Dissecting tumor microenvironments of blood cancers using single-cell expression data

Lemvigh, Camilla Koldbæk

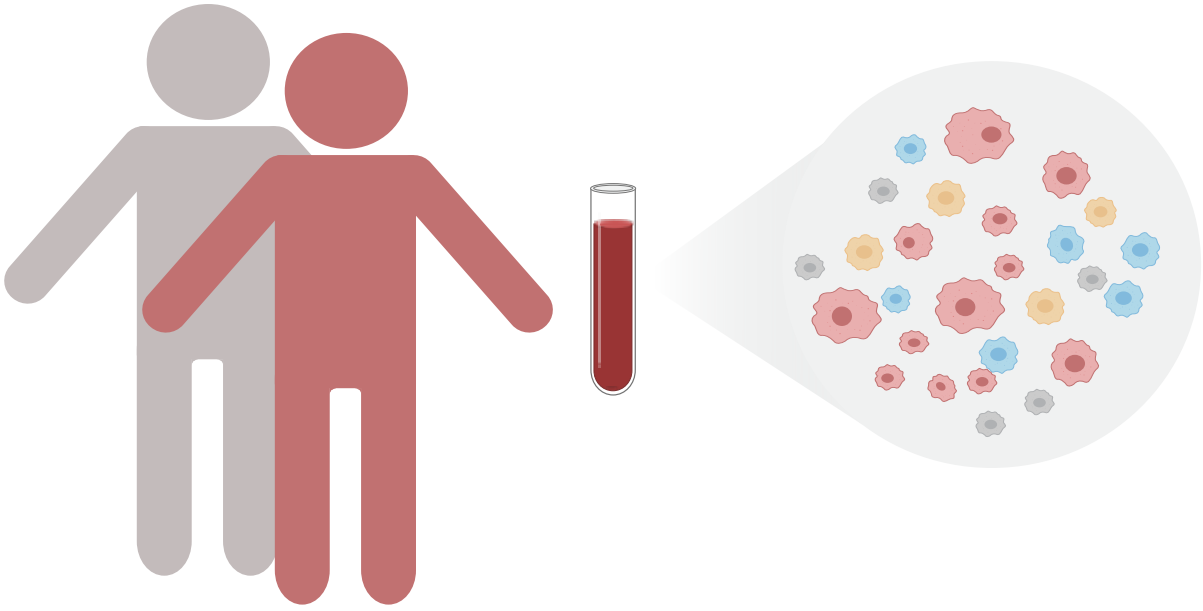
Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

DTU Health Tech
Section for Bioinformatics

Dissecting tumor microenvironments of blood cancers using single-cell expression data

PhD Thesis



Camilla Koldbæk Lemvigh
December 2022

Preface

This thesis presents work carried out from February 2019 to December 2022 in a collaboration between the Wu Lab from Dana-Farber Cancer Institute and the Single Cell Omics group, Section for Bioinformatics, DTU Health Tech.

The work was supervised by Associate Professor Lars Rønn Olsen (DTU) and Professor Catherine J. Wu, (Dana-Farber Cancer Institute), and likewise jointly financed. The majority of the work was conducted at the Section for Bioinformatics, DTU. Part time was also spent on a research stay with co-supervisor and collaborators in the Wu Lab in Boston.

Camilla Koldbæk Lemvigh

Kongens Lyngby, December 2022

Abstract

Cancer is a major global issue claiming the lives of millions every year. It is an extremely complex and diverse disease spanning numerous cancer types, however, they are all defined by one specific hallmark: abnormal cell growth. Historically, researchers have focused on the cancer itself, but it is now apparent that several other factors contribute to the disease. Cancer cells and immune cells constantly exert mutual selective pressure on each other during cancer development. Immune cells posses the ability to kill cancer cells, nevertheless cancer cells can escape this by creating immunosuppressive environments. The introduction of immunotherapies has changed the cancer treatment paradigm. Instead of using highly toxic treatment options such as chemotherapy and radiation, immunotherapies aim at unleashing the potential of immune cells to eliminate the cancer cells. Due to varying response rates, there is a strong need for patient stratification in order to determine who will benefit from treatment.

With the advances of single-cell technologies such single-cell RNA (scRNA) sequencing, dissection of heterogeneous cell systems such as the tumor microenvironment is now widely adopted. This thesis is comprised of four studies with a common goal of elucidating various aspects of the tumor microenvironment primarily employing scRNA sequencing. The first study presented highlights transcriptional differences between healthy immune cells and cancer patients with either chronic lymphocytic leukemia (CLL) or the precursor stage thereof.

Secondly, a study seeking to discover the determinants of response in Richter's syndrome (RS) patients following PD1 checkpoint blockade is presented. This showed that response associated with a $CD8^+$ T cell population marked by the expression of the transcription factor *ZNF683* that appeared to regulate key pathways of T cell activation and differentiation.

The third included in this thesis observes no transcriptional changes of the tumor cells of CLL patients going from precursor stage to disease. This is in concordance with the DNA methylation patterns presented in this study, that is detected already at the precursor stage and is persistent through the disease course. Finally, scRNA sequencing was utilized to detect clonal evolution of CLL cells transforming into an aggressive secondary lymphoma, RS. The findings of this study also includes molecular events driving this transformation.

Dansk resumé

Kræft er et stort globalt sundhedsproblem og en sygdom, der er koster flere millioner liv hvert år. Kræftsygdommen er ekstrem kompleks, divers, og dækker over mange forskellige kræftformer. Dog har de alle ét fælles karaktertræk: unkontrollerbar cellevækst. Historisk har forskere foukseret på selve kræften, men det er nu etableret, at en række andre faktorer også har indflydelse på sygdommen. Kræftceller og immunceller er i konstant kontakt og udøver selektivt pres på hinanden under sygdomsudviklingen. Kroppens eget immunforsvar har evnen til at eliminere kræften, men kræften kan undslippe dette ved at skabe et immundæmpende miljø omkring sig. Introducering af immunterapi har ændret behandlingen af kræftpatienter. I stedet for at anvende toksiske behandlinger såsom kemoterapi og strålebehandling, er målet med immunterapi at undslippe immunforsvarets egne celler til at eliminere kræften. Effektiviteten af immunterapi varierer en del, og der er derfor et stort fokus på at identificere hvilke patienter, der vil reagere positivt på behandlingen. Udviklingen af enkeltcelle-baserede teknologier, såsom enkeltcelle RNA-sekventering, har muliggjort undersøgelser af heterogene cellesystemer og bliver rutinemæssigt udført. Denne PhD afhandling indeholder fire forskellige studier med ét overodnet fælles mål at belyse forskellige aspekter af miljøet omkring kræften ved brug af enkeltcelle RNA-sekventering. Det første studie viser transkriptionelle forskelle mellem raske immunceller og kræftpatienters immunceller. Dernæst præsenteres et studdie der har til formål at undersøge mulige afgørende faktorer for et positivt respons til immunterapi med PD-1-blokade i kræftpatienter med Richter's syndrom (RS). Her viste vi, at respons er associaseret med en population af $CD8^+$ T-celler, der er makeret af ekspressionen af transkrriptionsfaktoren *ZNF683*. Studiet indikerer, at *ZNF683* regulerer essentielle "pathways" indenfor aktivering og differentiereing af T-celler. Et tredje studie observerer ingen transkripttionelle forskelle mellem kræftceller fra patienter med kronisk lymfatisk leukæmi samt forstadiet hertil. Det er i overensstemmelse med DNA methylering, der også er præsenteret i dette studie. Abnormal DNA methylering er dektekteret allerede på forstadiet og vedbliver gennem sygdomsforløbet. Slutteligt, anvendes enkeltcelle RNA-sekventering til at undersøge mekanismerne bag transformationen fra kronisk lymfatisk leukæmi til den aggresive og sekundære lymfom, RS.

Acknowledgements

The work presented in this thesis could not have been done without invaluable guidance and support throughout the last couple of years. Firstly, I would like to thank my two co-supervisors Lars Rønn Olsen, DTU and Catherine J. Wu, Wu lab, Dana-Farber Cancer Institute. Lars has both challenged me and supported my research throughout this thesis. I would like to thank him for trusting me to do this work and always encouraging me. Your never-ending optimism and passion for science is a true inspiration. Thank you for teaching me what good coffee should taste like. Catherine welcomed me into her lab at Dana-Farber Cancer Institute in Boston as a young master student back in 2018. Ever since she had full trust in my capabilities, allowing me to analyze high impact data and always made sure to credit me whenever it fitted. I really enjoyed visiting you multiple times throughout the years. You made sure to push me out of my comfort zone confident that I would be successful. Your passion and knowledge is inspiring and translates to everyone around you. Secondly, I would like to thank Erin M. Parry, an ambitious and extremely talented fellow from Wu lab, who has been my partner in one of my larger projects. We met back in 2018, and she always believed in me and my capabilities - even when I did not. We had many great (and also confusing) times both physically, but also online. You were always available for a talk, even if it was just five minutes.

Through my collaboration with the Wu lab, I have been fortunate to work with several skilled researchers. Noelia Purroy and Satyen H. Gohil deserves a thanks for our collaborations and productive discussions. Thanks to Peter Kharcenko for guidance and introducing me to the analysis of single-cell RNA sequencing. I would also like to thank my colleagues from the Single Cell Omics group at DTU, Christina Bligaard Pedersen, Giorgia Moranzoni, Søren Helweg Dam, Kristoffer Vitting-Seerup, and Nanna Møller Barnknob along with my office peers Clara Drachmann and Trine Zachariasen for good times, scientific debates and pep talks.

A thanks also goes to my friends and family for their love and support in this process. Lastly, thanks to my husband, Magnus, who's always been my biggest supporter and for always being patient with me. And to my daughter, Olivia, who always puts a smile on my face. I could not have done it without you.

Table of Contents

Research papers

Research papers included in the thesis

PAPER I: Purroy, N. Z.*, Tong, Y. E.*, **Lemvigh, C. K.***, Cieri, N., Li, S., Parry, E. M., Zhang, W., Rassenti, L. Z., Kipps, T. J., Slager, S. L., Kay, N. E., Lesnick, C., Shanafelt, T. D., Ghia, P., Scarfò, L., Livak, K. J., Kharchenko, P. V., Neuberg, D., Olsen, L. R., Fan, J., Gohil, S. H., Wu, C. J. Single cell analysis reveals immune dysfunction from the earliest stages of CLL that can be reversed by ibrutinib. *Blood 2022, https://doi.org/10.1182/blood.2021013926*

PAPER II: Parry, E. M.*, **Lemvigh, C. K.***, Deng, S., Dangle, N., Ruthen, N., Knisbacher, B. A., Broséus, J., Hergalant, S., Guièze, R., Li, S., Zhang, W., Long, J., Yin, S., Werner, L., Anandappa, A., Purroy, N. Z., Gohil, S. H., Oliveira, G., Bachireddy, P., Shukla, S. A., Huang, T., Livak, K. J., Gad Getz, Neuberg, D., Feugier, P., Kharchenko, P., Wierda, W., Olsen, L. R., Jain, N., Wu, C. J. ZNF683 marks a CD8+ T cell population associated with anti-tumor immunity following anti-PD-1 therapy for Richter syndrome. *Manuscript in review, Cancer Cell*

PAPER III: Kretzmer, H., Biran, A.*, Purroy, N. Z.*, **Lemvigh, C. K.***, Clement, K*., Gruber, M.*, Gu, H., Rassenti, L., Mohammad, A. W., Lesnick, C., Slager, S. L., Braggio, E., Shanafelt,

T. D., Kay, N. E., Fernandes, S. M., Brown, J. R., Wang, L., Li, S., Livak, K. J., Neuberg, D. S., Klages, S., Timmermann, B., Kipps. T. J., Campo, E., Gnirke, A., Wu, C. J.**, Meissner, A.** (2021). Preneoplastic Alterations Define CLL DNA Methylome and Persist through Disease Progression and Therapy. *Blood cancer discovery, 2(1), 54–69. https://doi.org/10.1158/2643-3230.BCD-19-0058*

PAPER IV: Parry, E. M.*, Leshchiner, I.*, Guièze, R.*, Johnson, C., Tausch, E., Parikh, S. A., **Lemvigh, C. K.**, Broséus, J., Hergalant, S., Messer, C., Utro, F., Levovitz, C., Rhrissorrakrai, K., Li, L., Rosebrock, D., Yin, S., Deng, S., Slowik, K., Jacobs, R., Huang, T., Li, S., Fell, G., Redd, R., Lin, Z., Knisbacher, B. A., Livitz, D., Schneider, C., Ruthen, N., Elagina, L., Taylor-Weiner, A., Persaud, B., Martinez, A., Fernandes, S. M., Purroy, N. Z., Anandappa, A., Ma, J., Hess, J., Rassenti, L. Z., Kipps, T. J., Jain, N., Wierda, W., Cymbalista, F., Feugier, P., Kay, N. E., Livak, K. J., Danysh, B. P., Stewart, C., Neuberg, D., Davids, M. S., Brown, J. R., Parida, L., Stilgenbauer, S.**, Getz, G.**, Wu, C. J.**. Evolutionary history of transformation from chronic lymphocytic leukemia to Richter syndrome. *Accepted, Nature Medicine*

*,** = equal contribution

Research papers not included the thesis

PAPER 1: Lee, P. C., Klaeger, S.*, Le, P. M.*, Korthauer, K.*, Cheng, J., Ananthapadmanabhan, V., Frost, T. C., Wong, A. Y., Iorgulescu, J. B., Tarren, A., Chea, V. A., Carulli, I. P., **Lemvigh, C. K.**, Pedersen, C. B., Sarkizova, S., Wright, K. T., Li, L. W., Nomburg, J., Li, S., Huang, T., Liu, X., Pomerance, L., Doherty, L. M., Apffel, A., Wallace, L., Rachimi, S., Felt, K. D., Wolff, J., Witten, E., Zhang, W., Neuberg, D., Zhang, G., Olsen, L. R., Thakuria, M., Rodig, S. J., Clauser, K. R., Starrett, G. J., Doench, J. G., Buhrlage, S. J., Carr, S. A., DeCaprio, J. A.**, Wu, C. J.**, Keskin, D. B.**. Reversal of viral and epigenetic HLA class I repression in Merkel cell carcinoma. *Journal of Clinical Investigation*

PAPER 2: Moranzoni, G., **Lemvigh, C. K.**, Mantas, P., Barnkob, N. M., Pedersen, C. B., Jessen, L. E., Kladis, G., Gohil, S. H., Vitting-Seerup, K., Barnkob, M. B.*, Olsen, L. R.*. Systematic computational evaluation of chimeric antigen receptor therapy targets. *Manuscript in preparation*

PAPER 3: Treon, S. P., Kotton, C. N., Park, D., Gathe, J. C., Varughese, T., Barnett, C. F., Belenchia, J. M., Clark, N. M., Farber, C. M., Olsen, L. R., Moranzoni, G., **Lemvigh, C. K.**, Keskin, D. B., Wu, C. J., Patterson, C., Guerrera, M. L., Hunter, Z. R., Soumerai, J., Chea, V., Carulli, I., Southard, J., Li, S., Livak, K. J., Holmgren, E., Kim, P., Shi, C., Lin, H., Ramakrishnan, V., Olszewski, S., Tankersley, C., Zimmerman, T., Dhakal, B. A Randomized, Placebo-Controlled Trial of Zanubrutinib in Hospitalized Patients With COVID-19 Respiratory Distress: Biomarker and Clinical Findings. *Submitted, JAMA Network Open*

*,** = equal contribution

1

Introduction

Cancer is the second leading cause of death worldwide responsible for almost 10 million deaths in 2020 [1,2]. The disease has an increasing incidence as the population ages. Every third Danish person will be diagnosed with some type of cancer before they reach the age of 75 years old, and two out of three will be close to somebody with cancer. However, with growing research and treatment options, prognosis for cancer patients is improving with about 67% surviving cancer [3], although some cancer types have better prognoses than others.

Cancer is an umbrella term covering more than 100 distinct cancer types, spanning both solid and liquid tumors. Due to heterogeneity within a single tumor, the understanding of cancer becomes even more complicated.

The immune system presents itself as a promising resource to utilize in killing cancer cells. However, cancer cells can evolve to evade immune killing making it a complex interplay. To heighten our understanding of cancer, it is therefore crucial to analyze both malignant and im-

mune cells and how they communicate, interact and co-evolve. It is becoming increasingly appreciated to profile tumor microenvironments using single-cell expression data. Accordingly, computational tools to robustly analyze this complex high-dimensional data is of outmost importance.

Thesis scope

The focus of this thesis is the bioinformatic analysis of data pertaining to the complex interplay between cancer cells, immune cells and and understanding how these interactions shift in the context of disease evolution and therapy primarily using single-cell transcriptomics data. Both computational analyses and biological interpretations of this will be included.
This provided a deeper understanding of the cancer cells in relation to immune cells, disease progression and during immunotherapy, while also covering the computational workflow of data analysis. The thesis includes a discussion of how to integrate data for increased knowledge extraction.

Thesis structure

First, I start out by setting the theoretical background necessary for the presented work by introducing cancer, the immune system, immunotherapies along with response predictions. Subsequently, a general introduction of cancer profiling using expression data, focusing on single-cell levels, and how to analyse this is provided. Following, are three chapters covering various facets of cancer and the tumor microenvironment:

1. **Single-cell profiling of immune cells in chronic lymphocytic leukemia**: Cancer cells co-evolve with immune cells in their environment and they mutually exert selective pressure on each other. In PAPER I, we studied how the immune cells of patients diagnosed with chronic lymphocytic leukemia evolve with disease progression, and the consequence of treatment regarding cell-cell communication patterns.

2. **Single-cell profiling of immune cells in Richter's syndrome during immunotherapy**: Richter's syndrome is the transformation of chronic lymphocytic leukemia into an ag-

gressive secondary lymphoma with poor prognosis and treatment options. Clinical trials are showing how these patients are responding to immmunotherapies, e.g. PD-1 checkpoint blockade. In PAPER II, we interrogated single-cell expression data of responders and non-responders to detect markers of response.

3. **Single-cell profiling of blood cancers**: The disease spectrum of chronic lymphocytic leukemia spans both a precursor stage, disease stage and sometimes a transformation into Richter's syndrome. In PAPER III we showed how the cancer cells evolve during the natural progression from precursor stage to chronic lymphocytic leukemia. Molecular characteristics defining disease transition into RS was highlighted in PAPER IV.

Then I will present a conclusion summarizing the work carried out in this thesis and how it sets in relation to future work. Finally, the four research papers included in this thesis is attached.

2

Theoretical background

2.1 Cancer is a complex disease

The key characteristic of cancer is the presence of cells growing uncontrolled. Cancer cells arise from the body's healthy cells through mutational changes enabling this abnormal growth. Abnormal growth is achieved through a set of characterized features, namely the *hallmarks of cancer*. As we heighten our understanding of cancer biology, the features also expand. In 2000, the first set hallmarks consisting of six biological properties was proposed [4]. These were further expanded to eight hallmarks with an addition of two enabling characteristics [5]. Hallmarks of cancer circa 2022 now includes 10 hallmarks and four enabling characteristics [6]. Briefly, the hallmarks include sustaining proliferative signaling, evading growth suppressors, avoiding immune destruction, enabling replicative immortality, tumor-promoting inflammation, activating invasion and metastasis, inducing or accessing vasculature, genome instability and mutation,

resisting cell death (apoptosis), deregulating cellular metabolism, unlocking phenotypic plasticity, non-mutational epigenetic reprogramming, senescent cells and polymorphic microbiomes. One that is especially interesting in the context of the work presented in this thesis is the ability to avoid immune destruction.

A cancer cell is an experiment in evolution, and it evolves through acquisition of mutations making them quite different from healthy cells [7, 8]. During cell proliferation, the genome is copied, and in this process mistakes can be introduced. Most often these will be corrected and have no effect, but it is not always the case. When errors are not corrected they are passed on to the progeny. The majority of mutations are harmless, but a subset leaves the cell one step closer to becoming cancerous [9]. A single mutation is rarely adequate, and some might be introduced by extrinsic factors [10] such as tobacco and alcohol [7, 11, 12] and UV light [13, 14]. Mutations can ultimately lead to altered gene products, which can be apparent at the RNA level [15].

For a long time, the focus has primarily been on the cancer cell itself. A tumor is not just a collection of cancerous cells, but rather a complex heterogeneous system [16]. Cancer cells on the outer side of a solid tumor may differ considerably from cells on the inside, as these are the cells that are in contact with the microenvironment surrounding the tumor. Transcriptional heterogeneity within the tumor is increasingly recognized as a driver of progression, metastasis and treatment outcome [17]. Several studies have shown an, at first surprisingly, high variation in transcriptomic profiles of cancer cells within the same tumor [17]. Also, it is becoming more appreciated that the non-cancerous cells, such as immune cells, in the tumor microenvironment (TME) play a vital role [18].

2.2 The immune system helps protect our body

Our body is constantly exposed to potential infectious agents and toxins, so why are we not sick all the time? The immune system is our body's defense against foreign matters that detects non-self from self and eliminates it. It is a highly complex system consisting of a variety of specialized effector cells and molecules [19, 20]. Overall, the immune system is categorized into the innate and the adaptive immune system, both of which depend on the activity of white blood cells, namely leukocytes.

The majority of immune cells arise from the bone marrow, where they can also develop and mature. Mature immune cells can either occupy peripheral tissues or circulate the blood stream or the lymphatic system. All leukocytes as well as the other components of blood including red blood cells (erythrocytes) and platelets, derive from hematopoietic stem cells (HSCs) from the bone marrow. HSCs divide into two types of stem cells; a common lymphoid and a common myeloid progenitor cell. The lymphoid progenitor gives rise to the lymphoid lineage of leukocytes; innate lymphoid cells, natural killer (NK) cells and T- and B cells. The myeloid progenitor gives rise to the rest of the leukocytes. Among the myeloid leukocytes are dendritic cells (DCs), macrophages, neutrophils, and mast cells. Figure 2.1 provides a brief overview of the how the cells arise from the HSCs in the bone marrow.

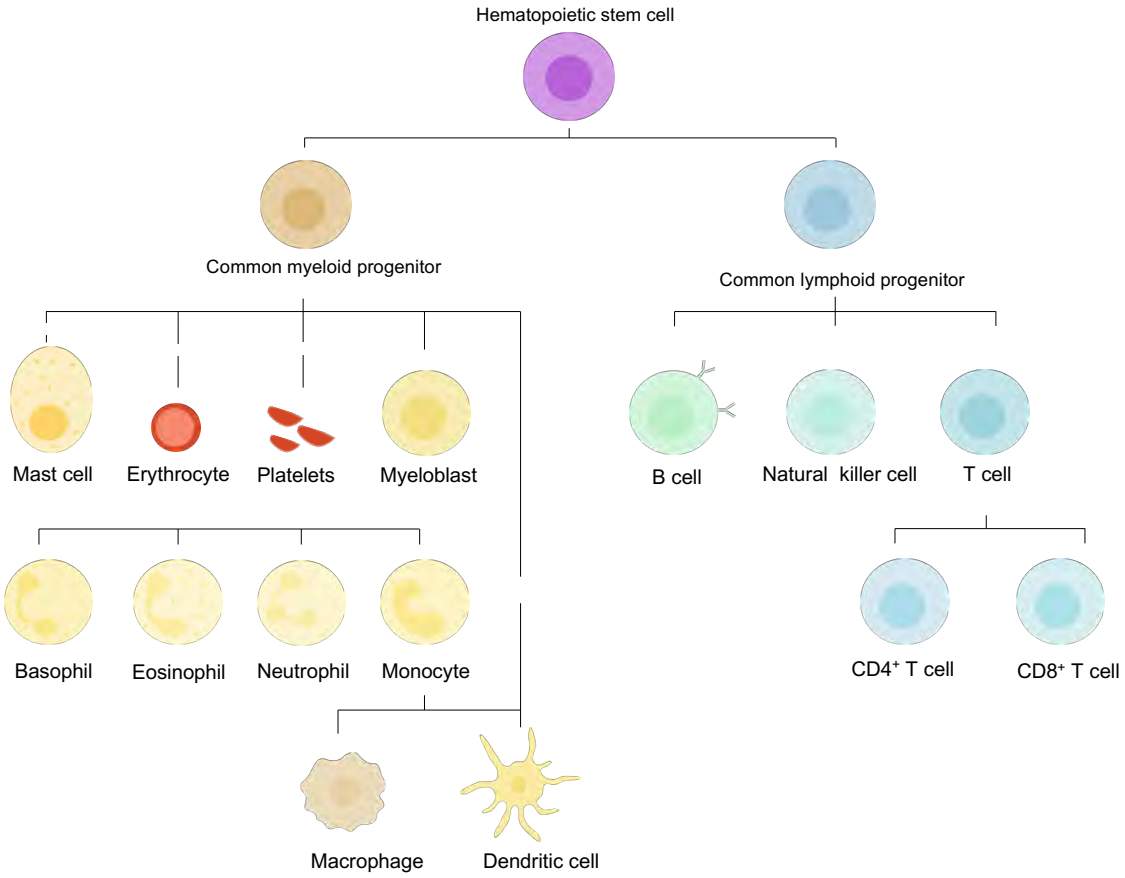


Figure 2.1 | Overview of the immune cells derived from hematopoietic stem cells of the bone marrow with cells split into the lymphoid and myeloid lineages. Some intermediate cell types have been omitted represented by broken lines. Figure adapted from [19, 20]

Innate immunity

The innate immune system is responsible for the immediate response to a wide range of pathogens. If a pathogen succeeds in breaching both anatomical (epithelial surfaces such as the skin) and chemical barriers within the host, it will encounter the components of the innate immune system. Once detected, inflammation is induced by the release of cytokines and chemokines attracting additional immune cells.

