells has also been known to enhance

immune-suppressiveness (Glaria and Valledor 2020). Their rescue after treatment addition, could also imply expansion of a treatment resistant CLL clone.

Subpopulation with characteristic high expression of ICOS surface marker was identified in the CyTOF analysis of the CD4+ T cell type. Lymph node samples contributing a higher proportion of this subpopulation were of IGHV-mutated CLL type. There was a significant difference in contribution of ICOS+ cells between IGHV-mutated and IGHV-unmutated CLL cases (p-values for Wilcoxon rank sum test, performed by Dr. Murat Iskar: CD4+: 0.02). ICOS is known to be involved in development and reactivation of both T and B cells. They possibly contribute to survival of T cell responses (Mahajan, Cervera et al. 2007). This could be a reason for their enrichment in IGHV-mutated CLL cases and diminished expression in aggressive IGHV-unmutated cases.

Further analysis is underway to identify T cell subpopulations that actively engage with CLL cells. This is being performed by Laura Llao Cid and Dr. Murat Iskar utilizing tools like CellPhoneDB (Efremova, Vento-Tormo et al. 2020) and NicheNet (Browaeys, Saelens et al. 2020). Interacting populations will then be followed up for *in-vivo* studies.

Nine CD4+ T cell subpopulations were identified using scRNA sequencing. T-cytotoxic subpopulation was uniquely identified in the transcriptome data as compared to the CyTOF proteome dataset. Several effector-memory T cell subtypes were identified. When patient CD4+ CLL T cell data was compared to that from spleens of Eμ-TCL1 mice, it was found that the mouse TME did not have specific T-effector-memory subpopulations (no expression for *CCR6*, *KLRB1*, *CXCR5* was observed) (figure 3.16 b). Instead 7 out of 12 mouse subpopulations showed expression of T-cytotoxic markers like *GZMB*, *GZMH* and *NKG7*, pointing towards activated T cells potentially engaging with neoantigens. In addition, interferon responders were uniquely identified as one of CD4+ T cell subpopulations in mouse. The physiological and genetic predispositions of the CLL mouse model during disease development are very different from that of CLL in patients including the comparatively fast development of CLL in mice and minimum or no previous infections to generate T-effector-memory cells. These differences potentially contribute to a different TME in the mouse.

scTCR sequencing paired with scRNA seq for 3 CLL lymph node samples identified T cell clonotypes expanding in response to antigens including the ones presented by tumor cells. One of the expanded clonotypes identified using VDJdb (analysis by Laura Llao Cid) was previously reported to expand against DST2 (overexpressed on CLL B cells).

Lastly, human CLL lymph node scRNA profiles for CD4+ and CD8+ subpopulations were compared to lymph node profiles from publicly available breast cancer LN TME data (Azizi, Carr et al. 2018). Most of the subpopulations like Tem, Tregs, and naïve from CLL CD4+ and CD8+ cell types integrated to similar subpopulations in the breast cancer dataset. Interestingly, however a CLL specific subpopulation with mixed CD4+ and CD8+ subpopulations were observed (annotated as Tem1 integrated dataset of two cohorts), that clustered separately from the breast cancer dataset (figure 3.18). It could be speculated that these are cells in the TME of CLL, that have a modulated disease specific transcription profile. Also, lymph nodes of breast cancer patients showed an increased proportion of naïve cell population as compared to CLL lymph nodes. The reason for decreased proportion of naïve subpopulation and increase in effector T cells in CLL lymph nodes as compared to those of breast cancer patients, is that the lymph node is the primary site of tumor cells in CLL.

Limitations

Ibrutinib treatment study should be performed in Eμ-TCL1 mouse tumors with oligoclonal BCRs, to claim whether ibrutinib resistance has a role in modulating BCR evolution culminating in an entirely new ibrutinib resistant BCR clone.

Differences at single cell level in tumor microenvironment profiles of CLL from LN, BM and PB raise concerns about using BM and PB patient samples for studying CLL linked exhaustion. Another limitation of the present study was the absence of controls as reference for scRNA RNA profiles of CLL LNs.

Conclusion

In conclusion, in this thesis I have pointed towards the possible role of '*Myc*' amplification in predisposing the Eμ-TCL1 mouse to CLL. I inferred that there are three distinct patterns of tumor evolution in Eμ-TCL1 mouse tumors, characterized by dynamics of BCR clonotype and SNV-defined clones in primary and serially transplanted tumors. Oligoclonal BCRs were observed in 77% of mouse tumors. In addition, the previous role of autoantigens in CLL progression, was reinstated here by identification of stereotyped B cell receptor rearrangements in the in-house mouse tumor cohort. These observations point towards a heterogenous CLL course in mice. From my investigation of transcriptional profiles of ibrutinib treated tumors, the *Tbet* gene is now being followed up for its role in drug resistance. By the proteomic characterization of the T cell compartment in CLL patients and mouse model, I was able to identify several subpopulations that have roles in CLL pathogenesis, e.g. CD4+ Tem1 Ki67+ CD38+ with potential role in rescuing the immune suppressive CLL niche. Lastly, I also compared CLL LN T cell transcription profiles to those from Eμ-TCL1 mouse and publicly available breast cancer dataset. Comparison with mouse revealed much more heterogenous effector and effector-memory subsets as compared to patients. I was able to identify CLL unique effector-memory subpopulation on comparison with T cells from breast cancer patients.

Many of the observations in this thesis comply with existing knowledge about CLL patient and mouse BCRs, mutation landscape, and CLL TME. This proves that the computational approaches, thresholds, workflows and packages employed here were robust and could be used for similar analyses in the future. These findings are important considerations while designing mechanistic and drug treatment studies in the Eμ-TCL1 mouse, assessing their translational potential in the clinical setting, as well as in independent studies in CLL patients. Varied abundances of T lymphocyte exhausted and effector-memory subpopulations in the CLL LN, BM and PB are important to consider while testing novel CLL immunotherapies in patient samples.

5. Publications

1. Paul Y*, Öztürk S*, Afzal S, Gil-Farina I, Kalter V, Arseni L, Schmidt M, Lichter P, Zapatka M, Seiffert M. Clonal heterogeneity and evolution of malignant B cells in the Eμ-TCL1 mouse model of chronic lymphocytic leukemia. (In preparation)
   *shared co-authorship

2. Ratnaparkhe M, Wong JKL, Wei PC, Hlevnjak M, Kolb T, Simovic M, Haag D, PaulY, Devens F, Northcott P, Jones DTW, Kool M, Jauch A, Pastorczak A, Mlynarski W, Korshunov A, Kumar R, Downing SM, Pfister SM, Zapatka M, McKinnon PJ, Alt FW,Lichter P, Ernst A. Defective DNA damage repair leads to frequent catastrophic genomic events in murine and human tumors. *Nat Commun*. 2018 Nov 12

6. Supplementary Tables

6.1 Supplementary Table 1

Sample	Alternate_Id	Transfer	Gender	Matched control tissue
TCL1_217	A506	Primary	Male	Tail
TCL1_218	D729	Primary	Female	T cells from blood
TCL1_219	B741	Primary	Female	T cells from blood
TCL1_220	TCL1_774	Primary	Female	T cells from blood
TCL1_221 (pTCL1_217)	3A3	Secondary	Female	Same as primary tumor
TCL1_222 (pTCL1_219)	2B1-5	Secondary	Female	Same as primary tumor
TCL1_223 (pTCL1_218)	2D3	Secondary	Female	Same as primary tumor
TCL1_224 (pTCL1_220)	2J3	Secondary	Female	Same as primary tumor

Supplementary Table 1: Eight Eμ-TCL1 mouse tumor cohort used for assessing BCR clonal dynamics and SNV-defined clonotypes. 4 primary and their 4 serially transplanted tumors were used.

6.2 Supplementary Table 2

sample_id	type	treatment	TumorSite	IGHVstatus	gender	age
BC12PB	tumor	untreated	PeripheralBlood	NA	male	77
BC05PB	tumor	untreated	PeripheralBlood	mutated	female	46
BC09PB	tumor	untreated	PeripheralBlood	unmutated	male	53
BC13PB	tumor	untreated	PeripheralBlood	NA	male	67
BC14PB	tumor	treated	PeripheralBlood	unmutated	male	60
BC15PB	tumor	treated	PeripheralBlood	unmutated	female	56
BC1PB	tumor	untreated	PeripheralBlood	mutated	male	70
BC8PB	tumor	treated	PeripheralBlood	unmutated	male	73
BC3LN	tumor	untreated	LymphNode	unmutated	female	57
BC4LN	tumor	untreated	LymphNode	mutated	male	53
BC5LN1	tumor	untreated	LymphNode	mutated	female	46
BC8LN	tumor	treated	LymphNode	unmutated	male	73
BC9LN	tumor	untreated	LymphNode	unmutated	male	53
HD1LN	tumor	pretreated	LymphNode	mutated	male	72
HD2LN	tumor	untreated	LymphNode	mutated	male	77
HD3LN	tumor	untreated	LymphNode	mutated	male	69
HD4LN	tumor	untreated	LymphNode	mutated	male	79
HD5LN	tumor	untreated	LymphNode	unmutated	male	71
HD6LN	tumor	untreated	LymphNode	mutated	male	76
HD8LN	tumor	untreated	LymphNode	mutated	male	72
HD9LN	tumor	untreated	LymphNode	NA	male	75
BC12LN	tumor	untreated	LymphNode	NA	male	77
BC5LN2	tumor	untreated	LymphNode	mutated	female	46
BC13LN	tumor	untreated	LymphNode	NA	male	67
BC14LN	tumor	treated	LymphNode	unmutated	male	60
BC15LN1	tumor	treated	LymphNode	unmutated	female	56
BC15LN2	tumor	treated	LymphNode	unmutated	female	56
HD10LN	tumor	untreated	LymphNode	mutated	male	70
HD11LN	tumor	treated	LymphNode	NA	male	69
BC10LN	tumorHodgkin	untreated	LymphNode	unmutated	female	59
BC09BM	tumor	untreated	BoneMarrow	unmutated	male	53
BC11BM	tumor	treated	BoneMarrow	unmutated	male	52
BC3BM	tumor	untreated	BoneMarrow	unmutated	female	57
BC1LN	tumorAcc	untreated	LymphNode	mutated	male	70
BC2LN	tumorAcc	untreated	LymphNode	unmutated	male	69
RLN1	control	untreated	LymphNode	NA	male	36

RLN2	control	untreated	LymphNode	NA	male	32
RLN3	control	untreated	LymphNode	NA	male	50
RLN4	control	untreated	LymphNode	NA	female	39
RLN5	control	untreated	LymphNode	NA	male	20
RLN6	control	untreated	LymphNode	NA	male	49
RLN7	control	untreated	LymphNode	NA	male	18
RLN8	control	untreated	LymphNode	NA	male	33
RLN10	control	untreated	LymphNode	NA	female	52
RLN9	control	untreated	LymphNode	NA	male	20
RLN11	control	untreated	LymphNode	NA	female	34
RLN12	control	untreated	LymphNode	NA	male	30
RLN13	control	untreated	LymphNode	NA	male	59

Supplementary Table 2: 48 samples used for CyTOF analysis. Out of these 8 samples were from patient peripheral blood (PB), 3 from patient bone marrow (BM), 23 from patient lymph node (LN), and 13 from control lymoh nodes (rLN).2 LN samples were from the same patient.