Innate cells are activated by pattern recognition receptors (PRRs) that detect typical features of microbes, such as lipopolysaccharides. When a microorganism crosses the epithelial barrier and starts reproducing within host tissue, it will likely be recognized by resident phagocytic cells. Phagocytes encompasses macrophages, monocytes, granolucytes and DCs with macrophages being the dominant population in most normal tissues. They arise from either progenitor cells or from circulating monocytes exiting the blood stream. Monocytes develop in the bone marrow and circulate the bloodstream, and is typically classified as either classical or non-classical monocytes.

Innate lymphoid cells (ILCs) including NK cells develop in bone marrow from the same progenitor that develops B- and T cells. As effector cells they amplify signals by innate recognition, and are stimulated by cytokines produced by other innate cells. NK cells lack the antigen-specific receptors of T cells, but can exhibit an equivalent cytotoxic capacity. They are found in tissues and circulating the blood stream. Additionally, the NK cells will contain an infection while the adaptive immune response becomes activated [19].

Adaptive immunity

The adaptive immune response relies on specific recognition using highly specialized cell-surface receptors, and a key characteristic is that it provides immunological memory. The adaptive immunity is mainly dominated by two cell types: B- and T cells. Both of these harbour extremely variable antigen-specific receptors, and differentiate from a state of naïve cells to effector cells upon antigen-recognition. These receptors are generated by somatic rearrangement to provide them with a unique antigen-specificity. Activated B cells, plasma cells, secrete antibodies, and T cells will differentiate into various effector cells. A subset of activated cells, will differentiate into memory populations responsible for the long-term memory. These can be

reactivated by a following antigen exposure and differentiate into effector cells in order to clear the infection more rapidly. This specificity and immunological memory is why the adaptive immune response is slower, but also more powerful [20].

Two main types of T cells exist, characterized by the expression of the co-receptors: CD4 and CD8. Through their T cell receptor (TCR), they recognize fragments of antigens, namely epitopes, bound by major histocompability complexes (MHC) on the surface of other cells. $CD8^+$ T cells recognize epitopes bound to MHC class I, which is present on all nucleated cells. MHC class I presents intracellular peptides, and thereby $CD8^+$ T cells can detect intracellular pathogens. When detecting a foreign epitope, the $CD8^+$ T cell will become cytotoxic and kill the infected cells [20]. Contrary, $CD4^+$ T cells interact with epitopes bound to MHC class II, which are only found on antigen-presenting cells (APCs). Peptides presented on MHC class II are usually derived from the extracellular environment, which primes the $CD4^+$ T cells to become either helper T cells or regulatory T cells (T_{regs}). Helper T cells recruit and activate other immune cells such as B cells via cytokines to produce antibodies or induce macrophage killing. T_{regs} dampen the activity of other lymphocytes and help limit any potential damage of the immune response and maintain self-tolerance. They inhibit the extend of the immune activation and promote tolerance rather than clearing of an antigen. Initial activation of the $CD8^+$ T cells is depending on signals from innate cells, as it occurs through presentation by DCs. If the DC is not infected itself, it will take up peptides generated by other infected cells through a process named cross-presentation [21]. Once activated, the T cells induce apoptosis of infected cells presenting peptide-MHC complexes complimentary to their TCR. In order to ensure activation, interaction between the co-stimularoty molecule, CD28, on the T cell and B7 ligands on the APC must occur. Following, the T cells will increase expression of inhibitory molecules, such as CTLA-4 and PD-1, with higher affinity for B7 ligands. This will eventually stop the activation, and is an important feature of tolerance [22, 23].

T cells continuously exposed to antigens or inflammatory signals in chronic infections or cancers can become exhausted leading to progressive loss of effector functions, and expression of multiple inhibitory receptors (PD-1, CTLA-4, LAG3, TIM3 and TIGIT), dysregulated metabolism, poor memory response and homeostatic proliferation [24, 25]. Exhaustion is associated with inefficient control of chronic infections or cancers. A subset is also referred to as being terminally

exhausted, which are resistant to treatment [26].

Immune responses can both be beneficial but also extremely harmful, depending on the antigen. Allergies and autoimmune diseases are examples of a normal response directed at an inappropriate antigen. Cancer is an example of a deficient immune response against an appropriate antigen.

The role of immune cells in cancer

Immune cells play a crucial role in immunosurveillance, as both adaptive and innate immune cells infiltrate the TME [27], which plays a critical role in tumorigenesis, metastasis and drug resistance [28]. The relationship between immune cells and cancer cells is extremely complex, but three phases of tumor growth have been proposed: 1) elimination phase, where immune cells destroy potential cancer cells (immunosurveillance), 2) equilibrium phase, where cancer cells undergo changes that promote survival due to selection pressure by the immune system, and 3) escape phase in which cancer cells have acquired abilities to escape immune killing [29, 30]. Immunosurveillance is generally considered to be the undetectable phase of early cancer development [31]. During the second phase, cancer immunoediting takes place where the properties of cancer cells are continuously shaped to promote survival. Here, cancer cells continue to coexist with the immune cells.

One of the hallmarks of cancer as previously mentioned is immune evasion by creating an immunosuppressive environment [30,32]. There are several avenues to achieve this: low immunogenicity (no peptide-MHC complex, no adhesion molecules or no co-stimulatory molecules), treating the tumor as self (tumor peptides presented on MHC in the absence of co-stimulation), antigenic modulation (loosing epitopes that will be recognized by TCRs), immune suppression (secreting factors that inhibit immune cells directly or expressing them on the surface of the tumor, e.g. PD-L1) and creating a tumor-induced privileged site (secreting factors that will act as a physical barrier against the immune cells, such as collagen).

The down-regulation of MHC molecules on cancer cells can be recognized by NK cells. One important activator of NK cells is the loss of MHC expression, thus making cancer cells a target. Also, NK cells produce cytokines and chemokines that recruit DCs and influence T cell responses during inflammation [33]. Therefore, NK cells are useful targets to activate in thera-

pies [34].

Tumor-specific antigens arising from point mutations or gene rearrangements, may alter the epitope presented on the MHC molecule preserving the binding affinity, or *de novo* proteins will now bind and be presented. These are referred to as neoepitopes, newly immunogenic versions of normal peptides. Both may elicit an immune response by recognition by T cells.

A successful anti-tumor response requires the presence of both adaptive $CD8^+$ T cells, $CD4^+$ T cells, B cells along with innate lymphoid cells, e.g. NK cells [35, 36]. Cytotoxic $CD8^+$ T cells are highly responsible for anti-tumor activities and are associated with improved prognosis in virtually all cancers [35, 37]. $CD4^+$ T helper cells promote $CD8^+$ T cells and are crucial for B cell activation, expansion and differentiation into plasma cells and memory B cells [35, 38]. Historically, the main focus of anti-tumor activity has relied on $CD8^+$ T cell responses, as presence of these are strongly associated with higher survival. However, recent work highlights a key role of B cells in immunotherapies and a positive association with presence and prognosis has been observed in several cancer types [39–42]. B cells can act as APCs and activate $CD8^+$ T cells, but can also kill the cancer cells directly by secretion of toxic cytokines or indirectly by secreting immunostimulatory cytokines. Plasma cells produce antibodies that mediate antibody-dependent cell-mediated cytotoxicity targeting cancer cells for phagocytosis by NK cells [35]. Contrary, B cells also harbor tumor-promoting effects with regulatory B cells inhibition of effector functions and converting $CD4^+$ T cells into T_{regs}. Likewise, T_{regs} counteract the tumor-specific response by suppressing $CD8^+$ T cells, and are thus also considered to be tumor-promoting. The composition of tumor-infiltrating lymphocytes (TILs) present in the TME has shown to impact clinical outcomes [35].

2.3 Immunotherapies to combat cancer

Traditional cancer therapies include surgery, chemotherapy, and radiation and have generally focused on the tumor itself [43]. However effective, these treatments suffer from severe side effects, as they also harm the non-cancerous cells of the body. The current treatment paradigm has shifted towards harnessing the immune system.

Introduction of immunotherapy has revolutionized cancer treatments showing responses seen in up to 50% of patients with long-term response [44], with adoptive cell transfer and immune checkpoint blockade inhibitors resulting in durable clinical responses. However, efficacy's vary and not all patients respond [27, 45].

As eluded to previously, cancers can take advantage of the immune system's built-in checkpoint that normally dampens the immune responses, preventing autoimmune diseases and allergies, thus modulating the intensity of an immune response [46]. Checkpoint blockade inhibitors therapies aim to revert this by using monoclonal antibodies to down-regulate these inhibitory molecules. Many of these are targeting the reactivation of exhausted T cells in tumors [26, 47]. However, as described the anti-tumor response is complex and involves both innate and adaptive immune cells. Therefore, the innate cells are promising new targets as well [36, 48].

The introduction of checkpoint blockade inhibitors, anti-CTLA-4 and anti-PD-1 in particular, changed the management of many cancers. In advanced melanoma this means we have reached long-term control and tumor regression in about 50% of patients, which was around 10% before [49]. The checkpoint receptors CTLA-4 and PD-1 have been studied extensively. PD-1 is responsible for limiting T cell activity during an inflammatory response. Expression of PD-1 is induced on antigen-stimulated T cells, B cells, NK cells and a subset of DCs, and after interaction with the cognate ligand, PDLs, expressed by tumor cells (PD-L1 or PD-L2), decreased T cell activation and proliferation and eventually exhaustion is observed. PD-1 expression on TILs is associated with poor prognosis, and blocking PD-1 with an antibody will restore the activity of the inhibited immune cells. A PD-1 blockade can thus revert the exhausted state in immune cells present in the TME. Understanding the TME is important for effective therapy [50]. Blockade of PD-1 or PD-L1 has shown great clinical promise [51]. Three PD-1-antibodies, and three anti-PD-L1 antibodies are currently approved by the Food and Drug Administration [52]. Mechanisms of immune escape following immunotherapy is categorized as either primary, adaptive and acquired resistance. Patients with primary resistance do not respond to treatment, where acquired resistance sets in after an initial response leading to disease progression. In adaptive immune resistance, the cancer is recognized but not eliminated through antigenic escape and/or loss of immunogenicity [43, 44].

2.4 Characteristics of response

A critical part of expanding the implementation of immunotherapies is dissecting the features that might be predictive or prognostic of clinical response. To better identify patients benefiting from treatment, samples are usually collected at diagnosis, time of therapy response (or resistance), remission and relapse [53] in order to detect novel biomarkers of response using proteomic, genomic, and transcriptomic analyses [52]. Biomarkers predicting responders are highly desirable, as are biomarkers predicting immune-related adverse effects that can require treatment discontinuation. Insights into host intrinsic and extrinsic factors impacting the response and toxicity are needed [54], and our understanding of response and resistance to treatment is continuously evolving.

Overall, neoantigen load is thought to be a major biomarker of response in cancer immunotherapy [55]. A positive association between tumor mutational burden (TMB) and response across 27 cancer types has been shown [56]. Higher TMB in in non-small cell lung carcinoma is associated with improved response, durable clinical benefit, and progression-free survival [57, 58]. In melanoma increased TMB is likewise found in patients with durable responses to CTLA-4 antibodies [59, 60]. BRCA2 mutations are enriched in melanoma patients responding to anti-PD-1, however, the association is not predictive [61].

Expression of PD-L1 on cancer cells in melanoma, non-small cell lung carcinoma, renal cell carcinoma, colorectal carcinoma, or castration-resistant prostate cancer positively associated with response to anti-PD-1 therapy [62]. Also, IFN-γ predicts clinical response to anti-PD-1 in melanoma [63]. A study of 48 tumor samples from melanoma patients treated with checkpoint inhibitors, showed that expression of the transcription factor *TCF7* in $CD8^+$ T cells was predictive of positive clinical outcome [64], while resistance has been associated to MHC loss [65]. Additionally, a treatment signature score consisting of a set of 91 genes was proposed in lung cancer [66] in order to predict response. In colorectal cancer, mismatch repair deficiency has been associated with better response to anti-PD-1 [67].

Intratumoral microbes as well as gut microbiota has shown to impact anti-tumor responses as well. The gut microbiome of melanoma patients modulates response to anti-PD-1 [68] showing higher alpha diversity and relative abundance of certain bacteria (the *Ruminococcaceae* family).

Another study also emphasized a significant correlation between commensal microbiota and clinical response in metastatic melanoma patients highlighting abundance differences between responders and non-responders [69]. A third study showed correlation between gut microbiota and clinical response in epithelial tumors [68].

2.5 Profiling tumor microenvironments

Gene expression profiling is deciphering the total mRNA or protein levels in a cell or tissue [70, 71]. The transcriptome is the complete set of transcripts and their abundances [72] acting as a link between phenotypes and genotypes. Transcriptomic analyses are the studies of RNA molecules using high-throughput technologies [73], and has been used extensively for functional characterization of tumors [71, 74]. It is also used to test for biological differences in expression under different conditions, such as healthy versus disease or treatment versus no treatment.

Actively transcribed RNA is highly dynamic meaning that it provides a signature or snapshot of cell or tissue states. Transcriptome profiling has shown to be better for understanding molecular mechanisms behind cancer prognosis and drug resistance [73]. As a consequence of both genomic mutations and epigenetic changes, cancer cells show aberrant transcriptional patterns and the abnormal cancer pathways can be identified [73].

RNA sequencing has been around for more than a decade, and has been a paramount tool in transcriptomic analysis of differential gene expression and alternative splicing of mRNA [75, 76]. It is valuable for understanding dynamics of transcriptomics during development, comparisons between healthy and diseased tissue along with classification of disease states. RNA sequencing provides a precise quantitative measurement of transcripts and isoforms [72]. The technology has enabled characterization of cancer heterogeneity, evolution, drug resistance, biomarker discovery providing invaluable insights of cancer research and treatment [73, 76].

However, one drawback of RNA sequencing of bulk tissues is the assumption of homogeneity, thus ignoring cellular diversity by providing averaged data [77, 78]. This is particularly problematic for blood, which consists of multiple different cell types with distinct genetic programs, present in varying abundances.

Single-cell RNA (scRNA) sequencing was highlighted as "Method of the Year" in 2013 [79] and it enables transcriptomic profiling at single-cell resolution and has shown immense potential in cancer research [53, 78, 80–82]. It is also valuable for dissecting the immune system [83–85] and during immunotherapy [86]. The overall workflow of scRNA sequencing is similar to that of bulk RNA sequencing: 1) single-cell capture, 2) mRNA reverse transcription, 3) cDNA amplification and 4) library preparation [83, 87]. The two technologies differ in how the input is prepared. That is isolation of single cells in scRNA sequencing. Several methods to do so exists; valves, droplets and nanowells, with droplet-based technologies being commonly used [77, 88]. Droplet-based methods offers high-throughput, however with the cost of less control [88], fewer features and limit studies to gene-level only. Here the single cells are encapsulated in oil beads along with reagents and barcodes. Three main platforms use this technique: 10X Genomics Chromium, inDrop and DropSeq [89–92]. Each individual cell is tagged to a bead with a unique barcode. Each mRNA transcript is also tagged with a unique molecular identifier (UMI) to account for potential PCR biases [93]. The output is a matrix of the absolute counts for each transcript in a given cell.

Protocols also differs in how the sequencing libraries are constructed: i) full-length or ii) tag-based. Tags can be added to either the 5' or 3' end of the transcript allowing to be combined with UMIs. The Smart-seq2 [94] and Smart-seq3 [95] protocols can give full-length transcriptome profiles of single-cells.

10X Genomics uses the GemCode^{TM} technology for cell partitioning, where single cells are encapsulated in Gel Bead in Emulsion (GEMs) with enzymes. Each bead has their own barcode, and each transcript labelled with an UMI. Cell lysis and reverse transcription happens in each GEM. All cDNA is then pooled and amplified for short-read sequencing [90].

As described, tumors are complex biological systems that requires comprehensive analysis. So, no single approach can elucidate the entire tumor development and behavior, requiring multiple levels of information such as genomic, transcriptomic, proteomics and epigenetics [74]. Integrated analysis of these different molecular features, or modalities, is essential to creating a comprehensive overview of the disease.

Most single-cell technologies require different cells as an input to measure different modalities, which can be challenging when trying to computationally integrate the data [96]. A rising number of platforms do offer multimodal measurements at single-cell resolution.

Reverting to 10X Genomics, they offer paired full-length TCR (or BCR) sequences coupled with 5' gene expression, thus allowing identification of the T and B cell repertoire, detection of expanded clones and linking clones to phenotype. Furthermore, it is possible to simultaneously profile gene expression and open chromatin stretches within a given cell (single-cell ATAC sequencing). Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) [97] is another multimodal technology that allows for coupling gene expression with surface protein expression for a given cell. CITE-seq uses barcoded antibodies to integrate protein expression with transcriptomic measurements and is compatible with existing scRNA sequencing approaches such as 10X Genomics [97]. With scNMT-seq [98] joint chromatin accessibility, gene expression and DNA methylation can be analyzed. One disadvantage of scRNA sequencing is the necessary dissociation of tissues missing the spatial information and information on cell proximities [99]. Spatial transcriptomics [100] is an emerging technology that allows for profiling gene expression of a tissue sample while preserving spatial information [80, 101], which can provide an in-depth analysis of the TME [28, 99, 102]. This will help us interrogating how the tumor cells communicate [80].

The increase in multimodal platforms, demands an increase in computational methods capable of analyzing such data. However, all these modalities will help gain a better understanding of complex environments such as the TME, and will be crucial in personalized medicine of cancer patients [103].

2.6 Computational tools for transcriptomic analysis

Analysis of RNA sequencing data can broadly be separated into two overall workflows. The first, and probably most used historically, consist of aligning to reference genome, e.g. using STAR [104], followed by gene quantification, normalization and differential gene expression (DGE) testing, using DESeq2 [105] or edgeR [106], or differentially expressed transcripts. However, a secondary approach is becoming widely adopted. This approach utilizes alignment-

free tools such as kallisto [107] and Salmon [108] for rapidly quantifying abundances of transcripts by the use of *pseudoalignments*, again followed by normalization and DGE or differentially expressed transcripts.

As mentioned, bulk RNA sequencing dilutes signals from smaller population by averaging the expression across a sample. scRNA sequencing data have shown to be an indispensable tool in dissecting heterogeneous cell systems and tissues and *de novo* discoveries [87, 109]. However, the data is associated with certain caveats. Due to the lower amount of starting material and lower sequencing depth, the mRNA yield resulting from a scRNA sequencing run is lower comparing to what is obtained with bulk RNA sequencing. Some genes will fail to be detected, even if they are expressed, and the resulting expression matrix will contain a large number of zeroes, ranging from 50% up to 99% [110]. This is very much in contrast to bulk RNA sequencing, where the number is around 20% (or less) [110]. This data sparsity is both due to technical and biological noise [87].

Raw sequencing reads from 10X Genomics are usually processed with their CellRanger pipeline [90] consisting of: demultiplexing, alignment, filtering followed by barcode and UMI counting generating a feature-cell count matrix. Once the expression matrix is obtained, quality control is a necessary step to exclude poor quality data and ensure to include viable cells only. Quality control is often based on three measurements. The first two are the number of detected genes (library complexity) and transcripts per cell (library size) [111, 112]. The third is removal of dying cells, which will often display a high proportion of mRNA from the mitochondrial genome [113]. There are no golden standards for threshold settings of these, but is often depending on the cell types being analyzed. It can be done by detection of outliers.

Doublets, or even multiplets, can be another contaminating factor, often approximately 5% (75). These can be simulated and filtered using tools such as Scrublet [114], DoubletDecon [115] and DoubletFinder [116]. DoubletFinder has shown to provide highest accuracy detection [117]. Cleaned data will then be normalized to account for varying sequencing depths either by simple count depth or by scRNA sequencing specific models that also account for data sparsity, e.g. scTransform [118]. Another possible confounder to consider is proliferating cells that will show enrichment of upregulated genes related to the cell cycle. To compensate for this, one can either

remove those cells or regress out their signal using general linear models [110]. Other gene sets accounting for unwanted source of variation may also be detected.

scRNA sequencing data is high dimensional, but many of the genes expressed in a cell are house-keeping genes with low variability between cells, and many genes will have zero counts in a dataset. Therefore, in order to reduce the dimensionality a feature selection will often follow to detect the most variable genes in the dataset and these will be used for downstream analysis [112].

Batch effects is another major concern when analyzing multiple samples. Several methods exist to integrate and analyze scRNA sequencing data and thus overcoming these batch effects. Broadly, they can be separated into two categories: supervised (e.g. reference-based) and unsupervised [119]. Reference-based methods are usually faster; however, it does require a good representative reference, which can be challenging when seeking novel cell populations. Unsupervised methods will often perform integration two datasets at a time.

Following will be two rounds of dimensionality reduction, first using principal component analysis (PCA) and then Uniform manifold approximation and projection (UMAP) [120] (or t-SNE [121] or largeVis [122] reductions) for visualizations. This is then followed by clustering of the cells. Clusters have no biological insight themselves, so the next step is to annotate them into cell populations, which can be a daunting and time-consuming task as many populations can be further divided into subpopulations. Also, many exist as continuous rather than distinct populations. But it can be done in two ways: 1) manual annotation by the use of marker genes or 2) mapping to previously annotated reference datasets, either bulk or single-cell, using Seurat [123, 124] or singleR [125]. Finally, differential expression and abundance is usually performed. Conventional DGE methods developed for bulk RNA sequencing, such as DESeq2, are in current use and perform well, however, methods accounting for the characteristics of scRNA sequencing data have been developed [126–128].

scRNA sequencing provides a snapshot of a cell's gene expression, yet cell states are not static, but rather continuous during cell differentiation processes. Adding another layer of information, scRNA sequencing can be used to estimate cell differentiation and lineage tracing [129]. Modeling of this continuum is referred to as trajectory inference or pseudotime mapping.

Contrary to clustering, where each cell is assigned a categorical label, trajectory analyses aim

at labeling cells with continuous values, namely pseudotime [130]. The term 'pseudo-time' was first introduced with Monocle [131]. Pseudotime values quantify a cell's position along a given trajectory, and how far the cell is from a given progenitor state. A plethora of methods for trajectory inferences from scNRA sequencing data are being published with many currently being developed [131–135]. Inferred trajectories can either be linear or a non-linear branching, where most differentiation process includes branching [134], but novel tools also infer cyclic trajectories [136] or disconnected graphs [137].

RNA velocities, or the changes in mRNA abundance, is another approach for studying cellular dynamics by leveraging splicing kinetics of mRNA [138, 139]. RNA velocities across genes can be used to infer the future state of a given cell.

An array of workflows for analyzing scRNA sequencing data exists, and many are currently under development for either end-to-end analysis or focused parts. Examples include pagoda2 [140], conos [141], Seurat [123, 124], Scanpy [142] and Monocle3 [143]. The general workflow described above is depicted in Figure 2.2.