6.3 Supplementary Table 3

Antigen	Isotope	Clone	Manufacturer	
Exhaustion				
2B4	113 In	C1.7	Biolegend	
KLRG1	115 In	SA231A2	Biolegend	
CD278/ICOS	148Nd	C398.4A	Fluidigm	
TIGIT	153Eu	MBSA43	Fluidigm	
CD152 (CTLA-4)	170Er	14D3	Fluidigm	
CD279 (PD-1)	174Yb	EH12.2H7	Fluidigm	
CD38	144Nd	HIT2	Fluidigm	
CD47	209Bi	CC2C6	Fluidigm	
Enzymes				
CD73 (Ecto-5-nucleotidase)	168Er	AD2	Fluidigm	
CD39	160Gd	A1	Fluidigm	
General (for cell type discrimination)				
CD3	141Pr	UCHT1	Biolegend	T cells
CD4	145Nd	RPA-T4	Biolegend	CD4 T cells
CD8a	146Nd	RPA-T8	Biolegend	CD8 T cells
FoxP3	162Dy	259D/C7	Fluidigm	Tregs
CD25	169Tm	2A3	Fluidigm	Tregs (but not only)
CD45RA	143Nd	H100	Fluidigm	
CD45RO	164Dy	UCHL1	Fluidigm	
CD27	155Gd	L128	Fluidigm	
CD197 (CCR7)	167Er	G043H7	Biolegend	
CD7	147Sm	CD7-6B7	Fluidigm	
CD127 (IL-7Ra)	149Sm	A019D5	Fluidigm	
CD185 (CXCR5)	171Yb	51505	Fluidigm	
Cytokines				
FAS	152Sm	Mab11	Fluidigm	
granzyme K	142Nd	GM26E7	Biolegend	
OX40	158Gd	B27	Fluidigm	
CD44	166Er	MQ1-17H12	Fluidigm	
Transcription Factors				
TCF1	163Dy	7F11A10	Biolegend	
Tbet	161Dy	4B10	Fluidigm	
Eomes	175Lu	WD1928	ebioscience	
tox	150Nd	TXRX10	ebioscience	
HELIOS	156Gd	22F6	Biolegend	
Proliferation				

Ki-67	172Yb	B56	Fluidigm	
Technical				
CD85J	190BCKG			
Ba138	138Ba			
Ce140	Ce140			
HLA_DR	151Eu			
127I	127I			
DNA1	191Ir			
DNA2	193Ir			
207Pb	Pb207			
208Pb	Pb208			
195Pt	Pt195			
196Pt	Pt196			

Supplementary Table 3: Panel of 43 markers used for CyTOF measurements. They have been divided into sections depending upon their characteristic phenotype.

Acknowledgements

I am extremely grateful to Prof. Dr. Peter Lichter for giving me the opportunity to conduct my doctoral research in his division. He has been extremely kind and supportive. I thank him for his time, effort and invaluable suggestions during ma PhD.

I thank my first supervisor Prof. Dr. Benedikt Brors, to have taken out the time to provide his suggestions and supervision, during TAC meetings and otherwise.

In addition, I thank Dr. Sascha Dietrich for being available with his expert knowledge on the clinical aspects of CLL and his suggestions during TAC meetings. He as a part of University Klinikum, Heidelberg has been instrumental in providing CLL and control lymph node samples used for analysis in this thesis.

Furthermore, I extend my thanks to Prof. Dr. Stefan Wiemann and Dr. Michael Milsom for their willingness to participate in my thesis examination committee. I am grateful to them for their time and effort.

My heartfelt thanks goes to my immediate supervisor, Dr. Marc Zapatka for his unmatched patience, support and excellent supervision. He has helped me grow, learn and think independently. He has offered immense flexibility, encouragement and provided direction whenever needed. I thank him for believing in me and extending his support in helping me hone, not only my research abilities but also scientific communication. His guidance has been invaluable. I am grateful to him for all his efforts and also for allowing me to present my work at various national and international conferences.

I would like to thank Michael Hain and Rolf Kabbe for their outstanding technical assistance and immediate response to technical disruptions at all times.

I also thank our collaborators from University Clinic Barcelona (Dolors Colomer and Elías Campo) and Luxembourg Institute of Health (Marina Wierz, Etienne Moussay, Jérôme Paggetti) for providing us with samples and CyTOF raw data that have been analysed in this thesis.

I especially thank group leader Dr. Martina Seiffert for leading the experimentation part of this thesis, being enthusiastic for discussing and providing inputs related to biological aspects of the projects.

I immensely thank Dr. Selcen Özturk, Haniyeh Yazdanparast, Dr. Lavinia Arseni and Laura Llao Cid for contributing towards all the experimentation, mouse work and sample preparations that made it possible for me to perform all my analysis. They have been very helpful throughout and answered all my questions about the experimental part. I especially thank Laura for her biological insights on the single cell project at every step. It was very interesting to collaborate with her and shape the project.

I am grateful to my group members Christian Aichmüller, Dr. Murat Iskar, Dr. John Wong, Dr. Mario Hlevnjak, Dr. Markus Schulze, and Dr. Michael Fletcher for stimulating discussions, their collaborative spirit and a positive environment always. Christian was a very helpful fellow PhD student all throughout.

Furthermore, I thank all my colleagues in the B060, B061 for a supportive and friendly environment and inputs during seminars.

I want to thank my fiancé Akshay Patil for his patience and support at home, so that I could concentrate on my PhD. I want to thank my super encouraging and helpful friends Zahid, Manish, Jaya, Revati, Neha, Harsh, Kriti and Manu, who were readily available for the moral support I needed during the course of my PhD.

Lastly, I want to thank my sister Kanika Paul, and my parents who have been my ultimate pillars of strength and have helped me see the light at the end of the tunnel. This would have been impossible without their indispensable support, sacrifices and understanding.

Bibliography

References for links

1. https://apps.who.int/bookorders/anglais/detart1.jsp?codlan=1&codcol=76&codcch=16
2. https://www.healthypeople.gov/2020/topics-objectives/topic/cancer
3. https://www.ejcancer.com/article/S0959-8049(18)30955-9/abstract
4. https://www.who.int/cancer/prevention/en/
5. https://abucketfullofscience.wordpress.com/2017/01/23/the-biological-hallmarks-of-cancer-what-makes-a-cancer-cell-cancerous/
6. https://www.medicalnewstoday.com/articles/322700
7. https://en.wikipedia.org/wiki/List_of_cancer_mortality_rates_in_the_United_States
8. https://www.texasoncology.com/types-of-cancer/leukemia/chronic-lymphocytic-leukemia/relapsed-chronic-lymphocytic-leukemia
9. https://www.creative-diagnostics.com/t-cell-differentiation.htm
10. https://en.wikipedia.org/wiki/Exome_sequencing#Applications_of_exome_sequencing
11. https://en.wikipedia.org/wiki/Illumina_dye_sequencing#cite_note-Jeon-11
12. https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/
13. https://assets.ctfassets.net/an68im79xiti/3UCNucSkmZdeadsIM0Dxfz/aaa61ca86d5ab0b74b75eeb222ed7de5/CG000207_ChromiumNextGEMSingleCellV_D_J_ReagentKits_v1.1_UG_RevE.pdf
14. https://kasperdanielhansen.github.io/genbioconductor/html/limma.html
15. https://ase.tufts.edu/chemistry/walt/sepa/Activities/jaccardPractice.pdf
16. https://documentation.partek.com/display/FLOWDOC/Graph-based+clustering
17. https://neo4j.com/blog/graph-algorithms-neo4j-louvain-modularity/

References for journal articles

Abdin, S. M., D. M. Zaher, E. A. Arafa and H. A. Omar (2018). "Tackling Cancer Resistance by Immunotherapy: Updated Clinical Impact and Safety of PD-1/PD-L1 Inhibitors." Cancers (Basel) **10**(2).

Afzal, S. (2019). "Systematic Comparative Study of Computational Methods for T-cell Receptor Sequencing Data Analysis " Briefings in bioinformatics.

Ahn, I. E., C. Underbayev, A. Albitar, S. E. Herman, X. Tian, I. Maric, D. C. Arthur, L. Wake, S. Pittaluga, C. M. Yuan, M. Stetler-Stevenson, S. Soto, J. Valdez, P. Nierman, J. Lotter, L. Xi, M. Raffeld, M. Farooqui, M. Albitar and A. Wiestner (2017). "Clonal evolution leading to ibrutinib resistance in chronic lymphocytic leukemia." Blood **129**(11): 1469-1479.

Angelousi, A., E. Chatzellis and G. Kaltsas (2018). "New Molecular, Biological, and Immunological Agents Inducing Hypophysitis." Neuroendocrinology **106**(1): 89-100.

Arber, D. A., A. Orazi, R. Hasserjian, J. Thiele, M. J. Borowitz, M. M. Le Beau, C. D. Bloomfield, M. Cazzola and J. W. Vardiman (2016). "The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia." Blood **127**(20): 2391-2405.

Armstrong, R. A. (2014). "When to Use the Bonferroni Correction." Ophthalmic Physiol Opt.

Azizi, E., A. J. Carr, G. Plitas, A. E. Cornish, C. Konopacki, S. Prabhakaran, J. Nainys, K. Wu, V. Kiseliovas, M. Setty, K. Choi, R. M. Fromme, P. Dao, P. T. McKenney, R. C. Wasti, K. Kadaveru, L. Mazutis, A. Y. Rudensky and D. Pe'er (2018). "Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment." Cell **174**(5): 1293-1308 e1236.

Barnett, B. E., R. P. Staupe, P. M. Odorizzi, O. Palko, V. T. Tomov, A. E. Mahan, B. Gunn, D. Chen, M. A. Paley, G. Alter, S. L. Reiner, G. M. Lauer, J. R. Teijaro and E. J. Wherry (2016). "Cutting Edge: B Cell-Intrinsic T-bet Expression Is Required To Control Chronic Viral Infection." J Immunol **197**(4): 1017-1022.

Bendall, S. C., K. L. Davis, A. D. Amir el, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan and D. Pe'er (2014). "Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development." Cell **157**(3): 714-725.

Benjamin, D., T. Sato, K. Cibulskis, G. Getz, C. Stewart and L. Lichtenstein (2019). "Calling Somatic SNVs and Indels with Mutect2."

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society.

Bichi, R., S. A. Shinton, E. S. Martin, A. Koval, G. A. Calin, R. Cesari, G. Russo, R. R. Hardy and C. M. Croce (2002). "Human chronic lymphocytic leukemia modeled in mouse by targeted TCL1 expression." Proc Natl Acad Sci U S A **99**(10): 6955-6960.

Blondel, V. D. (2008). "Fast unfolding of communities in large networks." Journal of Statistical Mechanics: Theory and Experiment.

Bolotin, D. A., S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, E. V. Putintseva and D. M. Chudakov (2015). "MiXCR: software for comprehensive adaptive immunity profiling." Nat Methods **12**(5): 380-381.

Browaeys, R., W. Saelens and Y. Saeys (2020). "NicheNet: modeling intercellular communication by linking ligands to target genes." Nat Methods **17**(2): 159-162.

Bruggner, R. V., B. Bodenmiller, D. L. Dill, R. J. Tibshirani and G. P. Nolan (2014). "Automated identification of stratifying signatures in cellular subpopulations." Proc Natl Acad Sci U S A **111**(26): E2770-2777.

Burger, J. A. (2011). "Nurture versus Nature: The Microenvironment inChronic Lymphocytic Leukemia."

Butler, A., P. Hoffman, P. Smibert, E. Papalexi and R. Satija (2018). "Integrating single-cell transcriptomic data across different conditions, technologies, and species." Nat Biotechnol **36**(5): 411-420.

Chen, R. (2015). "Whole-Exome Enrichment with the Agilent SureSelect Human All Exon Platform." Cold Spring Harb Protoc.

Chen, S.-S. (2010). "Murine TCL1 CLL Cells with B-Cell Receptors Specific for the Autoantigen Phosphatidylcholine Have a Selective Advantage During Adoptive Transfer." Blood.