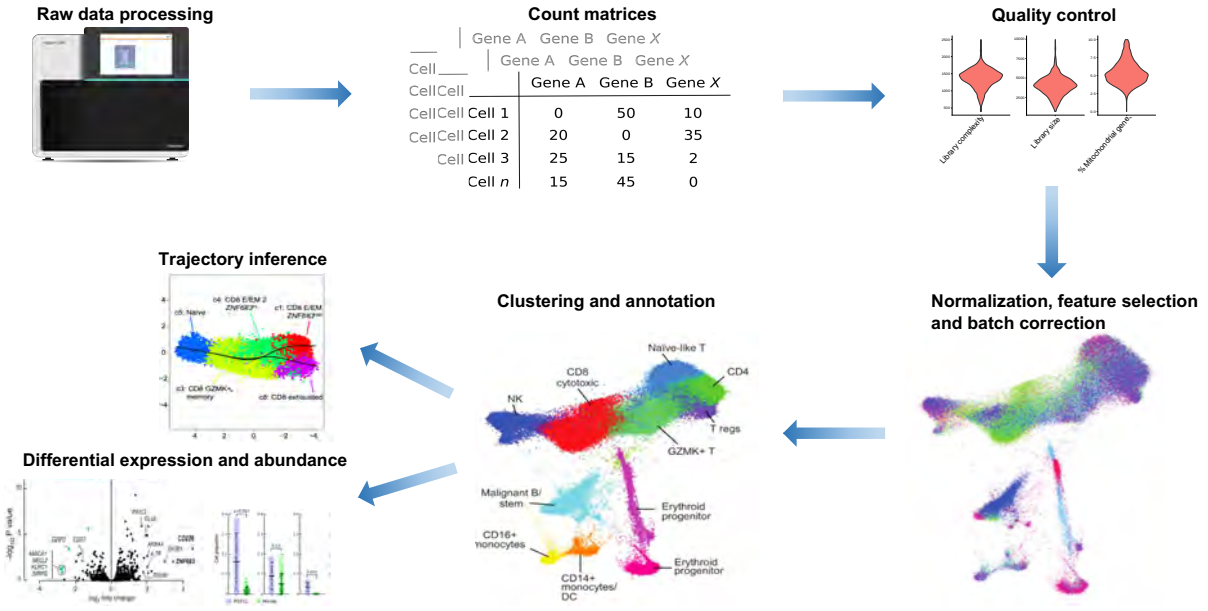


Figure 2.2 | Schematic of typical bioinformatic workflow of single-cell RNA sequencing data analysis of multiple samples.

3

Single-cell profiling of immune cells in chronic lymphocytic leukemia

Chronic lymphocytic leukemia (CLL) in the most common adult leukemia in Western countries. CLL has an incidence rate varying from less than 0.01% in Asia to 0.06% in the United States of America and Europe. The median age at diagnosis is 70 years old, and it is rarely seen in people below the age of 40, and is extremely rare in children. The prognosis of CLL patients is highly variable; some patients can live a full life without requiring treatment for long periods (or even at all), and some requires treatment imminently after diagnosis.

Historically, CLL patients were treated with chemotherapy or a combinational therapy of both chemotherapy and immunotherapy, such as anti-CD20 antibodies. During the past years the therapy has shifted from chemotherapy to using targeted agents, such as BTK inhibitors (Ibrutinib) and the BCL-2 inhibitor Venetoclax [144]. These are now the preferred initial treatments

of CLL patients in the United States [145, 146]. Ibrutinib irreversibly inhibits BTK, which is essential in the B cell receptor signaling cascade and once activated leads to increased proliferation, survival, and migration of the B cells [147]. CLL is a malignancy of $CD19^{+}CD5^{+}$ B cells that can circulate either in the bloodstream, bone marrow or secondary lymphoid tissues. Patients are diagnosed with CLL if there are more than 5,000 clonal B cells in a whole blood sample. Most CLL patients have more than 10,000 malignant B cells. CLL is divided into two subsets based on the mutational status of IGHV, with unmutated IGHV tending to lead to more aggressive disease courses. Monoclonal B-cell lymphocytosis (MBL) is the precursor stage of CLL [148, 149], and is characterized by less than 5,000 malignant cells in a whole blood sample with no additional signs of lymphoma [150]. MBL is categorized as either low-count ($<$500 cells) or high-count ($>$500 cells). Approximately 5% of adults with European ancestry at the age of 40 or above will have low-count MBL. Low-count MBL rarely progresses to CLL, although 1-2% of high-count patients will develop CLL per year [151].

CLL has several key characteristics making it an extraordinary model system to interrogate the co-evolution of immune and cancer cells. First, CLL is often associated with a slow disease progression that enables longitudinal studies. As CLL cells continuously circulate between the blood stream, bone marrow and lymph nodes [152], highly pure tumor samples can easily be drawn from the peripheral blood [153]. CLL is considered a prototype of a microenvironment dependent tumor, where cancerous cells co-evolve with host immune cells in the bone marrow or lymph node. Sampling from bone marrow and lymph nodes are despite being important microenvironments for CLL cells, more invasive and thus less common. Immune cells in CLL patients show a skewing towards generating a disease tolerant environment with phenotypes shifting towards exhausted states [154]. Even so, how immune and CLL cells co-evolve remains incompletely characterized [153].

Besides the direct effect of BTK inhibitors, such as Ibrutinib, on the malignant B cells, it is also evident that the anti-tumor effects are related to indirect effects on the TME. It is becoming increasingly recognized that Ibrutinib modulates the T cells in CLL patients by expanding memory T cells, and reducing the expression of the inhibitory molecules PD-1 and CTLA-4 [155, 156].

PAPER I

In PAPER I, we sought to investigate the circulating immune cells from peripheral blood mononuclear cells (PBMCs) in CLL through single-cell transcriptomics using 10X Genomics technologies. All single-cell data was analyzed using Seurat [123, 124]. First, two serial samples from 3 high-count MBL patients and 7 CLL patients were processed. This revealed that transcriptional dysfunction of immune cells occurs early in the disease setting, as no major differences distinguished immune cells of MBL and CLL patients. A second patient cohort was then interrogated: two serial samples from 4 MBL patients progressing to CLL along with two age-matched donors. Here, we observed transcriptomic differences between immune cells in CLL patients and age-matched healthy donors. Additionally, an increased number predicted interactions in MBL subjects, also in CLL, compared to the healthy donors was observed, especially within myeloid cells. These interactions included multiple inhibitory immune signals. Looking at an additional two patients receiving Ibrutinib, the interactions decreased to levels similar to healthy donors post treatment as depicted in Figure 3.3G. These findings warrants further assessment of the influence of myeloid cells in CLL patients.

Exploring data integration approaches

As previously mentioned, sample integration is a vital part of multi-sample analyses due to batch effects. The computational data integration carried out in relation to PAPER I spans a wide array of approaches. The following methods were tested during preliminary analysis and data exploration: 1) MUDAN [157], 2) Seurat using anchors using canonical correlation analysis, 3) Seurat using harmony [158], and finally 4) reference-based integration utilizing a CITE-seq reference of 162,000 cells measured with 228 antibodies [124] in Seurat. Detailed explanation of the final methodologies used can be found in PAPER I.

The importance of standardized sample preparation

During sample preparation, technical biases can be introduced that can impact the results that in most drastic cases could lead to false or biased discoveries. Therefore, it is very important to ensure standardized sampling conditions. Multiple confounders have shown to impact scRNA sequencing data: sample preparation, storage and processing [159–161]. A recent study [162] highlighted the effects of varying processing times of PBMC samples used for scRNA sequencing. The authors simulated common practices adopted in bio-banks and clinics by cryopreserving samples at varying time points, ranging from 0 to 48 hours. This resulted in an identification of a gene-set referred to initially as a cold-shock signature. This effect was consistent across all cell types as well as sequencing technologies. Sampling time accounted for the largest proportion of variability, and the described signature was overlapping with findings elucidated in bulk RNA sequencing [163]. PAPER I included samples collected at various sites, so therefore we investigated this proposed cold-shock signature our dataset. This displayed a clear distinction between samples from four patient cohorts included, as shown in Figure 3.1.

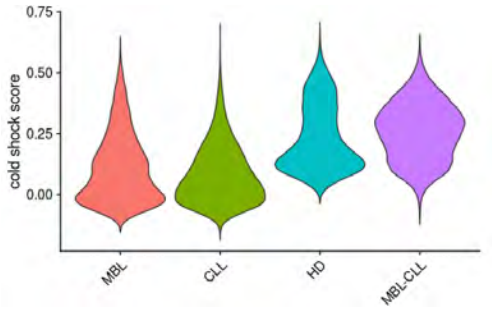


Figure 3.1 | Distribution of cold-shock signature scores for each patient cohort included in PAPER I. We analyzed 3 non-progressive MBL patients with 7 CLL patients due to similar cold-shock signature scores. Similarly, we analyzed 4 paired MBL-CLL patients with two healthy donors given their equivalent cold-shock signature scores. Signature scores were computed using Seurat [123, 124]. Figure adapted from supplementary material in PAPER I.

Methods to regress out such signals have been proposed as alluded to previously using cell cycle signals as an example. However, in this case simple linear regression was not enough to properly remove this tendency suggesting the presence of other confounding signatures. Although, this was not tested. In addition, regression of technical confounders comes with a risk of removing subtle variations and thus homogenizing cell subpopulations [162]. Consequently, the data analysis in PAPER I was split into two based on choosing samples with similar signature scores

computed in Seurat (Figure 3.2 and 3.3). The first patient cohort included two serial samples collected from 3 high-count MBL patients 7 CLL patients. Post quality control we identified 67,333 cells partitioned into 16 clusters (Figure 3.2).

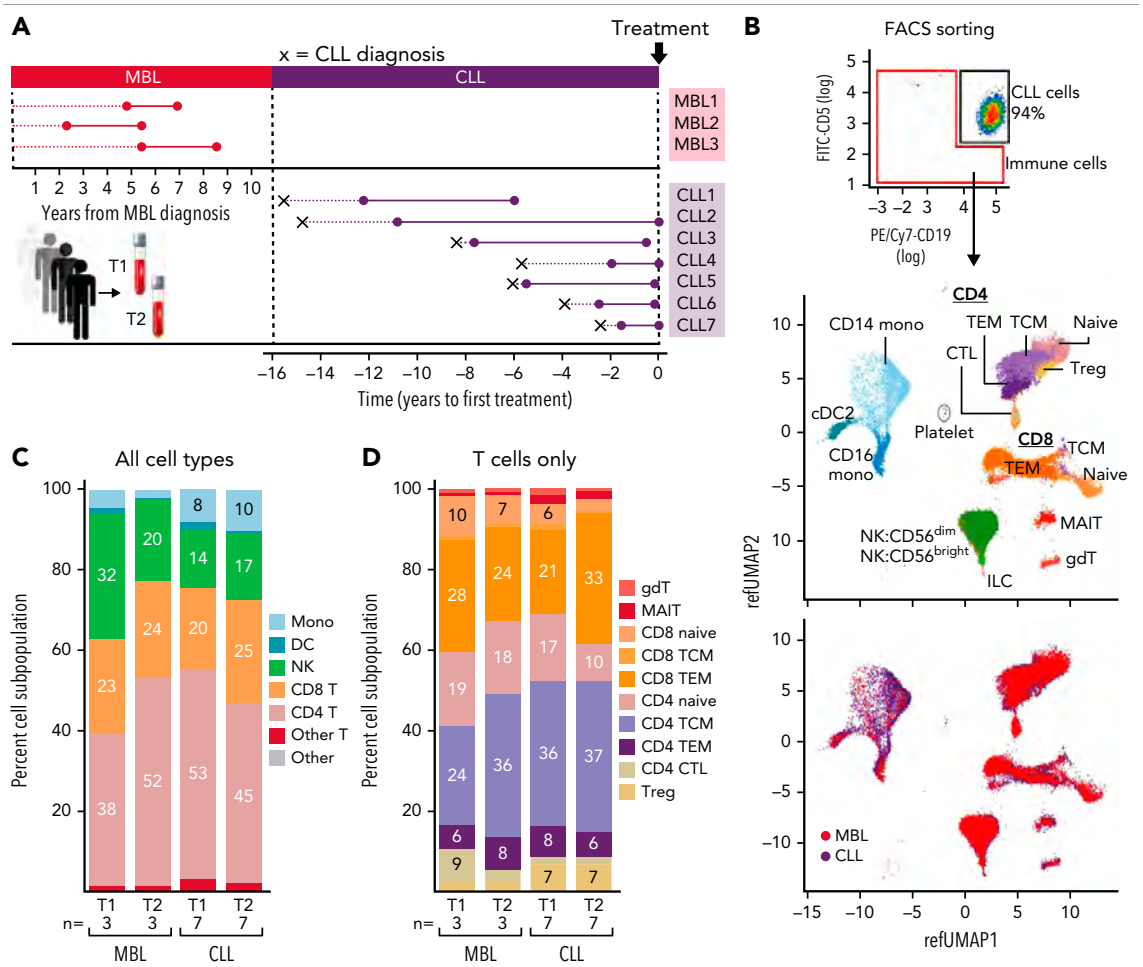


Figure 3.2 | Key findings from first patient cohort in PAPER I. **A.** PBMCs from 2 serial samples were collected for 3 patients with MBL and 7 with CLL. **B.** Non-CD19^{+}CD5^{+} cells were isolated by fluorescence-activated cell sorting (FACS) for scRNA sequencing. UMAP visualization of all immune cells colored by immune cell type (top) and CLL or MBL assignment (bottom). **C.** Immune cell type proportion at each time point in patients with MBL or CLL. **D.** T cell type proportion at each time point in patients with MBL or CLL. CTL = cytotoxic T lymphocyte, DC = dendritic cell, gdT = g-dT cells, ILC = innate lymphoid cell, MAIT = mucosa-associated invariant T cells, Mono = monocyte, NK = natural killer cell, pDC = plasmacytoid dendritic cell, T = T cell, TCM = central memory T cell, TEM = effector memory T cell, Treg = regulatory T cells. Figure adapted from PAPER I.

This highlighted that no major transcriptional or compositional differences between the two patient groups were observed, suggesting a disease tolerant environment. To confirm these findings, we evaluated the second patient cohort consisting of two healthy donors and four high-count MBL patients progressing to CLL (MBL-CLL). Figure 3.3 shows the key results from this analysis.

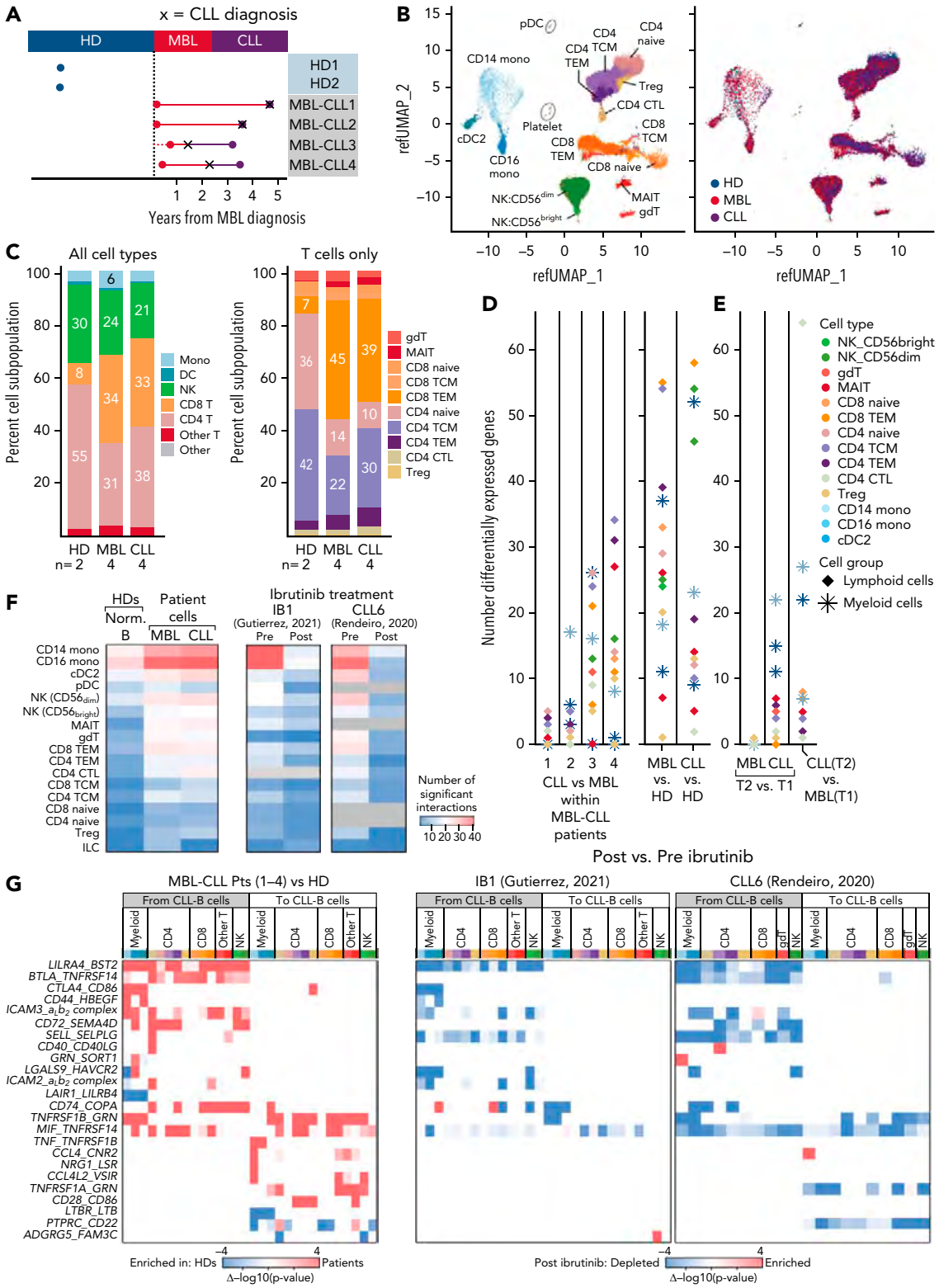


Figure 3.3 | Key results from second patient cohort in PAPER I. **A.** scRNA sequencing was performed on PBMCs collected from 4 patients with MBL (red) progressing to CLL (purple), and from 2 healthy donors (blue). X = CLL diagnosis. **B.** UMAP visualization of all immune cells colored by immune cell types (left) and by sample types (right). **C.** Proportion of immune cell types (left) and T-cell subtypes (right). **D.** Number of significant differentially expressed genes for each cell type by performing a comparison of paired samples within patients (left) or comparison between MBL samples or CLL samples vs healthy donors (right). **E.** Same analysis for significant differentially expressed genes was performed on 3 independent patients with non-progressive MBL and 7 with CLL. Figure adapted from PAPER I.

Figure 3.3 | **F.** Heatmaps with the number of the significant ligand-receptor interactions for each cell type under different conditions using CellPhoneDB (v2.1.7). Heatmap comparing the number of significant interactions between healthy donors and patient samples from either MBL stage or CLL stage (left). Heatmaps including samples before and after ibrutinib for 2 additional patients (right). Gray = insufficient cell numbers. **G.** Heatmaps representing the difference of P values for each ligand-receptor pair for specific cell types (x-axis). Interactions enriched in patients (red) or in healthy donors (blue) were calculated by subtracting 2log10(p-value) in healthy donors from 2log10 (p-value) in patients (left). The same interactions that are either enriched (red) or depleted (blue) after ibrutinib (right) are calculated by subtracting 2log10(p-value) in pre-ibrutinib from 2log10(p-value) in post-ibrutinib. Pts = patients. Figure adapted from PAPER I.

Again, we observed an absence of phenotypic changes in cell types transitioning from MBL to CLL. Contrary, immune cells of healthy donors were markedly distinguished from immune cells of both MBL and CLL patients. Particularly, the proportion of $CD8^+$ T cells was increased in CLL patients, and $CD4^+$ T cells decreased. However, there is a risk that these differences being owed to not capturing the biological variation within healthy donors due to the the small sample size. Although, we attempted to minimize this risk by inclusion of age-matched healthy donors.

Inferring cell-cell interactions

Cell-cell interactions are important in communication pathways, especially between immune and cancer cells. Quantification of cellular communication serves as an important tool across multiple disciplines [164]. The increased population of protein-protein interaction databases, has enabled inference of cellular interactions using gene expression (both bulk and single-cell). In particular, receptor-ligand pairs are used by detecting coordinate expression of cognate genes. Moreover, scRNA sequencing data has the advantage over bulk RNA sequencing that the cell types the signals originate from are known. This interrogation is also of interest within the TME in order to detect crosstalk between tumor and immune cells. This assumes that gene expression correlates to protein abundance, which is not always the case [165], and that protein abundance is enough to infer interactions, ignoring potential post-translational modifications [164]. Cell-cell interactions can be inferred with CellPhoneDB [166, 167], which is a publicly available repository encompassing about 900 ligand–receptor pairs from existing datasets.

Addressing cell-cell interactions between immune cells and malignant (or healthy) B cells in the second patient cohort revealed an enrichment of potential interactions in patients compared to healthy donors. This was observed for many of the immune cell types, in particular monocytes.

The majority of upregulated interactions involved inhibitory signals. By expanding the analysis with two additional CLL patients receiving Ibrutinib, we showed that the increased interactions was depleted upon treatment, as seen in Figure 3.3G (right).

Conclusion

To summarize, in PAPER I we presented both computational challenges and intriguing biological observations needing further investigations. This chapter exemplified the importance of standardizing sample preparations for scRNA sequencing by showing how it can impact and potentially misguide the analysis. The relatively small sample size is a clear limitation of this study, which may limit the power to observe differences along the disease trajectory and capture biological variation within healthy donors.

We showed that immune cells of healthy donors are markedly different from patients with either MBL or CLL, whereas no major transcriptional differences distinguished immune cells of MBL patients with immune cells of CLL patients. This suggests that immune deficits in CLL occurs early in the disease course.

4

Single-cell profiling of immune cells in Richter's syndrome during immunotherapy

Richter's syndrome (RS) is the transformation of CLL into an aggressive secondary lymphoma occurring in 2-10% cases of CLL patients. The aggressive nature of RS shows in a poor survival rate with a median survival of approximately one year [168, 169]. Currently, the treatment options available are limited, but typically includes an aggressive chemotherapy followed by a haemopoietic stem cell transplant for eligible patients. However, the response rates are low, remission is short and toxicity is severe, displaying a demanding need for treatment alternatives. Recent clinical trials have demonstrated unexpected responses to anti-PD-1 in RS patients with response rates of 42-65% [170–172]. However, there is a need for understanding the characteristics of response and predicting which patients will benefit from the treatment. In PAPER II we utilized data from patients enrolled in a clinical trial with PD-1 checkpoint blockade con-

currently with Ibrutinib (NCT 02420912).

PAPER II

In PAPER II we sought to discover the mechanisms of response in RS patients following PD-1 checkpoint blockade (NCT 02420912). We analyzed scRNA-sequencing data from 17 serial bone marrow samples from 6 RS patients (4 responders and 2 non-responders). The study showed that response was associated with a $CD8^+$ T cell effector/effector memory population marked by expression of the transcription factor *ZNF683*. Trajectory analysis predicted that said population to be an intermediate exhausted population evolving from stem-like memory cells and divergent from terminally exhausted cells.

Bulk RNA sequencing of peripheral blood samples from 7 independent RS patients pre-treatment (2 responders, 5 non-responders) confirmed the association of $ZNF683^{high}$ T cells and response. Furthermore, the signature overlapped with tumor-infiltrating populations from solid tumors and peripheral blood $CD8^+$ T cells from melanoma checkpoint blockade responders.

Through epigenetic analyses, we discovered that *ZNF683* directly impact key T cell genes (*TCF7*, *LMO2*, and *CD69*) and pathways for T cell cytotoxicity and activation.

Figure 4.1 summarizes the discovery cohort of which single-cell transcriptomes were analyzed during a clinical trial with PD-1 blockade treatment. The cohort included four responders (RS-R1, RS-R2, RS-R3 and RS-R4), two non-responders (RS-NR1 and RS-NR2), along with two CLL patients (CLL-1 and CLL-2) not responding to treatment. Sampling was done prior to therapy (except RS-R3), time of complete or partial response, and at progression. Bone marrow biopsies were flow-sorted to first isolate all viable lymphocytes, and subsequently only non-tumor cells ($CD5^-CD19^-$ cells). These were then prepped for scRNA sequencing with 10X Genomics, and processed and analyzed using pagoda2 [140] and conos [141] (Figure 4.1).

Tumor cells ($CD5^+CD19^+$ cells) were sorted separately, and processed, sequencing analyzed in the same manner, however, in this chapter I will focus on the results obtained from an analysis of the immune compartment. PAPER II highlights the results from the single-cell tumor analysis.