Darwiche, W., B. Gubler, J. P. Marolleau and H. Ghamlouch (2018). "Chronic Lymphocytic Leukemia B-Cell Normal Cellular Counterpart: Clues From a Functional Perspective." Front Immunol **9**: 683.

DeLuca, D. S., J. Z. Levin, A. Sivachenko, T. Fennell, M. D. Nazaire, C. Williams, M. Reich, W. Winckler and G. Getz (2012). "RNA-SeQC: RNA-seq metrics for quality control and process optimization." Bioinformatics **28**(11): 1530-1532.

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

Domenech, E., G. Gomez-Lopez, D. Gzlez-Pena, M. Lopez, B. Herreros, J. Menezes, N. Gomez-Lozano, A. Carro, O. Grana, D. G. Pisano, O. Dominguez, J. A. Garcia-Marco, M. A. Piris and M. Sanchez-Beato (2012). "New mutations in chronic lymphocytic leukemia identified by target enrichment and deep sequencing." PLoS One **7**(6): e38158.

Duhren-von Minden, M., R. Ubelhart, D. Schneider, T. Wossning, M. P. Bach, M. Buchner, D. Hofmann, E. Surova, M. Follo, F. Kohler, H. Wardemann, K. Zirlik, H. Veelken and H. Jumaa

(2012). "Chronic lymphocytic leukaemia is driven by antigen-independent cell-autonomous signalling." Nature **489**(7415): 309-312.

Efremov, D. G. (1996). "IgM-producing Chronic Lymphocytic Leukemia Cells Undergo Immunoglobulin Isotype-switching without Acquiring Somatic Mutations." J. Clin. Invest.

Efremova, M., M. Vento-Tormo, S. A. Teichmann and R. Vento-Tormo (2020). "CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes." Nat Protoc **15**(4): 1484-1506.

Eichhorst, B. F., R. Busch, G. Hopfinger, R. Pasold, M. Hensel, C. Steinbrecher, S. Siehl, U. Jager, M. Bergmann, S. Stilgenbauer, C. Schweighofer, C. M. Wendtner, H. Dohner, G. Brittinger, B. Emmerich, M. Hallek and C. L. L. S. G. German (2006). "Fludarabine plus cyclophosphamide versus fludarabine alone in first-line therapy of younger patients with chronic lymphocytic leukemia." Blood **107**(3): 885-891.

Elyahu, Y. (2019). "**Aging promotes reorganization oftheCD4 T cell landscape toward extreme regulatory andeffector phenotypes**." immunology.

Fais, F. (1996). "Examples of In Vivo Isotype Class Switching In IgM1 Chronic Lymphocytic Leukemia B Cells."

Ferrer, G. and E. Montserrat (2018). "Critical molecular pathways in CLL therapy." Mol Med **24**(1): 9.

Frankish, A., M. Diekhans, A. M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. Garcia Giron, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martinez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M. M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigo, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress and P. Flicek (2019). "GENCODE reference annotation for the human and mouse genomes." Nucleic Acids Res **47**(D1): D766-D773.

Gaidano, G. (2017). "The mutational landscape of chronic lymphocytic leukemia and its impact on prognosis and treatment." Hematology.

Garcia-Munoz, R., V. R. Galiacho and L. Llorente (2012). "Immunological aspects in chronic lymphocytic leukemia (CLL) development." Ann Hematol **91**(7): 981-996.

Glaria, E. and A. F. Valledor (2020). "Roles of CD38 in the Immune Response to Infection." Cells **9**(1).

Golubovskaya, V. and L. Wu (2016). "Different Subsets of T Cells, Memory, Effector Functions, and CAR-T Immunotherapy." Cancers (Basel) **8**(3).

Gong, S. (2015). "CD317 is over-expressed in B-cell chronic lymphocytic leukemia, but not B-cell acute lymphoblastic leukemia " Int J Clin Exp Pathology.

Grabowski, P., M. Hultdin, K. Karlsson, G. Tobin, A. Aleskog, U. Thunberg, A. Laurell, C. Sundstrom, R. Rosenquist and G. Roos (2005). "Telomere length as a prognostic parameter in chronic lymphocytic leukemia with special reference to VH gene mutation status." Blood **105**(12): 4807-4812.

Groneberg, C., T. Pickartz, A. Binder, F. Ringel, S. Srock, T. Sieber, D. Schoeler and F. Schriever (2003). "Clinical relevance of CD95 (Fas/Apo-1) on T cells of patients with B-cell chronic lymphocytic leukemia." Experimental Hematology **31**(8): 682-685.

Guieze, R. and C. J. Wu (2015). "Genomic and epigenomic heterogeneity in chronic lymphocytic leukemia." Blood **126**(4): 445-453.

Hafemeister, C. and R. Satija (2019). "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression." Genome Biol **20**(1): 296.

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

Hahne, F., N. LeMeur, R. R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain and R. Gentleman (2009). "flowCore: a Bioconductor package for high throughput flow cytometry." BMC Bioinformatics **10**: 106.

Hallek, M. (2019). "Chronic lymphocytic leukemia: 2020 update on diagnosis, risk stratification and treatment." Am J Hematol **94**(11): 1266-1287.

Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.

Hanna, B. S. (2020). "IL-10 receptor signaling alters chromatin accessibility in CD8+ T-cells preventing their exhaustion and immune escape in chronic lymphocytic leukemia."

Hanna, B. S., F. McClanahan, H. Yazdanparast, N. Zaborsky, V. Kalter, P. M. Rossner, A. Benner, C. Durr, A. Egle, J. G. Gribben, P. Lichter and M. Seiffert (2016). "Depletion of CLL-associated patrolling monocytes and macrophages controls disease development and repairs immune dysfunction in vivo." Leukemia **30**(3): 570-579.