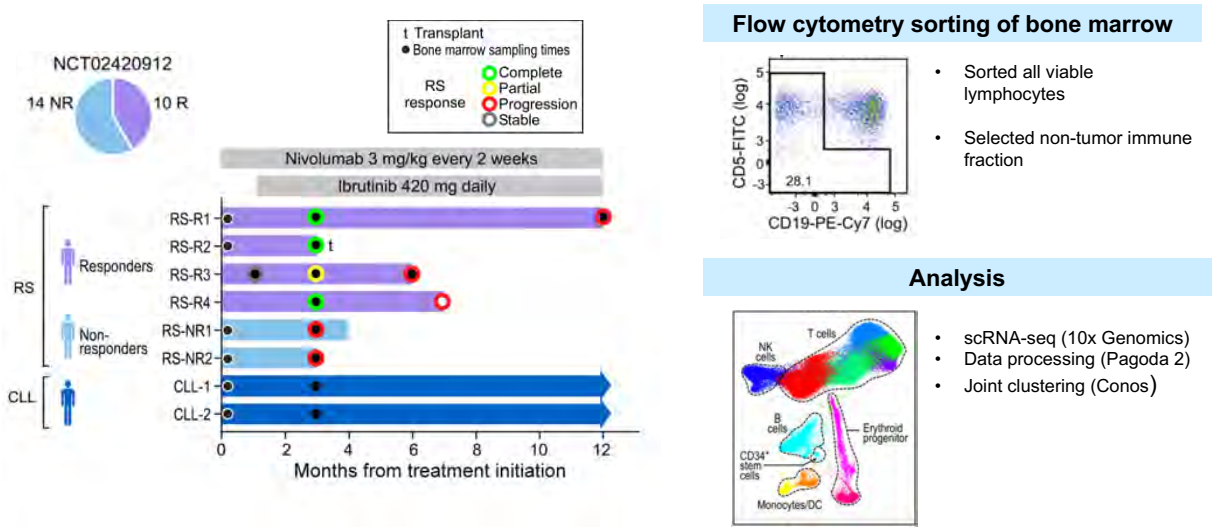


Figure 4.1 | Brief overview of discovery RS cohort, sampling times and analysis in PAPER II.

We analyzed in total 78,488 cells, of which we performed subclustering of lymphocytes (60,727 cells). Figure 4.2B-D show the phenotypic attributes of the 11 detected lymphocyte clusters. More detailed annotations of each identified cluster is provided in PAPER II. The $CD8^+$ T cells of the RS bone marrow samples revealed high cytotoxicity and exhausted states (Figure 4.2D). Furthermore, we included two age-matched healthy donors in the analysis, and an additional 28 healthy donors from publicly available datasets [173, 174]. Methodological descriptions of how additional samples were included in the analysis are detailed in the methods section of PAPER II. Inclusion of healthy donors revealed disease-specific clusters effector/effector memory T cells (cluster 1 and cluster 4), exhausted T cells (cluster 8) and T_{regs}. Contrary, healthy donors displayed a larger proportion of naïve T cells (cluster 5).

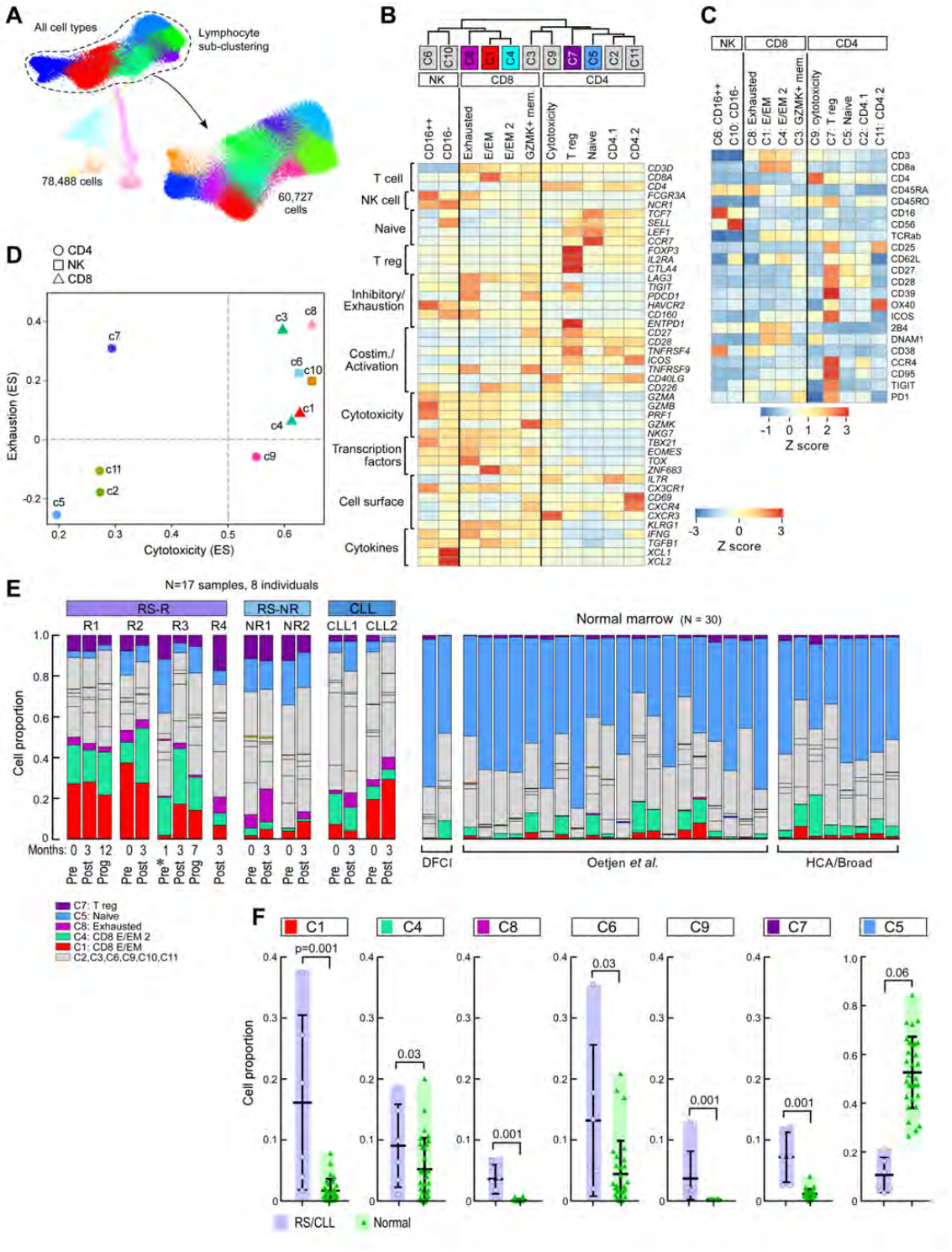


Figure 4.2 | **A.** Immune cell populations with subsequent subclustering of lymphocytes revealing 11 distinct transcriptional clusters. **B.** Heatmap of marker genes and cluster identities for 11 identified T and NK cell clusters. Cluster 1, red; Cluster 4, teal, Cluster 8, magenta. **C.** Heatmap of cell surface markers for 11 identified T and NK cell clusters using CITE-seq. **D.** Identified T and NK cell clusters plotted by Cytotoxicity and Exhaustion scores [175]. **E.** Bar graphs showing T and NK population distributions across serial marrow samples and normal bone marrow samples. **F.** Distribution of T and NK cell populations in RT and CLL marrows as compared to normal bone marrow samples (2 sample t-tests). Figure adapted from PAPER II.

We observed key transcriptional changes associated with PD-1 response, as depicted in Figure 4.3. However, one central observation is that RS responders show enrichment of $CD8^+$ T cells, primarily cluster 1, which is marked by the expression of *ZNF683*. Cluster 1, annotated as effector/effector memory $CD8^+$ T cells, showed quantitative differences already present at baseline, with a significantly larger proportion present in responders (p = 0.04).

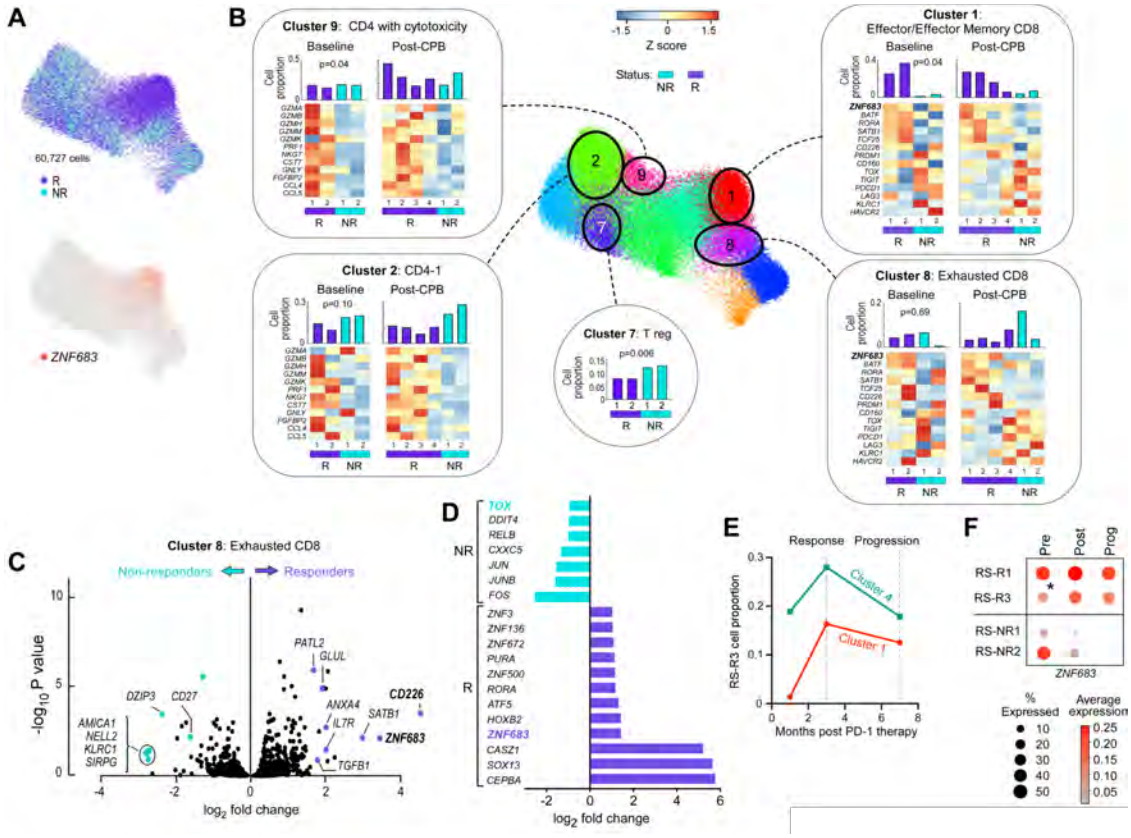


Figure 4.3 | Transcriptional changes associated with PD-1 response. **A.** Subclustering graph showing cells responders in purple and cells from non-responders in blue, as well as *ZNF683*-expressing cells (red). **B.** Cluster proportions for responders (purple) and non-responders (blue) in bar graphs on top of heatmaps showing representative cytotoxicity, exhaustion and expression changes within circled clusters. **C.** Volcano plot of gene expression differences between responder and non-responder cells in the exhausted T cell cluster (c8). **D.** Top transcription factors in comparison of all CD8 T cells between responders and non-responders. **E.** Kinetics of cluster 1 and 4 cell proportions showing expansion at time of response in RS-R3. **F.** Bubble plot showing percentage of expressing and average relative expression for RS patients with progression on PD-1 blockade. Figure adapted from PAPER II.

Within-cluster differential gene expression also highlighted *ZNF683* to mark response. Across all $CD8^+$ T cell clusters, *ZNF683* was the top up-regulated transcription factor in responders, whereas *TOX* was the dominant transcription factor of non-responders.

Linking functional T cell phenotypes to clonality

The TCR is a heterodimer consisting of two chains, the alpha and the beta chain. Single-cell approaches enable the detection of paired alpha and beta chains of TCRs, which is not possible using bulk RNA sequencing. T cell phenotypes and clonality can be linked using a combination of scRNA sequencing and single-cell TCR (scTCR) sequencing. This is extremely effective for mapping the phenotypic landscape of T cells during the course of immunotherapy [103]. In PAPER II, we performed scRNA sequencing coupled with scTCR sequencing for one RS responder and one RS non-responder pre and post treatment. This showed an enrichment of expanded clones in the responder, specifically one dominating T cell clone present both before and after treatment, as seen in Figure 4.4. This clone had high expression of *ZNF683* (Figure 4.4B), and occupied mainly cluster 1 and cluster 8. On the other hand, the clonal space of the non-responder revealed much higher diversity, with both contracting, expanding and emerging T cell clones.

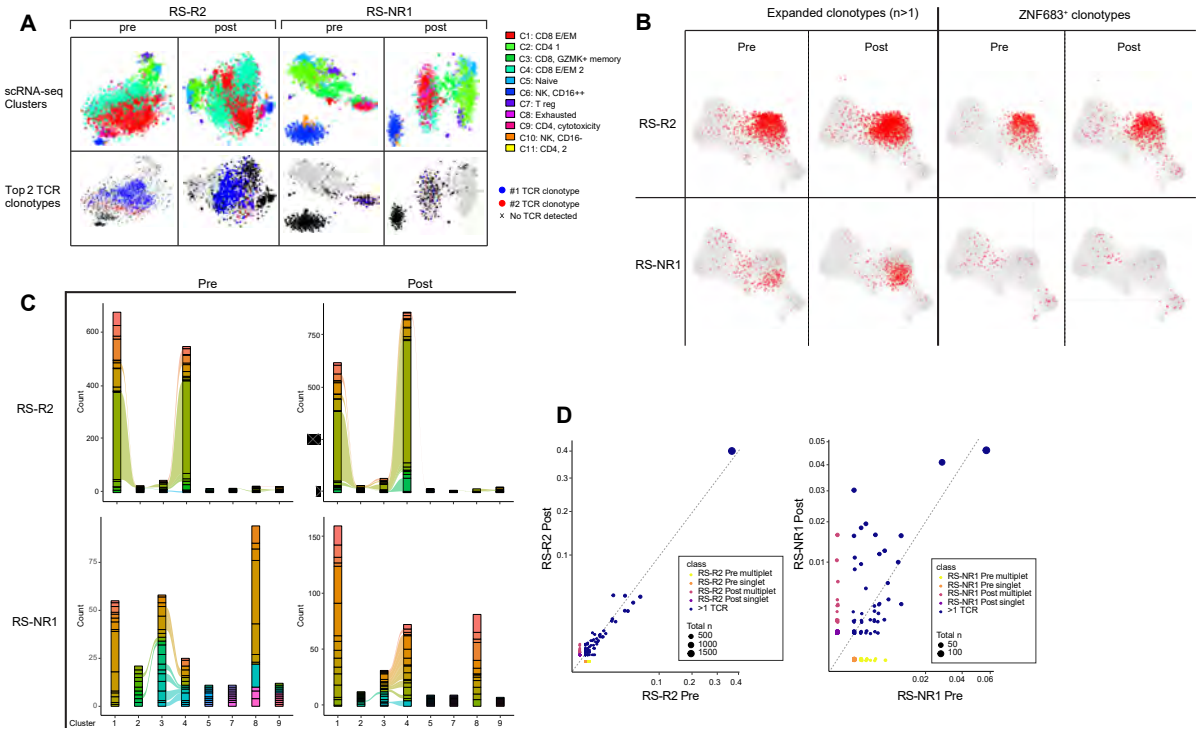


Figure 4.4 | Clonal tracing results from PAPER II. **A.** Cluster distribution for individual samples in 5' scRNA-seq performed with CITE-seq and TCR-seq (top). Top 2 expanded TCR clones in each sample (bottom). **B.** Expanded clonotypes (n > 1) for responder and non-responder pre and post treatment (left). *ZNF683*$^{+}$ clones in responder and non-responder pre and post treatment (right). C. Top 50 clones distribution per cluster. D. Scatterplot of clonotypes pre and post treatment for responder (left) and scatterplot of clonotypes pre and post treatment for non-responder (right). Plots in **C.** and *D.* were generated using scRepertoire [176]. Figure adapted from PAPER II.

Trajectory inference

As the overall distribution of CD8^{+} phenotypes suggested a continuum across tumor-specific T cell states, we performed trajectory inference of CD8^{+} T cells using slingshot [177] (Figure 4.5). Accordingly, the lineage was set to begin in the naïve T cell cluster (cluster 5).

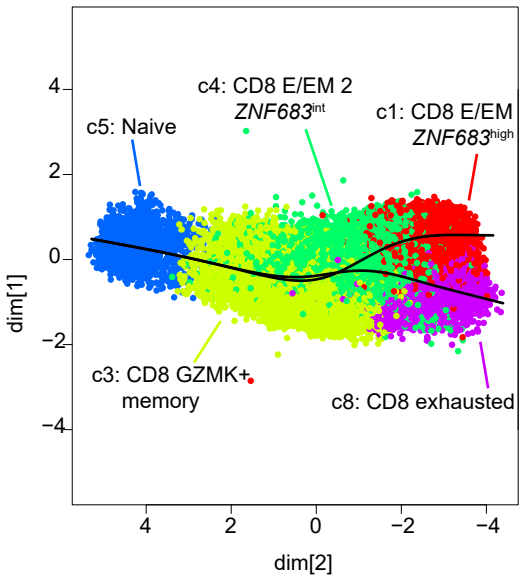


Figure 4.5 | Trajectory analysis of CD8$^+$ T cells using slingshot [177] from PAPER II. Trajectory was set to start in the naïve T cell cluster (c5), and shows a branching from cluster 4 into two divergent paths.

This showed a linear trajectory with bifurcation from the *ZNF683^{intermediate}* cluster (cluster 4) to either *ZNF683^{high}* effector/effector memory T cells (cluster 1) or terminally exhausted T cells (cluster 8). This suggests that the *ZNF683^{intermediate}* cluster serves as an alternate path towards exhausted effector function, which is divergent from terminal exhaustion.

Deconvolution of cell types in bulk RNA sequencing

scRNA sequencing is still relatively costly compared to bulk RNA, and large amounts of bulk RNA sequencing datasets exist. Therefore, it is possible to utilize signatures discovered in scRNA sequencing, and deconvoluting bulk RNA sequencing and interrogating for this signature. Deconvolution of cell type proportions from bulk RNA sequencing can readily be done using e.g. CIBERSORT [178], CIBERSORTx [179] or xCell [180].

The estimated cell type composition can then be used to dissect cell type specific expression patterns in bulk RNA sequencing, thus allowing to utilize the vast amount of bulk RNA sequencing datasets to confirm signatures detected in scRNA sequencing. This was exemplified in PAPER II, where cellular content was inferred using CIBERSORT with the LM22 signature matrix [178]. Proportions were then used to normalize gene expression data from an additional cohort comprising 35 RS patients in order to extract T cell specific expression of genes, such as

ZNF683 (Figure 4.6A (top)). Deconvolution of bulk RNA sequencing is highly essential, as the samples consist of mixtures of cells present in various abundances.

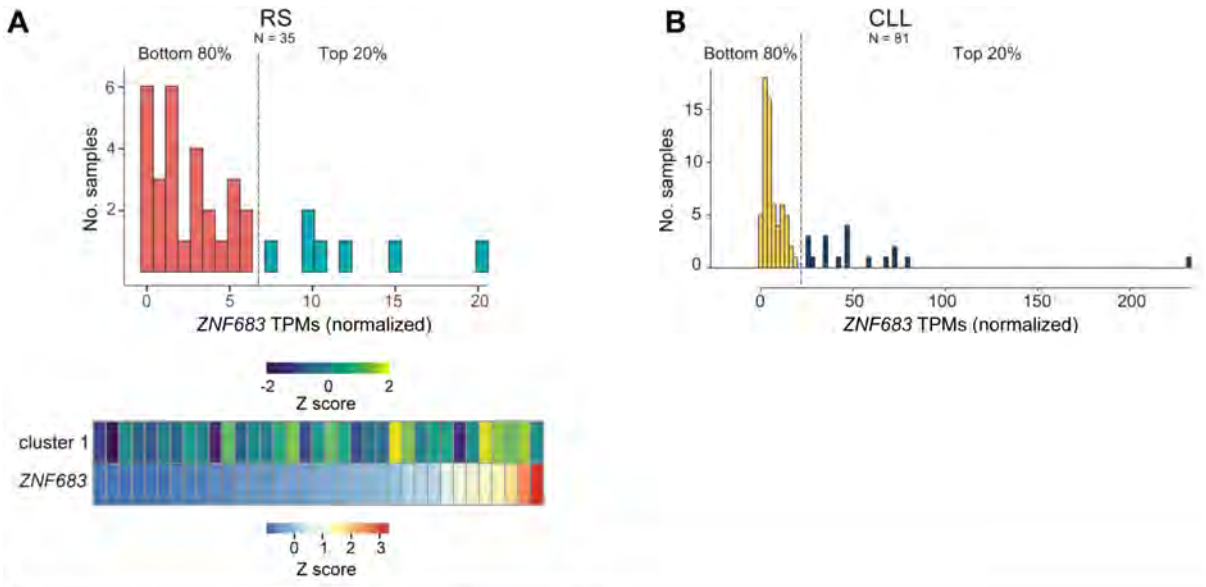


Figure 4.6 | **A.** *ZNF683* TPMs corrected for T cell content estimated by CIBERSORT [178] in RS bulk RNA seqeuncing samples (N=35). Top 20% of samples = blue, bottom 80% = red (top). Cluster 1 signature and T cell normalized *ZNF683* expression in bulk RNA sequencing data from 35 independent RS patients (bottom). **B.** *ZNF683* TPMs corrected for T cell content estimated by CIBERSORT [178] in CLL bulk RNA sequencing samples (N=81). Top 20% of samples = blue, bottom 80% = yellow. Figure adapted from PAPER II.

Additionally, the same analysis was performed on a cohort of bulk RNA sequencing samples from 81 CLL patients (Figure 4.6B). Both analyses showed that *ZNF683* expression is also detectable in bulk RNA sequencing with 20% showing high expression.

Furthermore, using top up-regulated genes marking the $ZNF683^{high}$ effector/effector memory T cell cluster (cluster 1), the enrichment for each RS bulk RNA sequencing sample was assessed using single-sample GSEA (ssGSEA) [181], showing overlap between high expression of *ZNF683* and high enrichment score (Figure 4.6A (bottom)).

Functional studies of *ZNF683*

In order to functionally annotate the transcription factor *ZNF683* and elucidate potential targets, experimental validation using Jurkat cell lines was performed by over-expression of *ZNF683* (+ doxycycline) (Figure 4.7A) followed by CUT&RUN [182] and bulk RNA sequencing (Figure 4.7B).

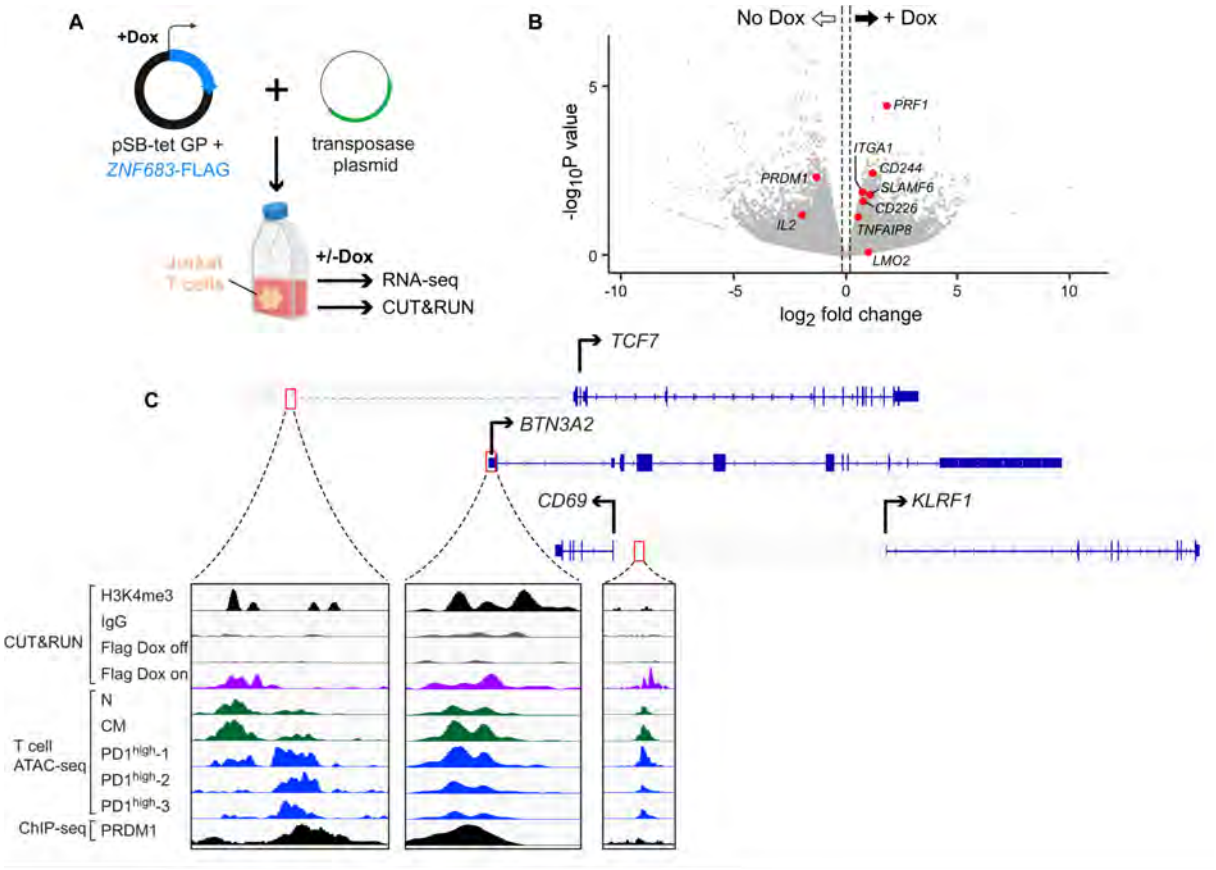


Figure 4.7 | **A.** Schematic of doxycycline-inducible expression of *ZNF683* in Jurkat cells. **B.** Volcano plot of differentially regulated genes by *ZNF683* induction by RNA sequencing. **C.** CUT&RUN data from Jurkat cell lines (top) shows binding of *ZNF683* at regions surrounding key immune genes that correspond to differential ATAC-sequencing peaks in T cell subsets and prior PRDM1 ChIP-seq data [183]. N = naïve, CM = central memory, PD-1high = PD-1 high tumor infiltrating $CD8^+$ T cells. Figure adapted from PAPER II.

Differential gene expression highlighted that over-expression of *ZNF683* resulted in up-regulation of *PRF1*, *ITGA1*, *CD244*, *CD226* and *IL2RB* as well as down-regulation of *PRDM1* and *IL2* (Figure 4.7B). Differential peaks and differential genes were integrated using CISTROME-GO [184] defining a set of potential targets of *ZNF683*. Some of these targets overlapped with publicly available ATAC-sequencing from mice [185] (Figure 4.7C). Detailed experimental and computational methodologies can be found in PAPER II.

In addition, data integration revealed enrichment of pathways involved in antigen binding and presentation, transcription factor activity, and T cell mediated cytotoxicity and cell killing upon *ZNF683* expression. These findings indicate that *ZNF683* is a regulator of key T cell function and immune response.

Implications of *ZNF683* expression

ZNF683 was also found to mark immune populations in other cancer types. A pan-cancer study [186] showed transcriptionally distinct populations highly overlapping with the signature presented in PAPER II (Figure 4.8, left).

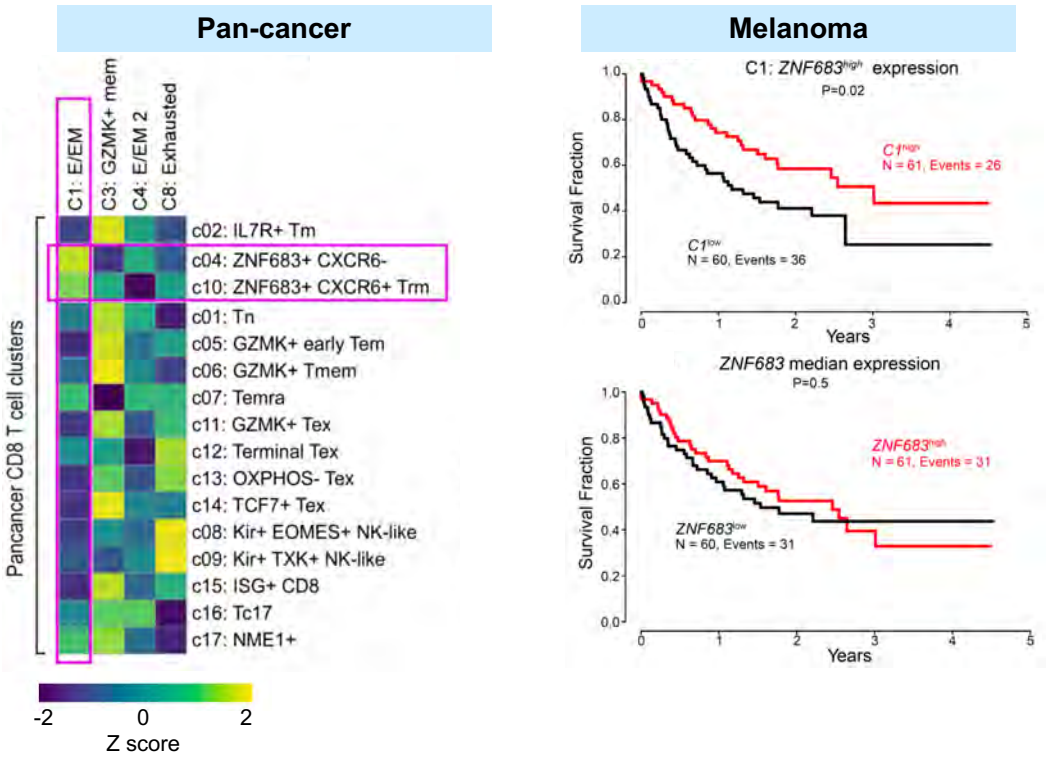


Figure 4.8 | Heatmap showing single-cell GSEA scores for identified CD8$^+$ T cell clusters as compared to Pan-cancer analysis CD8$^+$ T cell clusters [186] (left). Survival curve showing *ZNF683* expression and cluster 1 gene expression signature with overall survival in melanoma patients treated with PD-1 checkpoint blockade [187] (right). Figure adapted from PAPER II.

Further examination of bulk RNA sequencing data of melanoma patients treated with PD-1 checkpoint blockade [187] also showed the cluster 1 gene signature was associated with response ($p = 0.02$) (Figure 4.8, right).

In addition, we confirmed the signature to be present in the peripheral blood of RS patients by expanding the analysis to a validation cohort of bulk RNA sequencing of T cells from seven independent RS patients. Of this, we performed differential gene expression analysis between

two PD-1 checkpoint blockade responders and five non-responders. Again, *ZNF683* was among the top differentially expressed genes between the two groups (Figure 4.9A). We also detected overlaps with signatures associated with $CD8^+$ T cells in melanoma patients responding to PD-1 treatment [188] (Figure 4.9B), and with neoantigen-specific $CD8^+$ T cells early in PD-1 treatment detected in the peripheral blood of lung cancer patients [189] (Figure 4.9C).

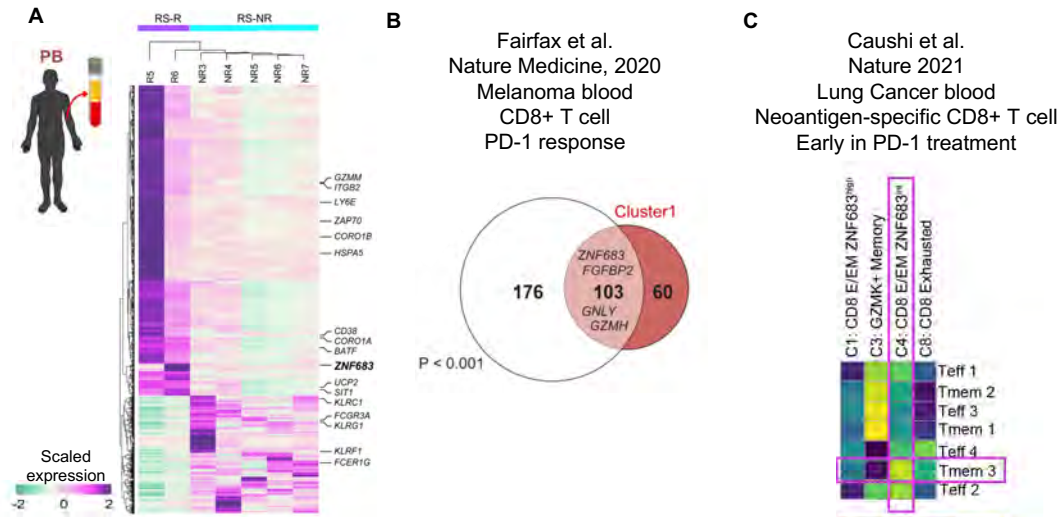


Figure 4.9 | **A.** Heatmap of differentially expressed genes from bulk RNA sequencing of peripheral blood human T cells from additional responders and non-responders highlighting *ZNF683* is associated with PD- response. **B.**. **C.** Heatmap of s. Figure adapted from PAPER II.

Conclusion

Response to PD-1-blockade appeared to be associated with an increased proportion of $ZNF683^{high}$ CD8 effector/effector memory T cells in bone marrow of RS patients. This $ZNF683^{high}$ population showed to be a divergent path from terminal exhaustion. Through functional studies, we detected that *ZNF683* regulates key pathways in T cell differentiation, activation and cytotoxicity. The $ZNF683^{high}$ signature was detectable in peripheral blood and seemingly correlates with response to checkpoint blockade. This presents a promising signal, that requires future studies in order to establish the predictive potential of such population.

5

Single-cell profiling of cancer cells

While chapter 3 and 4 focused on characterization of the immune cells in the TME, the following chapter will focus on profiling the cancer cells themselves in CLL and RS, respectively.
Cancer is thought to begin with changes in a single cell [53]. By selective forces, additional changes accumulate leading to heterogeneity among cancer cells both within a patient and between patients diagnosed with the same disease.
High-dimensional single-cell technologies are therefore a natural fit for studying cancer. It has already been widely adopted, and has transformed the understanding of cancer heterogeneity. This chapter will focus on analyzing cancer heterogeneity using single-cell transcriptomic profiles.

5.1 Profiling chronic lymphocytic leukemia cells

CLL has served as a model disease for studying cancer heterogeneity, and a vast amount of genetic characterizations have been made. Studies have shown a high genetic variability between CLL patients, although the overall tumor mutational burden is lower compared to solid tumors [153, 190–193]. Besides genetic aberrations, tumor heterogeneity also includes transcriptional and epigenetic changes along with interactions between cancer and immune cells [194–196]. Both bulk and single-cell sequencing technologies have facilitated longitudinal studies of clonal evolution in CLL patients [194].

However, some aspects of the disease and it's spectrum remains uncharted, and there is currently a poor understanding of why some cases of MBL progress into CLL and what the underlying mechanisms are. Accordingly, there is a need for research to get a better understanding of the disease course and what some of the risk factors are.

DNA methylation as an essential epigenetic mechanism in cancer

Cancer was long thought to be a genetic disease, however, it is now also evident that epigenetics influence the disease [197]. Epigenetics is describing changes that are not genetic, which are heritable through cell division [198, 199]. There are three main types of epigenetic changes: DNA methylation, genomic imprinting and histone modification. DNA methylation is the biological process of adding methyl groups to DNA, and is one of the essential epigenetic mechanisms that control cell proliferation, apoptosis, cell cycle and differentiation [200].

Abnormal DNA methylation in as disease hallmark of cancer presenting global hypomethylation and local hypermethylation [201]. Hypermethylation denotes the acquisition of methylation leading to transcriptional suppression and decreased gene expression. Hypomethylation is contrary lacking of methylation and is associated with chromosome stability. In cancers, hypermethylation is often observed in promoters regions, thereby silencing of tumor suppressor genes, which ultimately leads to the proliferation of cancer.

Differences in DNA methylation between CLL patients and normal B cells as well as within CLL subtypes have been established [202–206]. However, little is known about the methylation changes, if any, arises in the transition from a normal to a precursor to cancer stage.

PAPER III

The goal of PAPER III was to investigate the emergence and dynamics of the cancer methylome and transcriptome. Genome-wide DNA methylation in MBL and CLL patients was characterized, including serial samples collected across disease course. An aberrant tumor-associated methylation landscape at the time CLL diagnosis was observed, with no significantly differentially methylated regions in the transition of high-count MBL to CLL. Patient methylomes showed stability across natural disease and post-therapy progression. This was also apparent in methylation and transcriptomes of single CLL cells. This longitudinal study highlights that the cancer methylome emerges early and persists suggesting a key role in disease onset.

The work carried out in relation to this thesis consisted in preliminary and exploratory analysis of 5 serial scRNA sequencing samples from patients with MBL progressing to CLL (patients A-E) along with two age-matched healthy donors (HD_1, HD_2). Tumor cells from patient samples were sorted using Fluorescence-activated Cell Sorting ($CD5^+CD19^+$), and sequenced using 10X Genomics. The analysis included quality control assessment, normalization, and testing a set of computational integration approaches of 60,630 single-cells. However, all highlighted the same key results: malignant B cells are closer within patient and mostly almost overlapping than between patients and compared to healthy B cells, as seen in Figure 5.1B, and 5.1E.

Complimentary to a comprehensive methylation analysis, PAPER III profiled the transcriptomes of about 60,000 cells from two healthy donors and five matched MBL-to-CLL samples. Nine transcriptionally distinct clusters were identified, with four of them originating from the healthy donors (T cells, NK cells, myeloid cells and healthy B cells), and the remaining five comprised on of the malignant cells from each respective patient. Deconvolution of the composition within a patient cluster (Figure 5.1E) showed a remarkable overlap between MBL and CLL cells, highlighting the transcriptional similarity of the MBL and CLL cells.

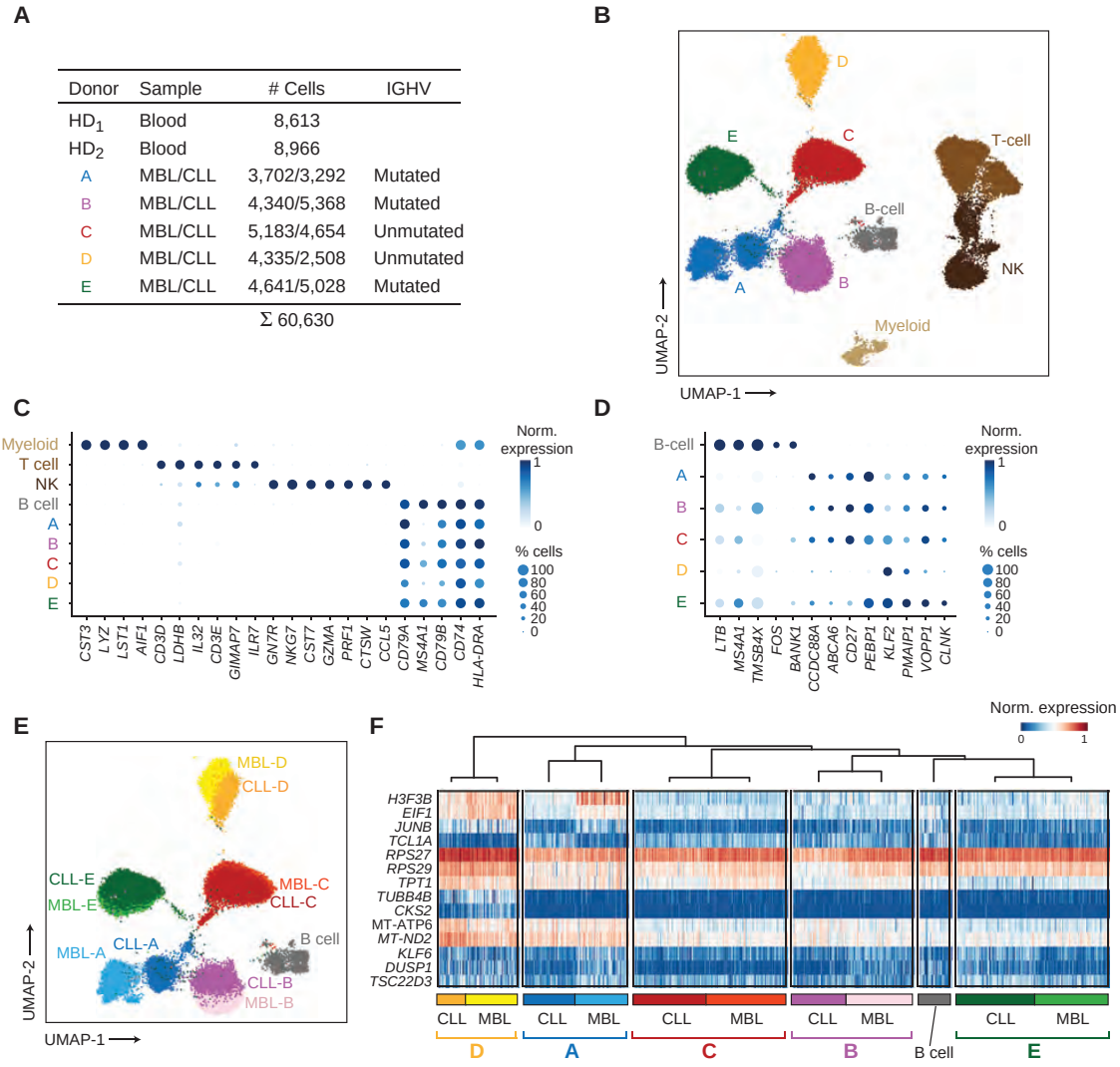


Figure 5.1 | Single-cell transcriptome analysis from PAPER III. **A.** Cohort overview. Patients were profiled in MBL as well as the CLL stage with scRNA sequencing using 10X Genomics. HD = healthy donor. **B.** UMAP of detected clusters. Healthy donor cells split into B cells, T cells, myeloid cells, and natural killer (NK) cells. The MBL and CLL cells of patients occupy distinct clusters, but are overlapping within a patient. **C.** Normalized gene expression and percentage of positive cells of marker genes. **D.** Normalized gene expression and percentage of positive cells of marker genes between healthy B cells and all patient cells. **E.** UMAP of MBL, CLL, and healthy B cells. **F.** Heatmap of marker genes between MBL and CLL cells within patient. Figure adapted from PAPER III.

Conclusion

The findings of PAPER III emphasize the cancer heterogeneity within patients of the same disease. The single-cell transcriptomic analysis captures this heterogeneity as malignant cells are more similar within patient compared to across patients. This in concordance with findings from other studies [207], highlighting the absence of strong selective pressure. In addition, based on patient methylomes and transcriptomes, it seemed that changes emerge early and are

persistent, suggesting a key role in early disease onset. This lack of heterogeneity over time also suggest a more flexible selection process of the cancer cells. Maybe it even indicates that there is less immune control, as the cancer cells are allowed to exist as they are and are not forced to evolve in order to escape immune killing. Perhaps this is also why many CLL patients can live with the disease over long time periods.

5.2 Clonal evolution of cancer cells in Richter's syndrome

RS, is as mentioned an aggressive secondary lymphoma developing in CLL patients. Despite deep characterizations of CLL [208, 209], understanding of driver mechanisms of RS remains unlimited. This is in part due to difficulties in acquiring matched CLL and RS cell samples challenging both evolutionary analyses and detection of molecular events underlying the transformation. Studies have shown alterations in genes such as *TP53, CDKN2A/B* and *MYC* [210,211]. Also, aberrations in *NOTCH1* has been detected both in whole-exome sequencing and bulk RNA sequencing [212]. A subset of RS is believed to be clonally unrelated to the underlying CLL [213], however, large whole-genome studies have not yet been performed to exclude any shared ancestry.

As RS biopsies contain admixtures of RS and CLL cells, either computational deconvolution is necessary. scRNA sequencing is another option for separation and study of these two cell types.

PAPER IV

In PAPER IV, we deciphered the genetic mechanisms underlying RS, by computationally deconvoluting admixtures of CLL and RS cells from 52 patients with RS, evaluating paired CLL/RS whole-exome sequencing data. RS-specific somatic driver mutations were uncovered along with recurrent copy number alterations, recurrent whole genome duplication and chromothripsis. This was confirmed in 45 independent RS cases and in an additional set of RS whole-genomes. We observed pathways that were dysregulated in RS compared to CLL. In addition, we were able to detect clonal evolution of the RS transformation at single-cell resolution and identifying intermediate cell states using scRNA sequencing. The study also defined distinct molecular subtypes of RS and

highlighted cell-free DNA analysis as a potential tool for early diagnosis and disease monitoring.

The work carried out in relation to this thesis comprised of patient-specific analysis of single-cell transcriptomes isolated from either bone marrow or lymph node biopsies of 5 RS patients. The data analysis included: quality control, data cleaning, dimensionality reduction and clustering followed by a preliminary analysis of copy number variations using *infercnv* [214].

From bulk RNA sequencing, PAPER IV identified s set of differentially expressed genes associated with transformation to RS (292 up-regulated and 111 down-regulated) (Figure 5.3a-b.). To further examine the transformation at high-resolution, we employed scRNA sequencing of samples from five patients at time of RS diagnosis. This revealed clonally related CLL and RS cells (Figure 5.3d-e). In addition, RS cells displayed a higher transcript abundance, both in bulk and scRNA sequencing. For the quality control of the scRNA sequencing data, this also meant that we increased the upper limit of the thresholds for library size to ensure the inclusion of RS cells.

Data cleaning

As briefly discussed, doublets in scRNA sequencing data occur frequently and can have substantial impact on the biological interpretation of the results. In PAPER IV, we detected several doublets by co-expression of otherwise exclusive cell type markers in a cluster (data not shown). Therefore, we estimated doublets using DoubletFinder [116]. In doing so, we removed between 5.2% and 7.4% doublets.

Any cell-free RNA within the input solution in droplet-based scRNA sequencing technologies are also captured, and sequencing of these constitute the background contamination that is another potential confounder influencing the biological interpretation of the data. This background noise is often referred to as ambient RNA, and methods to denoise data for this exist [215,216]. For the scRNA sequencing data included PAPER IV, we decided to denoise the data for ambient RNA using CellBender [215].

Finally, for some scRNA sequencing samples included in the patient cohort of PAPER IV, we

observed that initial clusterings were heavily confounded by cell cycle genes based on a published list of cell cycle markers [217]. An example is shown in Figure 5.2. Cell-cycle signals were subsequently removed with linear regression.

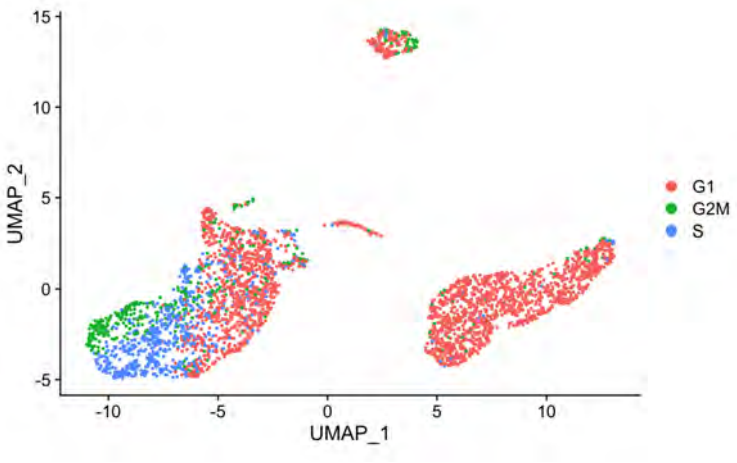


Figure 5.2 | UMAP of single-cell RNA sequencing data from one Richter's syndrome patient colored by cell cycle phase (G1, G2M, S) before correction. Cell cycle phases were determined with cell cycle scoring and regression in Seurat [123, 124].

Post data cleaning and subclustering of malignant B cells only, we identified transcriptionally distinct clusters that could be separated as CLL, RS or a transitional state, as seen in Figure 5.3e-d (middle). Based on previous studies highlighting cancer heterogeneity between patients (such as 5.1), we performed analysis on a per-patient basis.