Hanna, B. S., H. Yazdanparast, Y. Demerdash, P. M. Roessner, R. Schulz, P. Lichter, S. Stilgenbauer and M. Seiffert (2020). "Combining ibrutinib and checkpoint blockade improves CD8+ T-cell function and control of chronic lymphocytic leukemia in Em-TCL1 mice." Haematologica.

Hashimoto, K., T. Kouno, T. Ikawa, N. Hayatsu, Y. Miyajima, H. Yabukami, T. Terooatea, T. Sasaki, T. Suzuki, M. Valentine, G. Pascarella, Y. Okazaki, H. Suzuki, J. W. Shin, A. Minoda, I.

Taniuchi, H. Okano, Y. Arai, N. Hirose and P. Carninci (2019). "Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians." Proc Natl Acad Sci U S A **116**(48): 24242-24251.

Hayakawa, K., A. M. Formica, J. Brill-Dashoff, S. A. Shinton, D. Ichikawa, Y. Zhou, H. C. Morse, 3rd and R. R. Hardy (2016). "Early generated B1 B cells with restricted BCRs become chronic lymphocytic leukemia with continued c-Myc and low Bmf expression." J Exp Med **213**(13): 3007-3024.

Huhn, D., C. von Schilling, M. Wilhelm, A. D. Ho, M. Hallek, R. Kuse, W. Knauf, U. Riedel, A. Hinke, S. Srock, S. Serke, C. Peschel, B. Emmerich and G. German Chronic Lymphocytic Leukemia Study (2001). "Rituximab therapy of patients with B-cell chronic lymphocytic leukemia." Blood **98**(5): 1326-1331.

Ji, N., R. A. Sosa and T. G. Forsthuber (2011). "More than just a T-box: the role of T-bet as a possible biomarker and therapeutic target in autoimmune diseases." Immunotherapy **3**(3): 435-441.

Jimenez de Oya, N., M. De Giovanni, J. Fioravanti, R. Ubelhart, P. Di Lucia, A. Fiocchi, S. Iacovelli, D. G. Efremov, F. Caligaris-Cappio, H. Jumaa, P. Ghia, L. G. Guidotti and M. Iannacone (2017). "Pathogen-specific B-cell receptors drive chronic lymphocytic leukemia by light-chain-dependent cross-reaction with autoantigens." EMBO Mol Med **9**(11): 1482-1490.

Karolchik, D., A. S. Hinrichs and W. J. Kent (2009). "The UCSC Genome Browser." Curr Protoc Bioinformatics **Chapter 1**: Unit1 4.

Konen, J. M., J. J. Fradette and D. L. Gibbons (2019). "The Good, the Bad and the Unknown of CD38 in the Metabolic Microenvironment and Immune Cell Functionality of Solid Tumors." Cells **9**(1).

Lascano, V. (2013). "Chronic lymphocytic leukemia disease progression is accelerated by APRIL-TACI interaction in the TCL1 transgenic mouse model." Blood.

Law, C. W., M. Alhamdoosh, S. Su, X. Dong, L. Tian, G. K. Smyth and M. E. Ritchie (2016). "RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR." F1000Res **5**.

Levine, J. H., E. F. Simonds, S. C. Bendall, K. L. Davis, A. D. Amir el, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, R. Finck, A. L. Gedman, I. Radtke, J. R. Downing, D. Pe'er and G. P. Nolan (2015). "Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis." Cell **162**(1): 184-197.

Liu, X., W. Song, B. Y. Wong, T. Zhang, S. Yu, G. N. Lin and X. Ding (2019). "A comparison framework and guideline of clustering methods for mass cytometry data." Genome Biol **20**(1): 297.

Long, L. (2018). "The promising immune checkpoint LAG-3: from tumor microenvironment to cancer immunotherapy." Genes & Cancer.

Lou, X. (2007). "Polymer-Based Elemental Tags for Sensitive Bioassays." Angew Chem Int Ed Engl.

Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.

Mahajan, S., A. Cervera, M. MacLeod, S. Fillatreau, G. Perona-Wright, S. Meek, A. Smith, A. MacDonald and D. Gray (2007). "The role of ICOS in the development of CD4 T cell help and the reactivation of memory T cells." Eur J Immunol **37**(7): 1796-1808.

Marchetti, A., A. Di Lorito and F. Buttitta (2017). "Why anti-PD1/PDL1 therapy is so effective? Another piece in the puzzle." J Thorac Dis **9**(12): 4863-4866.

Matrai, Z. (2005). "CD38 as a prognostic marker in CLL." Hematology **10**(1): 39-46.

McClanahan, F., J. C. Riches, S. Miller, W. P. Day, E. Kotsiou, D. Neuberg, C. M. Croce, M. Capasso and J. G. Gribben (2015). "Mechanisms of PD-L1/PD-1-mediated CD8 T-cell dysfunction in the context of aging-related immune defects in the Emicro-TCL1 CLL mouse model." Blood **126**(2): 212-221.

Mittal, A. K., N. K. Chaturvedi, K. J. Rai, C. E. Gilling-Cutucache, T. M. Nordgren, M. Moragues, R. Lu, R. Opavsky, G. R. Bociek, D. D. Weisenburger, J. Iqbal and S. S. Joshi (2014). "Chronic lymphocytic leukemia cells in a lymph node microenvironment depict molecular signature associated with an aggressive disease." Mol Med **20**: 290-301.

Mittal, A. K., N. K. Chaturvedi, R. A. Rohlfsen, P. Gupta, A. D. Joshi, G. V. Hegde, R. G. Bociek and S. S. Joshi (2013). "Role of CTLA4 in the proliferation and survival of chronic lymphocytic leukemia." PLoS One **8**(8): e70352.

Multhoff, G., M. Molls and J. Radons (2011). "Chronic inflammation in cancer development." Front Immunol **2**: 98.

Nowicka, M., C. Krieg, H. L. Crowell, L. M. Weber, F. J. Hartmann, S. Guglietta, B. Becher, M. P. Levesque and M. D. Robinson (2017). "CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets." F1000Res **6**: 748.