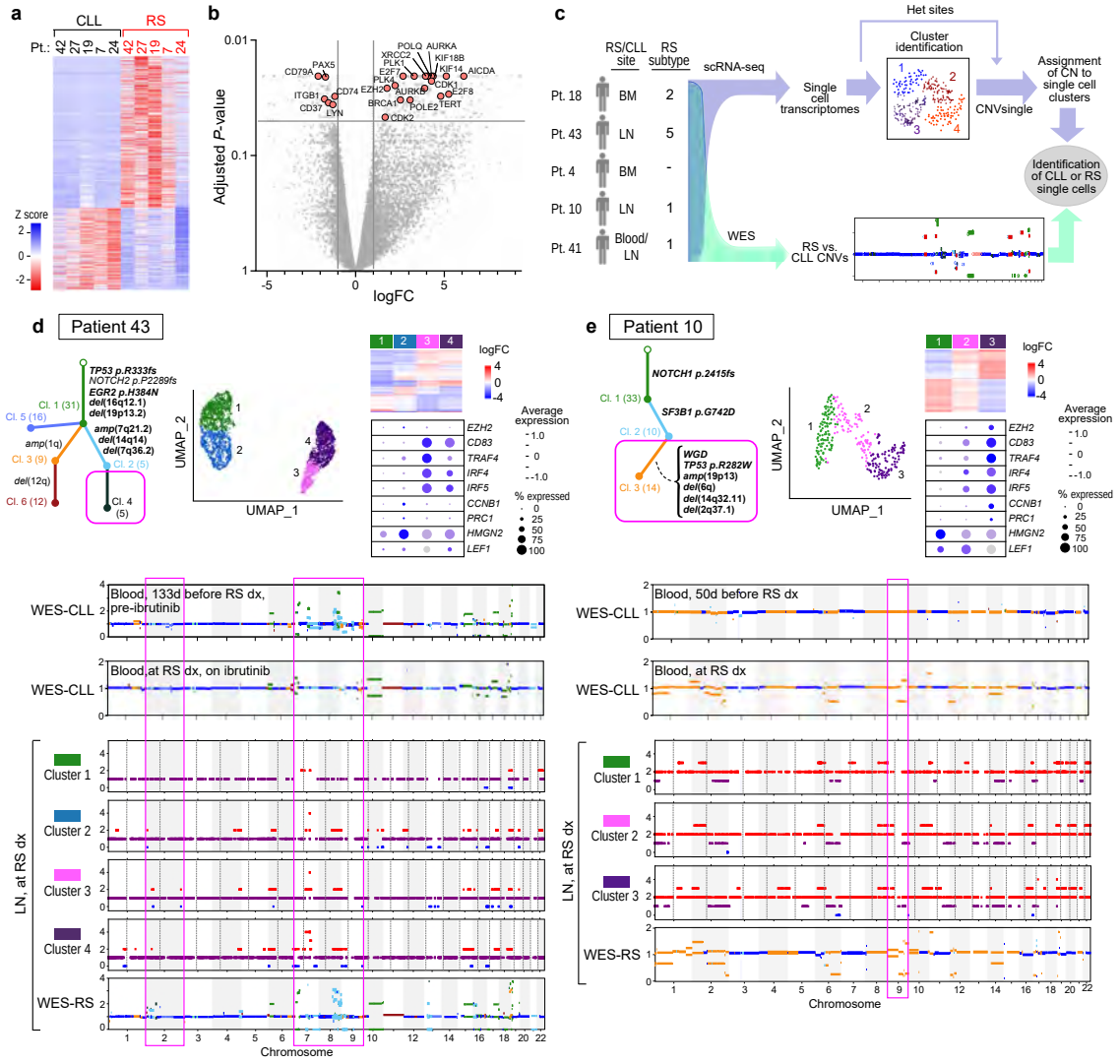


Figure 5.3 | Transformation to Richter's syndrome (RS) at single-cell resolution. **A.** Heatmap of differentially expressed genes (FDR<0.1, log2-foldchange>1) comparing paired RS and CLL samples. **B.** Volcano plot of gene expression changes in RS compared to CLL. Pink dots = selected relevant genes. **C.** Schematic of copy number changes assignment to single-cells enabling identification of CLL and RS cells. **D-E.** Single-cell RNA sequencing data shows transcriptional differences between RS and CLL from Pt 43 in **D.**, and Pt 10 in **E.**, and highlights intermediate states. Phylogenetic tree showing clonal structure of RS from WES data (top left) and UMAP visualization of RS and CLL single-cells (top middle). Heatmap of differentially expressed genes between clusters (top right) and dot plot showing cluster expression of representative genes in deregulated pathways. Inferred allelic copy number from *CNVsingle* for each single-cell cluster (bottom) depicted adjacent to WES allelic copy number plots color-coded to show copy number events assigned to CLL and RS clones. Figure adapted from PAPER IV.

Inferring copy number changes from single-cell transcriptomic data

Copy number variations (CNVs), defined as large stretches of DNA displaying copy number differences, is another avenue for achieving genetic diversity. CNVs are important drivers of rapid adaptive evolution and disease progression of cancers [218]. These variations can be

detected by a variety of approaches of next-generation sequencing including whole-exome-, whole-genome-, and cell-free DNA sequencing [219–222], however, CNVs can also be estimated from bulk RNA sequencing [223].

Now, computational tools capable of inferring CNVs from tumor scRNA sequencing data are emerging [214, 224]. In PAPER IV, a novel tool, *CNVsingle* [225], was developed for CNVs predictions that provides allele-specific copy number profiles for tumor cell clusters. Furthermore, *CNVsingle* does not rely on a reference, and it greatly reduces the signal-to-noise ratio compared to existing methods.

From the scRNA sequencing data, CNVs were predicted with *CNVsingle*, and these mapped to our detected transcriptionally distinct clusters (Figure 5.3d-e (bottom)). In-depth explanations and results can be further accessed in PAPER IV.

Conclusion

In PAPER IV, we presented a comprehensive evolutionary study of the largest series of paired CLL and RS specimens so far that integrates both genomic, cell-free DNA and single-cell analysis. The obtained results showed both distinct molecular events preceding and defining RS as well as identification of novel driver genes. These findings can improve future diagnosis and prognosis of RS patients. Furthermore, PAPER IV showed that scRNA sequencing data was able to distinguish CLL cells from intermediate states and RS cells with expression differences marking each of these. Although, the scRNA sequencing data analyzed in PAPER IV showed several issues that required attention including removal of doublets, ambient RNA and enrichment of cell cycle genes.

6

Conclusion

This thesis addresses various aspects of the bioinformatic analysis of the interplay between the cells of the tumor microenvironment. The main data type analyzed is gene expression data from either bulk RNA sequencing or scRNA sequencing, with a heavy focus on the latter. scRNA sequencing shows an immense potential for profiling cells of heterogeneous systems and discovering cell populations or signals that otherwise would have been missed using bulk technologies.

This serves as an extremely valuable tool for dissecting intra-tumor heterogeneity and immune cells present in the tumor microenvironment. In addition, coupled gene expression with CITE-seq data, scTCR- and scBCR sequencing data links phenotype to clonality providing a more comprehensive view of each single cell.

Elucidating the disease course of chronic lymhocytic leukemia

The work carried out in relation to this thesis have furthered the research of the co-evolution of cancer and immune cells along the disease spectrum of CLL spanning from the precursor stage, MBL, to the transformation into the aggressive, secondary lymphoma RS.

We have established in PAPER I and PAPER III that transcriptional changes both in immune and cancer cells happens early in CLL, as changes are already observed at the precursor stage compared to healthy donors. This has supported other findings highlighting that transition happens early observed in other studies and modalities, including aberrant DNA methylation patterns. Much focus has been on T cell deficits in CLL [155, 226], however, we have shown there is an increased number of interactions between the myeloid cells and the cancer cells including inhibitory molecules. Upon treatment these interactions are depleted, highly suggesting a myeloid role in the disease that should be studied further.

At the other end of the spectrum RS lies. A secondary lymphoma that is challenging to both diagnose and treat [227]. Several clinical trials are currently undertaken, and here accurate diagnosis is critical to ensure that trial outcomes are not influenced by underlying CLL. It is also important to ensure patients treated outside clinical trials receive the most optimal treatment. In PAPER IV, we identified novel molecular events that drive the transformation of CLL into RS, and suggested a framework that can be applied to other transformed cancers. Finally, we advocate for a non-invasive detection of RS by using cell-free DNA. This will improve diagnosis and prognosis for RS patients.

Exhaustion of $CD8^+$ T cells is a well-established phenomenon in many cancers, but the heterogeneity and variation among cancers is incompletely understood. In PAPER II, we have showed that an intermediate exhausted population expressing the transcription factor *ZNF683* served as an alternate path towards exhausted effector function divergent from terminal exhaustion. Our findings indicate that *ZNF683* plays an essential role in the choice of differentiation path in anti-tumor T cells, and that *ZNF683* was associated with response to checkpoint blockade warranting for further assessment of the predictive power of *ZNF683* both in RS but also other cancers. These two studies of the characteristics of RS, will improve and guide treatment of patients with otherwise poor prognoses.

Bioinformatics - where biology meets data science

The thesis provides insight into some of the essential computational aspects of handling scRNA sequencing data. Bioinformatic analysis of the scRNA sequencing requires strong knowledge within both biology and data science. Here, I have highlighted several key examples of where analytical choices are being made in order to continue the analysis.

First of all, when analyzing scRNA sequencing data, a lot of effort is being put into cleaning the data due to the large amount of noise present. Much often a publication utilizing scRNA sequencing will depict some sort of dimensionality reduction of the identified cell types, however, reaching this point is an extremely iterative process of cleaning and re-processing the data. As touched upon in section 2.6, initial quality control of scRNA sequencing data often includes filtering cells based on three metrics, in order to ensure inclusion of viable cells only. Yet, without prior knowledge of the cell types investigated this is a challenging task. RS cancer cells, as an example, are typically defined by a large morphology and containing more expressed transcripts at higher abundance compared to both CLL cells and immune cells, and therefore thresholds should be set keeping this in mind. Biological processes such as apoptosis and cell cycle can heavily influence the gene expression and dominate other potentially interesting signals. In PAPER IV, we showed how part of a dataset can be highly influenced by cell cycle genes, which needed to be handled.

Clustering is an important part of scRNA sequencing analysis and is often semi-supervised as it requires some sort of input affecting the clustering, such as resolution or the expected number of clusters. These features can be continuously modified and the resulting clustering should be re-assessed biologically. In addition, poor quality cells that have escaped initial quality control or other oddly behaving clusters can be identified, removed followed by a re-analysis of the cleaned up data.

The identification of an enrichment of a gene set related to sample handling in PAPER I is a strong example of how to detect confounding factors in the data followed by an attempt to reduce the effect of this. After realizing that it is not possible, we changed the analysis strategy to avoid misleading conclusions based on the detected confounder.

A common task in all analyses of scRNA sequencing carried out in this thesis included integra-

tion of multiple samples requiring batch correction. As mentioned, there are several ways of doing so, which is also exemplified here. The optimal method can vary between datasets [228] depending on which cell types are expected and what the goal of the study is. For larger datasets, it is often more feasible according to compute time and memory usage to use reference-based methods. Although, depending on the chosen reference, it can be challenging to identify rare and novel cell types with reference-based methods.

Even with the potential of scRNA sequencing, bulk RNA sequencing is still highly informative, and routinely used in clinical sampling. The nature of bulk RNA sequencing is an admixture of cells present in various abundances across samples. Therefore, when analyzing bulk RNA sequencing and looking for population specific expression, it is extremely important to estimate cell type proportions. These can then be used to obtain a ”corrected” expression value based on the cell type of interest. In PAPER II, I have completed deconvolution bulk RNA sequencing and used these to detect population specific expression patterns.
Finally, in PAPER II I demonstrated a strong cross-field analytical skills not only pertaining to RNA sequencing, but also epigenetic profiling regarding analysis ATAC-sequencing and CUT&RUN. Data obtained from these technologies combined with bulk RNA sequencing from an experiment of over-expressing a transcription factor were integrated in order to support the evidence the regulatory role of that transcription factor.

Future perspectives

There still exist a wide range of open-ended questions in the field of cancer research. A common goal is to optimize treatment often by the identification of biomarkers of response and disease markers that may lead to better and earlier diagnosis.
Multimodal single-cell technologies can assist in elucidating some of these questions. More modalities can now be measured per single cell giving a more comprehensive insight to the state of the cell. As these single-cell technologies continuously develop, we can also expect the size of datasets to increase. Recent examples of these large datasets include scRNA sequencing data from 500,000 cells across 24 tissues and organs [229]. Examples of large multimodal datasets include the publication of a pan-cancer T cell atlas consisting of coupled scRNA- and

scTCR sequencing of 316,000 T cells from 21 cancer types from 316 patients [186], and a CITE-seq study encompassing 162,000 cells and 228 antibodies measured [124]. Simultaneous quantification of gene and protein expression at single-cell resolution is eminently informative, as these two molecules are highly linked, but do not consistently correlate [230].
mRNA expression of cognate ligand-receptor pairs is not the only requirement of cellular communication. Interacting cells are usually found located close to each other, which is not captured with scRNA sequencing [164]. Cellular localization is crucial to improve predictions of cell-cell interactions. Therefore, spatial transcriptomics is the obvious next step in such an analysis. Showing that there is a potential for interaction both in terms of expression of the pair and that the cells are in fact in close contact is extremely valuable. However, this is only applicable in tissues and solid tumors.

An additional avenue to achieve diversity is through alternative splicing of mRNAs giving rise to divergent isoforms from the same gene. Alternative splicing or differential transcript usage have shown to be a key feature cancers [231]. However, with tag-based scRNA sequencing protocols, it is not possible to derive such isoforms as this requires transcripts of full-length. Currently, there is a lot of effort being put into long-read sequencing methods [232]. This might impact scRNA sequencing methods facilitating more feasible full-length transcripts and thereby enabling isoform studies on a single-cell level.

Cancer is such a multi-faceted disease, and there will never be a one-size-fits-all solution pushing the need for personalized medicine. A multitude of factors need to be taken into consideration: the tumor itself, the composition of the tumor microenvironment especially immune cells, but also intra-tumoral microbioata and several external factors. As more and more data is being produced and high-resolution technologies developed, more advanced analytical methods tailored to this data type are required. However, all will contribute in piecing the small pieces of the cancer puzzle together. The work presented here is a small step in that direction.

References

[1] World's Health Organization https://www.who.int/news-room/fact-sheets/detail/cancer Accessed: 2022-17-08.

[2] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021): Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries *CA: a cancer journal for clinicians,* **71**(3), 209–249.

[3] Kræftens Bekæmpelse https://www.cancer.dk/hjaelp-viden/fakta-om-kraeft/kraeft-i-tal/nogletal/ Accessed: 2022-17-08.

[4] Hanahan, D. and Weinberg, R. A. (2000): The hallmarks of cancer *cell,* **100**(1), 57–70.

[5] Hanahan, D. and Weinberg, R. A. (2011): Hallmarks of cancer: the next generation *cell,* **144**(5), 646–674.

[6] Hanahan, D. (2022): Hallmarks of cancer: new dimensions *Cancer discovery,* **12**(1), 31–46.

[7] Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., and Campbell, P. J. (2017): Universal patterns of selection in cancer and somatic tissues *Cell,* **171**(5), 1029–1041.

[8] Greaves, M. and Maley, C. C. (2012): Clonal evolution in cancer *Nature,* **481**(7381), 306–313.

[9] Martincorena, I. and Campbell, P. J. (2015): Somatic mutation in cancer and normal cells *Science,* **349**(6255), 1483–1489.

[10] Wu, S., Powers, S., Zhu, W., and Hannun, Y. A. (2016): Substantial contribution of extrinsic risk factors to cancer development *Nature,* **529**(7584), 43–47.

[11] Blot, W. J., McLaughlin, J. K., Winn, D. M., Austin, D. F., Greenberg, R. S., Preston-Martin, S., Bernstein, L., Schoenberg, J. B., Stemhagen, A., and Fraumeni Jr, J. F. (1988): Smoking and drinking in relation to oral and pharyngeal cancer *Cancer research,* **48**(11), 3282–3287.

[12] Kamangar, F., Chow, W.-H., Abnet, C. C., and Dawsey, S. M. (2009): Environmental causes of esophageal cancer *Gastroenterology Clinics of North America,* **38**(1), 27–57.

[13] Parkin, D., Mesher, D., and Sasieni, P. (2011): 13. Cancers attributable to solar (ultraviolet) radiation exposure in the UK in 2010 *British journal of cancer,* **105**(2), S66–S69.

[14] Koh, H. K., Geller, A. C., Miller, D. R., Grossbart, T. A., and Lew, R. A. (1996): Prevention and early detection strategies for melanoma and skin cancer: current status *Archives of dermatology,* **132**(4), 436–443.

[15] Calabrese, C., Davidson, N. R., Demircioğlu, D., Fonseca, N. A., He, Y., Kahles, A., Lehmann, K.-V., Liu, F., Shiraishi, Y., Soulette, C. M., et al. (2020): Genomic basis for RNA alterations in cancer *Nature,* **578**(7793), 129–136.

[16] Palucka, A. K. and Coussens, L. M. (2016): The basis of oncoimmunology *Cell,* **164**(6), 1233–1247.

[17] Barkley, D., Moncada, R., Pour, M., Liberman, D. A., Dryg, I., Werba, G., Wang, W., Baron, M., Rao, A., Xia, B., et al. (2022): Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment *Nature Genetics,* **54**(8), 1192–1201.

[18] Anderson, N. M. and Simon, M. C. (2020): The tumor microenvironment *Current Biology,* **30**(16), R921–R925.

[19] Murphy, K. and Weaver, C. (2017): Janeway's immunobiology, Garland science 9th edition.

[20] Haas, L. and Obenauf, A. C. (2019): Allies or enemies—the multifaceted role of myeloid cells in the tumor microenvironment *Frontiers in immunology,* **10**, 2746.

[21] Joffre, O. P., Segura, E., Savina, A., and Amigorena, S. (2012): Cross-presentation by dendritic cells *Nature Reviews Immunology,* **12**(8), 557–569.

[22] Waldman, A. D., Fritz, J. M., and Lenardo, M. J. (2020): A guide to cancer immunotherapy: from T cell basic science to clinical practice *Nature Reviews Immunology,* **20**(11), 651–668.

[23] Parry, R. V., Chemnitz, J. M., Frauwirth, K. A., Lanfranco, A. R., Braunstein, I., Kobayashi, S. V., Linsley, P. S., Thompson, C. B., and Riley, J. L. (2005): CTLA-4 and PD-1 receptors inhibit T-cell activation by distinct mechanisms *Molecular and cellular biology,* **25**(21), 9543–9553.

[24] Wherry, E. J. and Kurachi, M. (2015): Molecular and cellular insights into T cell exhaustion *Nature Reviews Immunology,* **15**(8), 486–499.

[25] Kurachi, M. (2019) CD8+ T cell exhaustion In *Seminars in immunopathology* Springer Vol. 41, pp. 327–337.

[26] Jiang, W., He, Y., He, W., Wu, G., Zhou, X., Sheng, Q., Zhong, W., Lu, Y., Ding, Y., Lu, Q., et al. (2021): Exhausted CD8+ T cells in the tumor immune microenvironment: new pathways to therapy *Frontiers in immunology,* **11**, 622509.

[27] Zhang, Y. and Zhang, Z. (2020): The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications *Cellular & molecular immunology,* **17**(8), 807–821.

[28] Wang, N., Li, X., Wang, R., and Ding, Z. (2021): Spatial transcriptomics and proteomics technologies for deconvoluting the tumor microenvironment *Biotechnology Journal,* **16**(9), 2100041.

[29] Kennedy, L. B. and Salama, A. K. (2020): A review of cancer immunotherapy toxicity *CA: a cancer journal for clinicians,* **70**(2), 86–104.

[30] Abbott, M. and Ustoyev, Y. (2019) Cancer and the immune system: the history and background of immunotherapy In *Seminars in oncology nursing* Elsevier Vol. 35, p. 150923.

[31] Oiseth, S. J. and Aziz, M. S. (2017): Cancer immunotherapy: a brief review of the history, possibilities, and challenges ahead *Journal of cancer metastasis and treatment,* **3**, 250–261.

[32] Chen, D. S. and Mellman, I. (2013): Oncology meets immunology: the cancer-immunity cycle *immunity,* **39**(1), 1–10.

[33] Chiossone, L., Dumas, P.-Y., Vienne, M., and Vivier, E. (2018): Natural killer cells and other innate lymphoid cells in cancer *Nature Reviews Immunology,* **18**(11), 671–688.

[34] Mitchison, T. J. (2021): So many ways to naturally kill a cancer cell *BMC biology,* **19**(1), 1–3.

[35] Paijens, S. T., Vledder, A., de Bruyn, M., and Nijman, H. W. (2021): Tumor-infiltrating lymphocytes in the immunotherapy era *Cellular & molecular immunology,* **18**(4), 842–859.

[36] Demaria, O., Cornen, S., Daëron, M., Morel, Y., Medzhitov, R., and Vivier, E. (2019): Harnessing innate immunity in cancer therapy *Nature,* **574**(7776), 45–56.

[37] Tumeh, P. C., Harview, C. L., Yearley, J. H., Shintaku, I. P., Taylor, E. J., Robert, L., Chmielowski, B., Spasic, M., Henry, G., Ciobanu, V., et al. (2014): PD-1 blockade

induces responses by inhibiting adaptive immune resistance *Nature,* **515**(7528), 568–571.

[38] Crotty, S. (2015): A brief history of T cell help to B cells *Nature Reviews Immunology,* **15**(3), 185–189.

[39] Kroeger, D. R., Milne, K., and Nelson, B. H. (2016): Tumor-Infiltrating Plasma Cells Are Associated with Tertiary Lymphoid Structures, Cytolytic T-Cell Responses, and Superior Prognosis in Ovarian CancerPlasma Cells, CD8 T Cells, and Survival in Ovarian Cancer *Clinical Cancer Research,* **22**(12), 3005–3015.

[40] Helmink, B. A., Reddy, S. M., Gao, J., Zhang, S., Basar, R., Thakur, R., Yizhak, K., Sade-Feldman, M., Blando, J., Han, G., et al. (2020): B cells and tertiary lymphoid structures promote immunotherapy response *Nature,* **577**(7791), 549–555.

[41] Garaud, S., Buisseret, L., Solinas, C., Gu-Trantien, C., de Wind, A., Van den Eynden, G., Naveaux, C., Lodewyckx, J.-N., Boisson, A., Duvillier, H., et al. (2019): Tumor-infiltrating B cells signal functional humoral immune responses in breast cancer *JCI insight,* **4**(18).

[42] Berntsson, J., Nodin, B., Eberhard, J., Micke, P., and Jirström, K. (2016): Prognostic impact of tumour-infiltrating B cells and plasma cells in colorectal cancer *International Journal of Cancer,* **139**(5), 1129–1139.

[43] Sharma, P., Hu-Lieskovan, S., Wargo, J. A., and Ribas, A. (2017): Primary, adaptive, and acquired resistance to cancer immunotherapy *Cell,* **168**(4), 707–723.

[44] Iorgulescu, J. B., Braun, D., Oliveira, G., Keskin, D. B., and Wu, C. J. (2018): Acquired mechanisms of immune escape in cancer following immunotherapy *Genome medicine,* **10**(1), 1–4.

[45] Hiam-Galvez, K. J., Allen, B. M., and Spitzer, M. H. (2021): Systemic immunity in cancer *Nature reviews cancer,* **21**(6), 345–359.

[46] Sun, C., Mezzadra, R., and Schumacher, T. N. (2018): Regulation and function of the PD-L1 checkpoint *Immunity,* **48**(3), 434–452.

[47] Barber, D. L., Wherry, E. J., Masopust, D., Zhu, B., Allison, J. P., Sharpe, A. H., Freeman, G. J., and Ahmed, R. (2006): Restoring function in exhausted CD8 T cells during chronic viral infection *Nature,* **439**(7077), 682–687.

[48] Lentz, R. W., Colton, M. D., Mitra, S. S., and Messersmith, W. A. (2021): Innate Immune Checkpoint Inhibitors: The Next Breakthrough in Medical Oncology? Innate Immune Checkpoint Inhibitors in Medical Oncology *Molecular cancer therapeutics,* **20**(6), 961–974.

[49] Carlino, M. S., Larkin, J., and Long, G. V. (2021): Immune checkpoint inhibitors in melanoma *The Lancet,* **398**(10304), 1002–1014.

[50] Binnewies, M., Roberts, E. W., Kersten, K., Chan, V., Fearon, D. F., Merad, M., Coussens, L. M., Gabrilovich, D. I., Ostrand-Rosenberg, S., Hedrick, C. C., et al. (2018): Understanding the tumor immune microenvironment (TIME) for effective therapy *Nature medicine,* **24**(5), 541–550.

[51] Xu-Monette, Z. Y., Zhou, J., and Young, K. H. (2018): PD-1 expression and clinical PD-1 blockade in B-cell lymphomas *Blood, The Journal of the American Society of Hematology,* **131**(1), 68–83.

[52] Twomey, J. D. and Zhang, B. (2021): Cancer immunotherapy update: FDA-approved checkpoint inhibitors and companion diagnostics *The AAPS Journal,* **23**(2), 1–11.