Orbanz, P. and Y. W. Teh "Bayesian Nonparametric Models."

Ornatsky, O. (2008). "Study of Cell Antigens and Intracellular DNA by Identification of Element-Containing Labels and Metallointercalators Using Inductively Coupled Plasma Mass Spectrometry." Analytical Chemistry.

Pekarsky, Y., A. Drusco, P. Kumchala, C. M. Croce and N. Zanesi (2015). "The long journey of TCL1 transgenic mice: lessons learned in the last 15 years." Gene Expr **16**(3): 129-135.

Petrova, V. N., L. Muir, P. F. McKay, G. S. Vassiliou, K. G. C. Smith, P. A. Lyons, C. A. Russell, C. A. Anderson, P. Kellam and R. J. M. Bashford-Rogers (2018). "Combined Influence of B-Cell

Receptor Rearrangement and Somatic Hypermutation on B-Cell Class-Switch Fate in Health and in Chronic Lymphocytic Leukemia." Front Immunol **9**: 1784.

Qin, L., T. C. Waseem, A. Sahoo, S. Bieerkehazhi, H. Zhou, E. V. Galkina and R. Nurieva (2018). "Insights Into the Molecular Mechanisms of T Follicular Helper-Mediated Immunity and Pathology." Front Immunol **9**: 1884.

Raczkowski, F., A. Rissiek, I. Ricklefs, K. Heiss, V. Schumacher, K. Wundenberg, F. Haag, F. Koch-Nolte, E. Tolosa and H. W. Mittrucker (2018). "CD39 is upregulated during activation of mouse and human T cells and attenuates the immune response to Listeria monocytogenes." PLoS One **13**(5): e0197151.

Reppert, S., I. Boross, M. Koslowski, O. Tureci, S. Koch, H. A. Lehr and S. Finotto (2011). "A role for T-bet-mediated tumour immune surveillance in anti-IL-17A treatment of lung cancer." Nat Commun **2**: 600.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Res **43**(7): e47.

Robinson, M. D. (2010). "A scaling normalization method for differentialexpression analysis of RNA-seq data." Genome Biology.

Roth, A., J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Cote and S. P. Shah (2014). "PyClone: statistical inference of clonal population structure in cancer." Nat Methods **11**(4): 396-398.

S.Hanna, B. (2019). "Beyond bystanders: Myeloid cells in chronic lymphocytic leukemia." Molecular Immunology.

Santegoets, S. J., C. L. Duurland, E. S. Jordanova, J. J. van Ham, I. Ehsan, S. L. van Egmond, M. J. P. Welters and S. H. van der Burg (2019). "Tbet-positive regulatory T cells accumulate in oropharyngeal cancers with ongoing tumor-specific type 1 T cell responses." J Immunother Cancer **7**(1): 14.

Sarkar, M., Y. Liu, J. Qi, H. Peng, J. Morimoto, C. Rader, N. Chiorazzi and T. Kodadek (2016). "Targeting Stereotyped B Cell Receptors from Chronic Lymphocytic Leukemia Patients with Synthetic Antigen Surrogates." J Biol Chem **291**(14): 7558-7570.

See, P., J. Lum, J. Chen and F. Ginhoux (2018). "A Single-Cell Sequencing Guide for Immunologists." Front Immunol **9**: 2425.

Seidel, J. A., A. Otsuka and K. Kabashima (2018). "Anti-PD-1 and Anti-CTLA-4 Therapies in Cancer: Mechanisms of Action, Efficacy, and Limitations." Front Oncol **8**: 86.

Shen, R. R. (2006). "Dysregulated TCL1 requires the germinal center and genome instability for mature B-cell transformation." Blood.

Shugay, M., D. V. Bagaev, I. V. Zvyagin, R. M. Vroomans, J. C. Crawford, G. Dolton, E. A. Komech, A. L. Sycheva, A. E. Koneva, E. S. Egorov, A. V. Eliseev, E. Van Dyk, P. Dash, M. Attaf, C. Rius, K. Ladell, J. E. McLaren, K. K. Matthews, E. B. Clemens, D. C. Douek, F. Luciani, D. van Baarle, K. Kedzierska, C. Kesmir, P. G. Thomas, D. A. Price, A. K. Sewell and D. M. Chudakov (2018). "VDJdb: a curated database of T-cell receptor sequences with known antigen specificity." Nucleic Acids Res **46**(D1): D419-D427.

Shugay, M., O. V. Britanova, E. M. Merzlyak, M. A. Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. A. Bolotin, D. B. Staroverov, E. V. Putintseva, K. Plevova, C. Linnemann, D. Shagin, S. Pospisilova, S. Lukyanov, T. N. Schumacher and D. M. Chudakov (2014). "Towards error-free profiling of immune repertoires." Nat Methods **11**(6): 653-655.

Slupsky, J. R. (2014). "Does B cell receptor signaling in chronic lymphocytic leukaemia cells differ from that in other B cell types?" Scientifica (Cairo) **2014**: 208928.

Soares, A., L. Govender, J. Hughes, W. Mavakla, M. de Kock, C. Barnard, B. Pienaar, E. Janse van Rensburg, G. Jacobs, G. Khomba, L. Stone, B. Abel, T. J. Scriba and W. A. Hanekom (2010). "Novel application of Ki67 to quantify antigen-specific in vitro lymphoproliferation." J Immunol Methods **362**(1-2): 43-50.