[53] Wang, L., Livak, K. J., and Wu, C. J. (2018): High-dimension single-cell analysis applied to cancer *Molecular aspects of medicine,* **59**, 70–84.

[54] Morad, G., Helmink, B. A., Sharma, P., and Wargo, J. A. (2021): Hallmarks of response, resistance, and toxicity to immune checkpoint blockade *Cell,* **184**(21), 5309–5337.

[55] Coulie, P. G., Van den Eynde, B. J., Van Der Bruggen, P., and Boon, T. (2014): Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy *Nature Reviews Cancer,* **14**(2), 135–146.

[56] Yarchoan, M., Hopkins, A., and Jaffee, E. M. (2017): Tumor mutational burden and response rate to PD-1 inhibition *New England Journal of Medicine,* **377**(25), 2500–2501.

[57] Sha, D., Jin, Z., Budczies, J., Kluck, K., Stenzinger, A., and Sinicrope, F. A. (2020): Tumor mutational burden as a predictive biomarker in solid tumors *Cancer discovery,* **10**(12), 1808–1825.

[58] McGranahan, N., Furness, A. J., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., Jamal-Hanjani, M., Wilson, G. A., Birkbak, N. J., Hiley, C. T., et al. (2016): Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade *Science,* **351**(6280), 1463–1469.

[59] Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., Walsh, L. A., Postow, M. A., Wong, P., Ho, T. S., et al. (2014): Genetic basis for clinical response to CTLA-4 blockade in melanoma *New England Journal of Medicine,* **371**(23), 2189–2199.

[60] Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Geukes Foppen, M. H., Goldinger, S. M., et al. (2015): Genomic correlates of response to CTLA-4 blockade in metastatic melanoma *Science,* **350**(6257), 207–211.

[61] Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., et al. (2016): Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma *Cell,* **165**(1), 35–44.

[62] Taube, J. M., Klein, A., Brahmer, J. R., Xu, H., Pan, X., Kim, J. H., Chen, L., Pardoll, D. M., Topalian, S. L., and Anders, R. A. (2014): Association of PD-1, PD-1 Ligands, and Other Features of the Tumor Immune Microenvironment with Response to Anti–PD-

1 TherapyAssociation of PD-1 and Ligands with Response to Anti–PD-1 *Clinical cancer research,* **20**(19), 5064–5074.

[63] Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D. R., Albright, A., Cheng, J. D., Kang, S. P., Shankaran, V., et al. (2017): IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade *The Journal of clinical investigation,* **127**(8), 2930–2940.

[64] Sade-Feldman, M., Yizhak, K., Bjorgaard, S. L., Ray, J. P., de Boer, C. G., Jenkins, R. W., Lieb, D. J., Chen, J. H., Frederick, D. T., Barzily-Rokni, M., et al. (2018): Defining T cell states associated with response to checkpoint immunotherapy in melanoma *Cell,* **175**(4), 998–1013.

[65] Sade-Feldman, M., Jiao, Y. J., Chen, J. H., Rooney, M. S., Barzily-Rokni, M., Eliane, J.-P., Bjorgaard, S. L., Hammond, M. R., Vitzthum, H., Blackmon, S. M., et al. (2017): Resistance to checkpoint blockade therapy through inactivation of antigen presentation *Nature communications,* **8**(1), 1–11.

[66] Jiang, Z., Zhou, Y., and Huang, J. (2021): A Combination of Biomarkers Predict Response to Immune Checkpoint Blockade Therapy in Non-Small Cell Lung Cancer *Frontiers in immunology,* **12**.

[67] Le, D. T., Durham, J. N., Smith, K. N., Wang, H., Bartlett, B. R., Aulakh, L. K., Lu, S., Kemberling, H., Wilt, C., Luber, B. S., et al. (2017): Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade *Science,* **357**(6349), 409–413.

[68] Routy, B., Le Chatelier, E., Derosa, L., Duong, C. P., Alou, M. T., Daillère, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M. P., et al. (2018): Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors *Science,* **359**(6371), 91–97.

[69] Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M.-L., Luke, J. J., and Gajewski, T. F. (2018): The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients *Science,* **359**(6371), 104–108.

[70] Buzdin, A., Sorokin, M., Garazha, A., Glusker, A., Aleshin, A., Poddubskaya, E., Sekacheva, M., Kim, E., Gaifullin, N., Giese, A., et al. (2020) RNA sequencing for research and diagnostics in clinical oncology In *Seminars in Cancer Biology* Elsevier Vol. 60, pp. 311–323.

[71] Latha, N. R., Rajan, A., Nadhan, R., Achyutuni, S., Sengodan, S. K., Hemalatha, S. K., Varghese, G. R., Thankappan, R., Krishnan, N., Patra, D., et al. (2020): Gene expression signatures: A tool for analysis of breast cancer prognosis and therapy *Critical Reviews in Oncology/Hematology,* **151**, 102964.

[72] Wang, Z., Gerstein, M., and Snyder, M. (2009): RNA-Seq: a revolutionary tool for transcriptomics *Nature reviews genetics,* **10**(1), 57–63.

[73] Cieślik, M. and Chinnaiyan, A. M. (2018): Cancer transcriptome profiling at the juncture of clinical translation *Nature Reviews Genetics,* **19**(2), 93–109.

[74] Hanash, S. (2004): Integrated global profiling of cancer *Nature reviews Cancer,* **4**(8), 638–644.

[75] Stark, R., Grzelak, M., and Hadfield, J. (2019): RNA sequencing: the teenage years *Nature Reviews Genetics,* **20**(11), 631–656.

[76] Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., Xie, S.-J., Xiao, Z.-D., and Zhang, H. (2020): RNA sequencing: new technologies and applications in cancer research *Journal of hematology & oncology,* **13**(1), 1–16.

[77] Kulkarni, A., Anderson, A. G., Merullo, D. P., and Konopka, G. (2019): Beyond bulk: a review of single cell transcriptomics methodologies and applications *Current opinion in biotechnology,* **58**, 129–136.

[78] Lei, Y., Tang, R., Xu, J., Wang, W., Zhang, B., Liu, J., Yu, X., and Shi, S. (2021): Applications of single-cell sequencing in cancer research: progress and perspectives *Journal of Hematology & Oncology,* **14**(1), 1–26.

[79] (2014): Method of the Year 2013 *Nature Methods,* **11**(1), 1.

[80] Li, X. and Wang, C.-Y. (2021): From bulk, single-cell to spatial RNA sequencing *International Journal of Oral Science,* **13**(1), 1–6.

[81] Zhang, Y., Wang, D., Peng, M., Tang, L., Ouyang, J., Xiong, F., Guo, C., Tang, Y., Zhou, Y., Liao, Q., et al. (2021): Single-cell RNA sequencing in cancer research *Journal of Experimental & Clinical Cancer Research,* **40**(1), 1–17.

[82] Baslan, T. and Hicks, J. (2017): Unravelling biology and shifting paradigms in cancer with single-cell sequencing *Nature Reviews Cancer,* **17**(9), 557–569.

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

[84] Hu, X. and Zhou, X. (2022): Impact of single-cell RNA sequencing on understanding immune regulation *Journal of Cellular and Molecular Medicine,* **26**(17), 4645–4657.

[85] Brodin, P. and Davis, M. M. (2017): Human immune system variation *Nature reviews immunology,* **17**(1), 21–29.

[86] Erfanian, N., Derakhshani, A., Nasseri, S., Fereidouni, M., Baradaran, B., Tabrizi, N. J., Brunetti, O., Bernardini, R., Silvestris, N., and Safarpour, H. (2022): Immunotherapy of cancer in single-cell RNA sequencing era: A precision medicine perspective *Biomedicine & Pharmacotherapy,* **146**, 112558.

[87] Hedlund, E. and Deng, Q. (2018): Single-cell RNA sequencing: technical advancements and biological applications *Molecular aspects of medicine,* **59**, 36–46.

[88] Prakadan, S. M., Shalek, A. K., and Weitz, D. A. (2017): Scaling by shrinking: empowering single-cell'omics' with microfluidic devices *Nature Reviews Genetics,* **18**(6), 345–361.

[89] Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., and Wang, J. (2019): Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems *Molecular cell,* **73**(1), 130–142.

[90] Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017): Massively parallel digital transcriptional profiling of single cells *Nature communications,* **8**(1), 1–12.

[91] Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015): Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells *Cell,* **161**(5), 1187–1201.

[92] Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015): Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets *Cell,* **161**(5), 1202–1214.

[93] Grün, D. and van Oudenaarden, A. (2015): Design and analysis of single-cell sequencing experiments *Cell,* **163**(4), 799–810.

[94] Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013): Smart-seq2 for sensitive full-length transcriptome profiling in single cells *Nature methods,* **10**(11), 1096–1098.

[95] Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A. J., Faridani, O. R., and Sandberg, R. (2020): Single-cell RNA counting at allele and isoform resolution using Smart-seq3 *Nature Biotechnology,* **38**(6), 708–714.

[96] Gong, B., Zhou, Y., and Purdom, E. (2021): Cobolt: integrative analysis of multimodal single-cell sequencing data *Genome biology,* **22**(1), 1–21.

[97] Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017): Simultaneous epitope and transcriptome measurement in single cells *Nature methods,* **14**(9), 865–868.

[98] Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., et al. (2018): scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells *Nature communications,* **9**(1), 1–9.

[99] Longo, S. K., Guo, M. G., Ji, A. L., and Khavari, P. A. (2021): Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics *Nature Reviews Genetics,* **22**(10), 627–644.

[100] Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., et al. (2016): Visualization and analysis of gene expression in tissue sections by spatial transcriptomics *Science,* **353**(6294), 78–82.

[101] Rao, A., Barkley, D., França, G. S., and Yanai, I. (2021): Exploring tissue architecture using spatial transcriptomics *Nature,* **596**(7871), 211–220.

[102] Ji, A. L., Rubin, A. J., Thrane, K., Jiang, S., Reynolds, D. L., Meyers, R. M., Guo, M. G., George, B. M., Mollbrink, A., Bergenstråhle, J., et al. (2020): Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma *Cell,* **182**(2), 497–514.

[103] Gohil, S. H., Iorgulescu, J. B., Braun, D. A., Keskin, D. B., and Livak, K. J. (2021): Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy *Nature Reviews Clinical Oncology,* **18**(4), 244–256.

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

[105] Love, M. I., Huber, W., and Anders, S. (2014): Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 *Genome biology,* **15**(12), 1–21.

[106] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010): edgeR: a Bioconductor package for differential expression analysis of digital gene expression data *bioinformatics,* **26**(1), 139–140.

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

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

[109] Xu, G., Liu, Y., Li, H., Liu, L., Zhang, S., and Zhang, Z. (2020): Dissecting the human immune system with single cell RNA sequencing technology *Journal of leukocyte biology,* **107**(4), 613–623.

[110] Andrews, T. S., Kiselev, V. Y., McCarthy, D., and Hemberg, M. (2021): Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data *Nature protocols,* **16**(1), 1–9.

[111] Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., and Luo, Y. (2022): Single-cell RNA sequencing technologies and applications: A brief overview *Clinical and Translational Medicine,* **12**(3), e694.

[112] Luecken, M. D. and Theis, F. J. (2019): Current best practices in single-cell RNA-seq analysis: a tutorial *Molecular systems biology,* **15**(6), e8746.

[113] Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. (2016): Classification of low quality cells from single-cell RNA-seq data *Genome biology,* **17**(1), 1–15.

[114] Wolock, S. L., Lopez, R., and Klein, A. M. (2019): Scrublet: computational identification of cell doublets in single-cell transcriptomic data *Cell systems,* **8**(4), 281–291.

[115] DePasquale, E. A., Schnell, D. J., Van Camp, P.-J., Valiente-Alandí, Í., Blaxall, B. C., Grimes, H. L., Singh, H., and Salomonis, N. (2019): DoubletDecon: deconvoluting doublets from single-cell RNA-sequencing data *Cell reports,* **29**(6), 1718–1727.

[116] McGinnis, C. S., Murrow, L. M., and Gartner, Z. J. (2019): DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors *Cell systems,* **8**(4), 329–337.

[117] Xi, N. M. and Li, J. J. (2021): Benchmarking computational doublet-detection methods for single-cell RNA sequencing data *Cell systems,* **12**(2), 176–194.

[118] Hafemeister, C. and Satija, R. (2019): Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression *Genome biology,* **20**(1), 1–15.

[119] Johansen, N. and Quon, G. (2019): scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data *Genome biology,* **20**(1), 1–21.

[120] McInnes, L., Healy, J., and Melville, J. (2018): Umap: Uniform manifold approximation and projection for dimension reduction *arXiv preprint arXiv:1802.03426,*.

[121] Van der Maaten, L. and Hinton, G. (2008): Visualizing data using t-SNE. *Journal of machine learning research,* **9**(11).

[122] Tang, J., Liu, J., Zhang, M., and Mei, Q. (2016) Visualizing large-scale and high-dimensional data In *Proceedings of the 25th international conference on world wide web* pp. 287–297.

[123] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019): Comprehensive integration of single-cell data *Cell,* **177**(7), 1888–1902.

[124] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021): Integrated analysis of multimodal single-cell data *Cell,* **184**(13), 3573–3587.

[125] Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., et al. (2019): Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage *Nature immunology,* **20**(2), 163–172.

[126] Zhang, Z., Cui, F., Lin, C., Zhao, L., Wang, C., and Zou, Q. (2021): Critical downstream analysis steps for single-cell RNA sequencing data *Briefings in Bioinformatics,* **22**(5), bbab105.

[127] Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014): Bayesian approach to single-cell differential expression analysis *Nature methods,* **11**(7), 740–742.

[128] Rosati, D. and Giordano, A. (2021): Single-cell RNA sequencing and bioinformatics as tools to decipher cancer heterogenicity and mechanisms of drug resistance *Biochemical Pharmacology,* p. 114811.

[129] Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018): Using single-cell genomics to understand developmental processes and cell fate decisions *Molecular systems biology,* **14**(4), e8046.

[130] Weiler, P., Van den Berge, K., Street, K., and Tiberi, S. (2021): A guide to trajectory inference and RNA velocity *bioRxiv,*.

[131] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014): Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions *Nature biotechnology,* **32**(4), 381.

[132] Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016): Diffusion pseudotime robustly reconstructs lineage branching *Nature methods,* **13**(10), 845–848.

[133] Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'er, D. (2016): Wishbone identifies bifurcating developmental trajectories from single-cell data *Nature biotechnology,* **34**(6), 637–645.

[134] Cannoodt, R., Saelens, W., and Saeys, Y. (2016): Computational methods for trajectory inference from single-cell transcriptomics *European journal of immunology,* **46**(11), 2496–2506.

[135] Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019): A comparison of single-cell trajectory inference methods *Nature biotechnology,* **37**(5), 547–554.

[136] Liu, Z., Lou, H., Xie, K., Wang, H., Chen, N., Aparicio, O. M., Zhang, M. Q., Jiang, R., and Chen, T. (2017): Reconstructing cell cycle pseudo time-series via single-cell transcriptome data *Nature communications,* **8**(1), 1–9.

[137] Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F. J. (2019): PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells *Genome biology,* **20**(1), 1–9.

[138] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. (2018): RNA velocity of single cells *Nature,* **560**(7719), 494–498.

[139] Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020): Generalizing RNA velocity to transient cell states through dynamical modeling *Nature biotechnology,* **38**(12), 1408–1414.

[140] Barkas, N., Petukhov, V., Kharchenko, P., and Biederstedt, E. (2021): pagoda2: single cell analysis and differential expression *R package version,* **102**.

[141] Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K., and Kharchenko, P. V. (2019): Joint analysis of heterogeneous single-cell RNA-seq dataset collections *Nature methods,* **16**(8), 695–698.

[142] Wolf, F. A., Angerer, P., and Theis, F. J. (2018): SCANPY: large-scale single-cell gene expression data analysis *Genome biology,* **19**(1), 1–5.

[143] Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., et al. (2019): The single-cell transcriptional landscape of mammalian organogenesis *Nature,* **566**(7745), 496–502.

[144] Kipps, T. J., Stevenson, F. K., Wu, C. J., Croce, C. M., Packham, G., Wierda, W. G., O'brien, S., Gribben, J., and Rai, K. (2017): Chronic lymphocytic leukaemia *Nature reviews Disease primers,* **3**(1), 1–22.

[145] Rhodes, J. M. and Barrientos, J. C. (12, 2020): Chemotherapy-free frontline therapy for CLL: is it worth it? *Hematology,* **2020**(1), 24–32.

[146] Patel, K. and Pagel, J. M. (2021): Current and future treatment strategies in chronic lymphocytic leukemia *Journal of Hematology & Oncology,* **14**(1), 1–20.

[147] Natalia, T. and Varsha, G. (2021): Ibrutinib combinations in CLL therapy: scientific rationale and clinical results *Blood Cancer Journal,* **11**(4).

[148] Dagklis, A., Fazi, C., Scarfò, L., Apollonio, B., and Ghia, P. (2009): Monoclonal B lymphocytosis in the general population *Leukemia & lymphoma,* **50**(3), 490–492.

[149] Rawstron, A. C., Bennett, F. L., O'Connor, S. J., Kwok, M., Fenton, J. A., Plummer, M., de Tute, R., Owen, R. G., Richards, S. J., Jack, A. S., et al. (2008): Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia *New England Journal of Medicine,* **359**(6), 575–583.

[150] Landgren, O., Albitar, M., Ma, W., Abbasi, F., Hayes, R. B., Ghia, P., Marti, G. E., and Caporaso, N. E. (2009): B-cell clones as early markers for chronic lymphocytic leukemia *New England Journal of Medicine,* **360**(7), 659–667.

[151] Strati, P. and Shanafelt, T. D. (2015): Monoclonal B-cell lymphocytosis and early-stage chronic lymphocytic leukemia: diagnosis, natural history, and risk stratification *Blood, The Journal of the American Society of Hematology,* **126**(4), 454–462.

[152] Calissano, C., Damle, R. N., Hayes, G., Murphy, E. J., Hellerstein, M. K., Moreno, C., Sison, C., Kaufman, M. S., Kolitz, J. E., Allen, S. L., et al. (2009): In vivo intraclonal and interclonal kinetic heterogeneity in B-cell chronic lymphocytic leukemia *Blood, The Journal of the American Society of Hematology,* **114**(23), 4832–4842.

[153] Purroy, N. and Wu, C. J. (2017): Coevolution of leukemia and host immune cells in chronic lymphocytic leukemia *Cold Spring Harbor perspectives in medicine,* **7**(4), a026740.

[154] Arruga, F., Gyau, B. B., Iannello, A., Vitale, N., Vaisitti, T., and Deaglio, S. (2020): Immune response dysfunction in chronic lymphocytic leukemia: dissecting molecular mechanisms and microenvironmental conditions *International journal of molecular sciences,* **21**(5), 1825.

[155] Long, M., Beckwith, K., Do, P., Mundy, B. L., Gordon, A., Lehman, A. M., Maddocks, K. J., Cheney, C., Jones, J. A., Flynn, J. M., et al. (2017): Ibrutinib treatment improves T cell number and function in CLL patients *The Journal of clinical investigation,* **127**(8), 3052–3064.

[156] Mhibik, M., Wiestner, A., and Sun, C. (2019): Harnessing the effects of BTKi on T cells for effective immunotherapy against CLL *International Journal of Molecular Sciences,* **21**(1), 68.

[157] MUDAN https://github.com/JEFworks/MUDAN.

[158] Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019): Fast, sensitive and accurate integration of single-cell data with Harmony *Nature methods,* **16**(12), 1289–1296.

[159] Denisenko, E., Guo, B. B., Jones, M., Hou, R., De Kock, L., Lassmann, T., Poppe, D., Clément, O., Simmons, R. K., Lister, R., et al. (2020): Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows *Genome biology,* **21**(1), 1–25.

[160] van den Brink, S. C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C. S., Robin, C., and Van Oudenaarden, A. (2017): Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations *Nature methods,* **14**(10), 935–936.

[161] Hanamsagar, R., Reizis, T., Chamberlain, M., Marcus, R., Nestle, F. O., de Rinaldis, E., and Savova, V. (2020): An optimized workflow for single-cell transcriptomics and repertoire profiling of purified lymphocytes from clinical samples *Scientific reports,* **10**(1), 1–15.

[162] Massoni-Badosa, R., Iacono, G., Moutinho, C., Kulis, M., Palau, N., Marchese, D., Rodríguez-Ubreva, J., Ballestar, E., Rodriguez-Esteban, G., Marsal, S., et al. (2020): Sampling time-dependent artifacts in single-cell genomics studies *Genome biology,* **21**(1), 1–16.

[163] Baechler, E., Batliwalla, F., Karypis, G., Gaffney, P., Moser, K., Ortmann, W., Espe, K., Balasubramanian, S., Hughes, K., Chan, J., et al. (2004): Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation *Genes & Immunity,* **5**(5), 347–353.

[164] Armingol, E., Officer, A., Harismendy, O., and Lewis, N. E. (2021): Deciphering cell–cell interactions and communication from gene expression *Nature Reviews Genetics,* **22**(2), 71–88.

[165] Buccitelli, C. and Selbach, M. (2020): mRNAs, proteins and the emerging principles of gene expression control *Nature Reviews Genetics,* **21**(10), 630–644.

[166] Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., Park, J.-E., Stephenson, E., Polański, K., Goncalves, A., et al. (2018): Single-cell reconstruction of the early maternal–fetal interface in humans *Nature,* **563**(7731), 347–353.

[167] Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2020): CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes *Nature protocols,* **15**(4), 1484–1506.

[168] Parikh, S. A., Kay, N. E., and Shanafelt, T. D. (2014): How we treat Richter syndrome *Blood, The Journal of the American Society of Hematology,* **123**(11), 1647–1657.

[169] Ding, W. (2018): Richter transformation in the era of novel agents *Hematology 2014, the American Society of Hematology Education Program Book,* **2018**(1), 256–263.

[170] Ding, W., LaPlant, B. R., Call, T. G., Parikh, S. A., Leis, J. F., He, R., Shanafelt, T. D., Sinha, S., Le-Rademacher, J., Feldman, A. L., et al. (2017): Pembrolizumab in patients with CLL and Richter transformation or with relapsed CLL *Blood, The Journal of the American Society of Hematology,* **129**(26), 3419–3427.

[171] Jain, N., Ferrajoli, A., Basu, S., Thompson, P. A., Burger, J. A., Kadia, T. M., Estrov, Z. E., Pemmaraju, N., Lopez, W., Thakral, B., et al. (2018): A phase II trial of nivolumab combined with ibrutinib for patients with Richter transformation *Blood,* **132**, 296.

[172] Younes, A., Brody, J., Carpio, C., Lopez-Guillermo, A., Ben-Yehuda, D., Ferhanoglu, B., Nagler, A., Ozcan, M., Avivi, I., Bosch, F., et al. (2019): Safety and activity of ibrutinib in combination with nivolumab in patients with relapsed non-Hodgkin lymphoma or chronic lymphocytic leukaemia: a phase 1/2a study *The Lancet Haematology,* **6**(2), e67–e78.

[173] Oetjen, K. A., Lindblad, K. E., Goswami, M., Gui, G., Dagur, P. K., Lai, C., Dillon, L. W., McCoy, J. P., and Hourigan, C. S. (2018): Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry *JCI insight,* **3**(23).

[174] Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A., and Teichmann, S. A. (2017): The Human Cell Atlas: from vision to reality *Nature,* **550**(7677), 451–453.

[175] Oliveira, G., Stromhaug, K., Klaeger, S., Kula, T., Frederick, D. T., Le, P. M., Forman, J., Huang, T., Li, S., Zhang, W., et al. (2021): Phenotype, specificity and avidity of antitumour CD8+ T cells in melanoma *Nature,* **596**(7870), 119–125.

[176] Borcherding, N., Bormann, N. L., and Kraus, G. (2020): scRepertoire: An R-based toolkit for single-cell immune receptor analysis *F1000Research,* **9**.

[177] Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018): Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics *BMC genomics,* **19**(1), 1–16.

[178] Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. (2015): Robust enumeration of cell subsets from tissue expression profiles *Nature methods,* **12**(5), 453–457.

[179] Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., et al. (2019): Determining cell type abundance and expression from bulk tissues with digital cytometry *Nature biotechnology,* **37**(7), 773–782.

[180] Aran, D., Hu, Z., and Butte, A. J. (2017): xCell: digitally portraying the tissue cellular heterogeneity landscape *Genome biology,* **18**(1), 1–14.

[181] Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., et al. (2009): Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1 *Nature,* **462**(7269), 108–112.