Souers, A. J., J. D. Leverson, E. R. Boghaert, S. L. Ackler, N. D. Catron, J. Chen, B. D. Dayton, H. Ding, S. H. Enschede, W. J. Fairbrother, D. C. Huang, S. G. Hymowitz, S. Jin, S. L. Khaw, P. J. Kovar, L. T. Lam, J. Lee, H. L. Maecker, K. C. Marsh, K. D. Mason, M. J. Mitten, P. M. Nimmer, A. Oleksijew, C. H. Park, C. M. Park, D. C. Phillips, A. W. Roberts, D. Sampath, J. F. Seymour, M. L. Smith, G. M. Sullivan, S. K. Tahir, C. Tse, M. D. Wendt, Y. Xiao, J. C. Xue, H. Zhang, R. A. Humerickhouse, S. H. Rosenberg and S. W. Elmore (2013). "ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets." Nat Med **19**(2): 202-208.
Spitzer, M. H. and G. P. Nolan (2016). "Mass Cytometry: Single Cells, Many Features." Cell **165**(4): 780-791.

Stuart, T. (2019). "Comprehensive Integration of Single-Cell Data." Cell.

Stuart, T., A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, 3rd, Y. Hao, M. Stoeckius, P. Smibert and R. Satija (2019). "Comprehensive Integration of Single-Cell Data." Cell **177**(7): 1888-1902 e1821.

Takeuchi, A. and T. Saito (2017). "CD4 CTL, a Cytotoxic Subset of CD4(+) T Cells, Their Differentiation and Function." Front Immunol **8**: 194.

Talevich, E., A. H. Shain, T. Botton and B. C. Bastian (2016). "CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing." PLoS Comput Biol **12**(4): e1004873.

Ten Hacken, E. and J. A. Burger (2016). "Microenvironment interactions and B-cell receptor signaling in Chronic Lymphocytic Leukemia: Implications for disease pathogenesis and treatment." Biochim Biophys Acta **1863**(3): 401-413.

Turchaninova, M. A. (2016). "High-quality Full-Length Immunoglobulin Profiling With Unique Molecular Barcoding " Nature Protocols.

Van Gassen, S., B. Callebaut, M. J. Van Helden, B. N. Lambrecht, P. Demeester, T. Dhaene and Y. Saeys (2015). "FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data." Cytometry A **87**(7): 636-645.

Vecellio, M., C. J. Cohen, A. R. Roberts, P. B. Wordsworth and T. J. Kenna (2018). "RUNX3 and T-Bet in Immunopathogenesis of Ankylosing Spondylitis-Novel Targets for Therapy?" Front Immunol **9**: 3132.

Verdeil, G. (2016). "MAF drives CD8(+) T-cell exhaustion." Oncoimmunology **5**(2): e1082707.

Wang, K., M. Li and H. Hakonarson (2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic Acids Res **38**(16): e164.

Weber, L. M. and M. D. Robinson (2016). "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data." Cytometry A **89**(12): 1084-1096.

Weigmann, B. and M. F. Neurath (2002). "T-bet as a possible therapeutic target in autoimmune disease." Expert Opin Ther Targets **6**(6): 619-622.

Wherry, E. J. (2011). "T cell exhaustion." Nat Immunol **12**(6): 492-499.

Wherry, E. J., S. J. Ha, S. M. Kaech, W. N. Haining, S. Sarkar, V. Kalia, S. Subramaniam, J. N. Blattman, D. L. Barber and R. Ahmed (2007). "Molecular signature of CD8+ T cell exhaustion during chronic viral infection." Immunity **27**(4): 670-684.

Wherry, E. J. and M. Kurachi (2015). "Molecular and cellular insights into T cell exhaustion." Nat Rev Immunol **15**(8): 486-499.

Wierz, M. (2018). "Dual PD1/LAG3 immune checkpoint blockade limits tumor development in a murine model of chronic lymphocytic leukemia." Blood.

Wiestner, C. S. a. A. (2017). "CLL kinetics in the tumor microenvironment." Oncotarget.

Wilkerson, M. D. and D. N. Hayes (2010). "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking." Bioinformatics **26**(12): 1572-1573.

Xu, C. and Z. Su (2015). "Identification of cell types from single-cell transcriptomes using a novel clustering method." Bioinformatics **31**(12): 1974-1980.

Yan, X. J., E. Albesiano, N. Zanesi, S. Yancopoulos, A. Sawyer, E. Romano, A. Petlickovski, D. G. Efremov, C. M. Croce and N. Chiorazzi (2006). "B cell receptors in TCL1 transgenic mice resemble those of aggressive, treatment-resistant human chronic lymphocytic leukemia." Proc Natl Acad Sci U S A **103**(31): 11713-11718.

Yao, Y. and W. Dai (2014). "Genomic Instability and Cancer." J Carcinog Mutagen **5**.

Yosifov, D. Y., C. Wolf, S. Stilgenbauer and D. Mertens (2019). "From Biology to Therapy: The CLL Success Story." Hemasphere **3**(2): e175.

Yuille, M. R. (2001). "TCL1 Is Activated by Chromosomal Rearrangementor by Hypomethylation." Wiley-Liss, Inc.

Zaborsky, N., F. J. Gassner, J. P. Hopner, M. Schubert, D. Hebenstreit, R. Stark, D. Asslaber, M. Steiner, R. Geisberger, R. Greil and A. Egle (2019). "Exome sequencing of the TCL1 mouse model for CLL reveals genetic heterogeneity and dynamics during disease development." Leukemia **33**(4): 957-968.

Zaki, M. (2000). "Disruption of the IFN-gcytokine network in chronic lymphocyticleukemia contributes to resistance of leukemic B cells toapoptosis." Leukemia Research.

Zhang, X., T. Li, F. Liu, Y. Chen, J. Yao, Z. Li, Y. Huang and J. Wang (2019). "Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems." Mol Cell **73**(1): 130-142 e135.

Zheng, C., L. Zheng, J. K. Yoo, H. Guo, Y. Zhang, X. Guo, B. Kang, R. Hu, J. Y. Huang, Q. Zhang, Z. Liu, M. Dong, X. Hu, W. Ouyang, J. Peng and Z. Zhang (2017). "Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing." Cell **169**(7): 1342-1356 e1316.

Zheng, G. X., J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson and J. H. Bielas (2017). "Massively parallel digital transcriptional profiling of single cells." Nat Commun **8**: 14049.