[182] Skene, P. J., Henikoff, J. G., and Henikoff, S. (2018): Targeted in situ genome-wide profiling with high efficiency for low cell numbers *Nature protocols,* **13**(5), 1006–1019.

[183] Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., et al. (2018): The Encyclopedia of DNA elements (ENCODE): data portal update *Nucleic acids research,* **46**(D1), D794–D801.

[184] Li, S., Wan, C., Zheng, R., Fan, J., Dong, X., Meyer, C. A., and Liu, X. S. (2019): Cistrome-GO: a web server for functional enrichment analysis of transcription factor ChIP-seq peaks *Nucleic acids research,* **47**(W1), W206–W211.

[185] Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015): ATAC-seq: a method for assaying chromatin accessibility genome-wide *Current protocols in molecular biology,* **109**(1), 21–29.

[186] Zheng, L., Qin, S., Si, W., Wang, A., Xing, B., Gao, R., Ren, X., Wang, L., Wu, X., Zhang, J., et al. (2021): Pan-cancer single-cell landscape of tumor-infiltrating T cells *Science,* **374**(6574), abe6474.

[187] Liu, D., Schilling, B., Liu, D., Sucker, A., Livingstone, E., Jerby-Arnon, L., Zimmer, L., Gutzmer, R., Satzger, I., Loquai, C., et al. (2019): Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma *Nature medicine,* **25**(12), 1916–1927.

[188] Fairfax, B. P., Taylor, C. A., Watson, R. A., Nassiri, I., Danielli, S., Fang, H., Mahé, E. A., Cooper, R., Woodcock, V., Traill, Z., et al. (2020): Peripheral CD8+ T cell characteristics associated with durable responses to immune checkpoint blockade in patients with metastatic melanoma *Nature medicine,* **26**(2), 193–199.

[189] Caushi, J. X., Zhang, J., Ji, Z., Vaghasia, A., Zhang, B., Hsiue, E. H.-C., Mog, B. J., Hou, W., Justesen, S., Blosser, R., et al. (2021): Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers *Nature,* **596**(7870), 126–132.

[190] Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., et al. (2010): A comprehensive catalogue of somatic mutations from a human cancer genome *Nature,* **463**(7278), 191–196.

[191] Fabbri, G., Rasi, S., Rossi, D., Trifonov, V., Khiabanian, H., Ma, J., Grunn, A., Fangazio, M., Capello, D., Monti, S., et al. (2011): Analysis of the chronic lymphocytic leukemia coding genome: role of NOTCH1 mutational activation *Journal of Experimental Medicine,* **208**(7), 1389–1401.

[192] Puente, X. S., Pinyol, M., Quesada, V., Conde, L., Ordóñez, G. R., Villamor, N., Escaramis, G., Jares, P., Beà, S., González-Díaz, M., et al. (2011): Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia *Nature,* **475**(7354), 101–105.

[193] Wang, L., Lawrence, M. S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D. S., Zhang, L., et al. (2011): SF3B1 and other novel cancer genes in chronic lymphocytic leukemia *New England Journal of Medicine,* **365**(26), 2497–2506.

[194] Kwok, M. and Wu, C. J. (2021): Clonal evolution of high-risk chronic lymphocytic leukemia: a contemporary perspective *Frontiers in oncology,* **11**.

[195] Black, J. R. and McGranahan, N. (2021): Genetic and non-genetic clonal diversity in cancer evolution *Nature Reviews Cancer,* **21**(6), 379–392.

[196] Nam, A. S., Chaligne, R., and Landau, D. A. (2021): Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics *Nature Reviews Genetics,* **22**(1), 3–18.

[197] Wu, Y., Sarkissyan, M., and Vadgama, J. V. (2015): Epigenetics in breast and prostate cancer *Cancer Epigenetics,* pp. 425–466.

[198] Dawson, M. A. and Kouzarides, T. (2012): Cancer epigenetics: from mechanism to therapy *cell,* **150**(1), 12–27.

[199] Feinberg, A. P. and Tycko, B. (2004): The history of cancer epigenetics *Nature Reviews Cancer,* **4**(2), 143–153.

[200] Pan, Y., Liu, G., Zhou, F., Su, B., and Li, Y. (2018): DNA methylation profiles in cancer diagnosis and therapeutics *Clinical and experimental medicine,* **18**(1), 1–14.

[201] Yuasa, Y. (2002): DNA methylation in cancer and ageing *Mechanisms of ageing and development,* **123**(12), 1649–1654.

[202] Cahill, N., Bergh, A.-C., Kanduri, M., Göransson-Kultima, H., Mansouri, L., Isaksson, A., Ryan, F., Smedby, K., Juliusson, G., Sundström, C., et al. (2013): 450K-array analysis of chronic lymphocytic leukemia cells reveals global DNA methylation to be relatively stable over time and similar in resting and proliferative compartments *Leukemia,* **27**(1), 150–158.

[203] Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R. P., Hong, C., Downey, S. L., Johnson, B. E., Fouse, S. D., Delaney, A., Zhao, Y., et al. (2010): Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications *Nature biotechnology,* **28**(10), 1097–1105.

[204] Kulis, M., Heath, S., Bibikova, M., Queiros, A. C., Navarro, A., Clot, G., Martinez-Trillos, A., Castellano, G., Brun-Heath, I., Pinyol, M., et al. (2012): Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia *Nature genetics,* **44**(11), 1236–1242.

[205] Pei, L., Choi, J.-H., Liu, J., Lee, E.-J., McCarthy, B., Wilson, J. M., Speir, E., Awan, F., Tae, H., Arthur, G., et al. (2012): Genome-wide DNA methylation analysis reveals novel epigenetic changes in chronic lymphocytic leukemia *Epigenetics,* **7**(6), 567–578.

[206] Landau, D. A., Clement, K., Ziller, M. J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., Li, S., et al. (2014): Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia *Cancer cell,* **26**(6), 813–825.

[207] Penter, L., Gohil, S. H., Lareau, C., Ludwig, L. S., Parry, E. M., Huang, T., Li, S., Zhang, W., Livitz, D., Leshchiner, I., et al. (2021): Longitudinal Single-Cell Dynamics of Chromatin Accessibility and Mitochondrial Mutations in Chronic Lymphocytic Leukemia Mirror Disease HistoryLongitudinal Chromatin Evolution and mtDNA Mutations in CLL *Cancer Discovery,* **11**(12), 3048–3063.

[208] Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., Munar, M., Rubio-Pérez, C., Jares, P., Aymerich, M., et al. (2015): Non-coding recurrent mutations in chronic lymphocytic leukaemia *Nature,* **526**(7574), 519–524.

[209] Chigrinova, E., Rinaldi, A., Kwee, I., Rossi, D., Rancoita, P. M., Strefford, J. C., Oscier, D., Stamatopoulos, K., Papadaki, T., Berger, F., et al. (2013): Two main genetic pathways lead to the transformation of chronic lymphocytic leukemia to Richter syndrome *Blood, The Journal of the American Society of Hematology,* **122**(15), 2673–2682.

[210] Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Böttcher, S., et al. (2015): Mutations driving CLL and their evolution in progression and relapse *Nature,* **526**(7574), 525–530.

[211] Fabbri, G., Khiabanian, H., Holmes, A. B., Wang, J., Messina, M., Mullighan, C. G., Pasqualucci, L., Rabadan, R., and Dalla-Favera, R. (2013): Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome *Journal of Experimental Medicine,* **210**(11), 2273–2288.

[212] Klintman, J., Appleby, N., Stamatopoulos, B., Ridout, K., Eyre, T. A., Robbe, P., Pascua, L. L., Knight, S. J., Dreau, H., Cabes, M., et al. (2021): Genomic and transcriptomic correlates of Richter transformation in chronic lymphocytic leukemia *Blood,* **137**(20), 2800–2816.

[213] Rossi, D., Spina, V., Deambrogi, C., Rasi, S., Laurenti, L., Stamatopoulos, K., Arcaini, L., Lucioni, M., Rocque, G. B., Xu-Monette, Z. Y., et al. (2011): The genetics of Richter syndrome reveals disease heterogeneity and predicts survival after transformation *Blood, The Journal of the American Society of Hematology,* **117**(12), 3391–3401.

[214] inferCNV of the Trinity CTAT Project https://github.com/broadinstitute/inferCNV.

[215] Fleming, S., Marioni, J., and Babadi, M. (2019): CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. bioRxiv *Preprint at https://doi. org/10.1101/791699,*.

[216] Young, M. D. and Behjati, S. (2020): SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data *Gigascience,* **9**(12), giaa151.

[217] Kowalczyk, M. S., Tirosh, I., Heckl, D., Rao, T. N., Dixit, A., Haas, B. J., Schneider, R. K., Wagers, A. J., Ebert, B. L., and Regev, A. (2015): Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells *Genome research,* **25**(12), 1860–1872.

[218] Lauer, S. and Gresham, D. (2019): An evolving view of copy number variants *Current Genetics,* **65**(6), 1287–1295.

[219] Nord, A., Salipante, S. J., and Pritchard, C. (2015): Copy Number Variant Detection Using Next-Generation Sequencing In *Clinical Genomics* pp. 165–187, Elsevier.

[220] Sathirapongsasuti, J. F., Lee, H., Horst, B. A., Brunner, G., Cochran, A. J., Binder, S., Quackenbush, J., and Nelson, S. F. (2011): Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV *Bioinformatics,* **27**(19), 2648–2654.

[221] Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., Quinlan, A. R., Nickerson, D. A., Eichler, E. E., Project, N. E. S., et al. (2012): Copy number variation detection and genotyping from exome sequence data *Genome research,* **22**(8), 1525–1532.

[222] de Araújo Lima, L. and Wang, K. (2017): PennCNV in whole-genome sequencing data *BMC bioinformatics,* **18**(11), 49–56.

[223] Serin Harmanci, A., Harmanci, A. O., and Zhou, X. (2020): CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data *Nature communications,* **11**(1), 1–16.

[224] Fan, J., Lee, H.-O., Lee, S., Ryu, D.-e., Lee, S., Xue, C., Kim, S. J., Kim, K., Barkas, N., Park, P. J., et al. (2018): Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data *Genome research,* **28**(8), 1217–1227.

[225] CNVsingle, Broad Institute https://github.com/broadinstitute/CNVsingle.

[226] Ramsay, A. G., Johnson, A. J., Lee, A. M., Gorgün, G., Le Dieu, R., Blum, W., Byrd, J. C., Gribben, J. G., et al. (2008): Chronic lymphocytic leukemia T cells show impaired immunological synapse formation that can be reversed with an immunomodulating drug *The Journal of clinical investigation,* **118**(7), 2427–2437.

[227] Soilleux, E. J., Wotherspoon, A., Eyre, T. A., Clifford, R., Cabes, M., and Schuh, A. H. (2016): Diagnostic dilemmas of high-grade transformation (Richter's syndrome) of chronic lymphocytic leukaemia: results of the phase II National Cancer Research Institute CHOP-OR clinical trial specialist haemato-pathology central review *Histopathology,* **69**(6), 1066–1076.

[228] Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Müller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022): Benchmarking atlas-level data integration in single-cell genomics *Nature methods,* **19**(1), 41–50.

[229] Consortium*, T. S., Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O., Quake, S. R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., et al. (2022): The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans *Science,* **376**(6594), eabl4896.

[230] Liu, Y., Beyer, A., and Aebersold, R. (2016): On the dependency of cellular protein levels on mRNA abundance *Cell,* **165**(3), 535–550.

[231] Vitting-Seerup, K. and Sandelin, A. (2017): The Landscape of Isoform Switches in Human CancersIsoform Switches in Cancer *Molecular Cancer Research,* **15**(9), 1206–1220.

[232] Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018): The third revolution in sequencing technology *Trends in Genetics,* **34**(9), 666–681.

PAPER I

Single cell analysis reveals immune dysfunction from the earliest stages of CLL that can be reversed by ibrutinib

Purroy, N. Z.*, Tong, Y. E.*, **Lemvigh, C. K.***, Cieri, N., Li, S., Parry, E. M., Zhang, W., Rassenti, L. Z., Kipps, T. J., Slager, S. L., Kay, N. E., Lesnick, C., Shanafelt, T. D., Ghia, P., Scarfò, L., Livak, K. J., Kharchenko, P. V., Neuberg, D., Olsen, L. R., Fan, J., Gohil, S. H., Wu, C. J.

Accepted, Blood, 2022

The work carried out in relation to this thesis included an extensive preliminary analysis of the scRNA sequencing data presented here. The work comprised investigation ways of combining batches, initial interaction predictions, identification of cold shock signature and interpretation of results.

* = equal contribution

blood®
Letter to *Blood*

TO THE EDITOR:

Single-cell analysis reveals immune dysfunction from the earliest stages of CLL that can be reversed by ibrutinib

Noelia Purroy,[1-3,*] Yuzhou Evelyn Tong,[2-4,*] Camilla K. Lemvigh,[1,5,*] Nicoletta Cieri,[1-3] Shuqiang Li,[3,6] Erin M. Parry,[1-3] Wandi Zhang,[1] Laura Z. Rassenti,[7] Thomas J. Kipps,[7] Susan L. Slager,[8] Neil E. Kay,[8,9] Connie Lesnick,[9] Tait D. Shanafelt,[10] Paolo Ghia,[11] Lydia Scarfò,[11] Kenneth J. Livak,[1,6] Peter V. Kharchenko,[12] Donna S. Neuberg,[13] Lars Rønn Olsen,[4,5] Jean Fan,[14] Satyen H. Gohil,[1-3,15] and Catherine J. Wu[1-3,16]

[1]Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA; [2]Harvard Medical School, Boston, MA; [3]Broad Institute, Cambridge, MA; [4]Program in Health Sciences and Technology, Harvard Medical School–Massachusetts Institute of Technology, Boston, MA; [5]Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark; [6]Translational Immunogenomics Laboratory, Dana Farber Cancer Institute, Boston, MA; [7]Moores Cancer Center, University of California San Diego, La Jolla, CA; [8]Department of Health Sciences Research and [9]Department of Medicine, Mayo Clinic, Rochester, MN; [10]Department of Medicine, Stanford University, Stanford, CA; [11]Division of Experimental Oncology, Department of Onco-Hematology, Università Vita-Salute San Raffaele–Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Ospedale San Raffaele, Milan Italy; [12]Department of Biomedical Informatics, Harvard Medical School, Boston, MA; [13]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA; [14]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD; [15]Department of Academic Haematology, University College London, United Kingdom; and [16]Division of Hematology, Department of Medicine, Brigham and Women's Hospital, Boston, MA

Chronic lymphocytic leukemia (CLL) is characterized by a clonal expansion of mature $CD19^+CD5^+$ B cells, which are highly dependent on microenvironmental cues for their survival.[1] This common adult leukemia is preceded by a precursor phase termed monoclonal B-cell lymphocytosis (MBL),[2,3] which has been characterized as indistinguishable from CLL at the genetic, transcriptomic, and epigenomic level.[4-6] However, how leukemia cells coevolve with immune cells in their circulating microenvironment during the onset of MBL and upon progression to CLL remains incompletely characterized.[7]

Recently, single-cell transcriptome sequencing (scRNA-seq) approaches have transformed our ability to gain a comprehensive evaluation of the spectrum of immune cells within the tumor microenvironment and of their potential cross talk with cancer cells.[8-14] In our study, we applied scRNA-seq to broadly characterize circulating immune cells coexisting with leukemic cells during natural CLL progression. Although we acknowledge the critical role of the bone marrow and lymph node microenvironments on CLL cells, the lack of feasibility for procuring serial specimens from these tissue compartments led us to focus our study on circulating immune cells. We therefore collected serial peripheral blood mononuclear cell (PBMC) samples from 3 individuals with high-count MBL who did not progress to CLL after a median follow-up of 7.0 years and 7 patients with CLL, whose genetic characterization of $CD19^+CD5^+$ cells over time by whole-exome sequencing, has been reported[15] (Figure 1A). We processed paired samples from all patients: the first samples were collected at time point 1 (T1), at a median of 4.96 years (range, 2.44-5.46) after MBL diagnosis or 2.54 years (range, 0.5-4.2) after CLL diagnosis; whereas the second group were collected at T2, a median of 2.97 years (range, 2.01-2.99) after T1 for the MBL patients and 4.75 years (range, 1.3-10.6) for the CLL patients. T2 samples for CLL patients were collected at a median of 0.2 years (range, 0-5.9) before the first treatment (supplemental Table 1, available on the *Blood* Web site).

Non-$CD19^+CD5^+$ cells were isolated by fluorescence-activated cell sorting, and samples from each patient were processed on the same day to minimize the batch effect. Cell suspensions were loaded on a GemCode Single-Cell Instrument (10× Genomics), and libraries were prepared as previously described[16] (supplemental Methods). Analysis was conducted using Seurat V4.0.0 selecting cells with gene count between 500 and 3000 and less than 10% mitochondrial reads. Using the trimmed data set, we isolated the nontumor population and assigned immune cell types by performing multimodal reference mapping, using a CITE-seq (cellular indexing of transcriptomes and epitope–sequencing) reference of 162 000 PBMCs measured with 228 antibodies.[17] B cells were excluded because of potential CLL contamination. After quality control, we obtained 67 333 single-cell transcriptomes (median number of cells per sample, 3711; range, 491-6633; Figure 1B; supplemental Table 1). For each sample, we evaluated the potential for processing and batch artifacts between samples and cohorts, and we selected cohorts with similar "cold-shock signature"[18] for comparison (supplemental Figure 1A). In total, we identified 16 clusters across 3 distinct lineages: T cells, natural killer cells, and myeloid cells (Figure 1B; top, UMAP [uniform manifold approximation and projection]). The distribution of immune cell types from MBL and CLL samples and across patients appeared to be balanced across the cell clusters (Figure 1B; bottom, UMAP; supplemental Figure 1B). Analysis of the proportions of immune cell types, including various T-cell subsets, between MBL and CLL samples revealed no differences, even across time points (T1 vs T2; Figure 1C-D; supplemental Table 2A).

To confirm the absence of major differences in immune cell proportions between MBL and CLL, we performed scRNA-seq on PBMCs collected from a separate cohort of 4 patients with high-count MBL that progressed to CLL (MBL-CLL1-4); the median time from MBL (T1) to CLL diagnosis was 2.68 years (range, 0.7-4.6) and from CLL diagnosis to T2 was 0.6 years (range, 0-1.8). We also evaluated 2 age-matched healthy donors (HDs, median

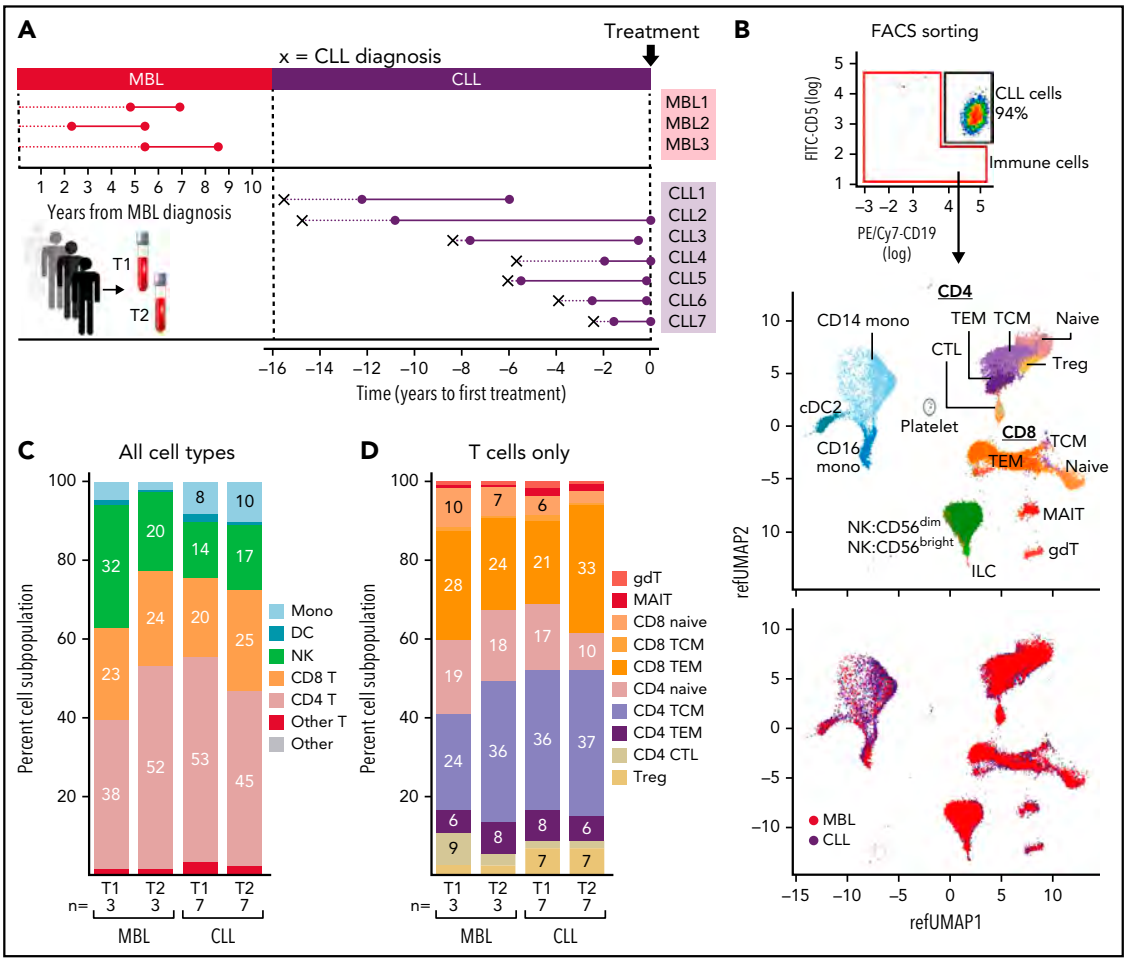


Figure 1. scRNA-seq analysis of immune cells from nonprogressive MBL patients and CLL patients. (A) PBMCs from 2 serial samples were collected for 3 patients with MBL and 7 with CLL. (B) Non-CD19^{+}CD5^{+} cells were isolated by fluorescence-activated cell sorting. UMAP visualization of all immune cells colored by immune cell type (top) and CLL or MBL assignment (bottom). (C) Proportion of immune cell types per time point in patients with MBL or CLL. (D) Proportion of T-cell types per time point in patients with MBL or CLL. Cell percentages were calculated after the number of cell from all samples. CTL, cytotoxic T lymphocyte; DC, dendritic cell; gdT, γ-δ T (cells); ILC, innate lymphoid cell; MAIT, mucosa-associated invariant T (cells); Mono, monocyte; NK, natural killer (cell); pDC, plasmacytoid dendritic cell; T, T cell; TCM, central memory T (cell); TEM, effector memory T (cell); Treg, regulatory T (cells).

number of cells per sample, 4400; range, 2630-7596 cells) using the same approach described above (Figure 2A-B). Again, we observed an absence of major compositional or phenotypic changes in immune cell populations in the transition from MBL to CLL, whereas marked differences in the composition in immune cell types were evident in patients with CLL compared with HDs. In particular, the proportion of CD8^{+} T cells was higher in patients with CLL than in HDs (33% vs 8%, $P = .037$), with a corresponding decrease in CD4^{+} T cells (Figure 2C, left; supplemental Table 2B). The CD4^{+} and CD8^{+} T-cell subtypes that contributed to these differences were naive, central memory CD4^{+} and terminal effector memory CD8^{+} cells (Figure 2C; right). A higher number of differentially expressed genes (adjusted $P < .05$ and $|\text{avg\_log}_2\ \text{FC}| > 0.6$) was observed between HDs and patients with MBL/CLL than between MBL and CLL at the time of progression (patients MBL-CLL-1 and -2; Figure 2D; supplemental Table 3). More differences in gene expression were seen in those paired CLL samples obtained at a time more distant from transition to CLL (patients MBL-CLL-3 and -4), suggesting further evolution of the immune response over time with CLL progression. Effector memory CD8^{+} T cells and CD56^{dim} natural killer cells consistently showed more differentially expressed genes in patients with MBL and CLL than in HDs (Figure 2D, right), which we also observed in a pseudobulk reanalysis of the same data (supplemental Figure 2). Comparable shifts in immune cell expression profiles were observed in the evaluation of independent MBL (MBL1-3, T1) vs CLL (CLL1-7, T2), but only minimal differences were observed in nonprogressing MBL (Figure 2E). We acknowledge that the low number of replicates (n = 2) did not provide sufficient power to detect the biological variability among HDs and that individual-specific variations may have confounded the observed

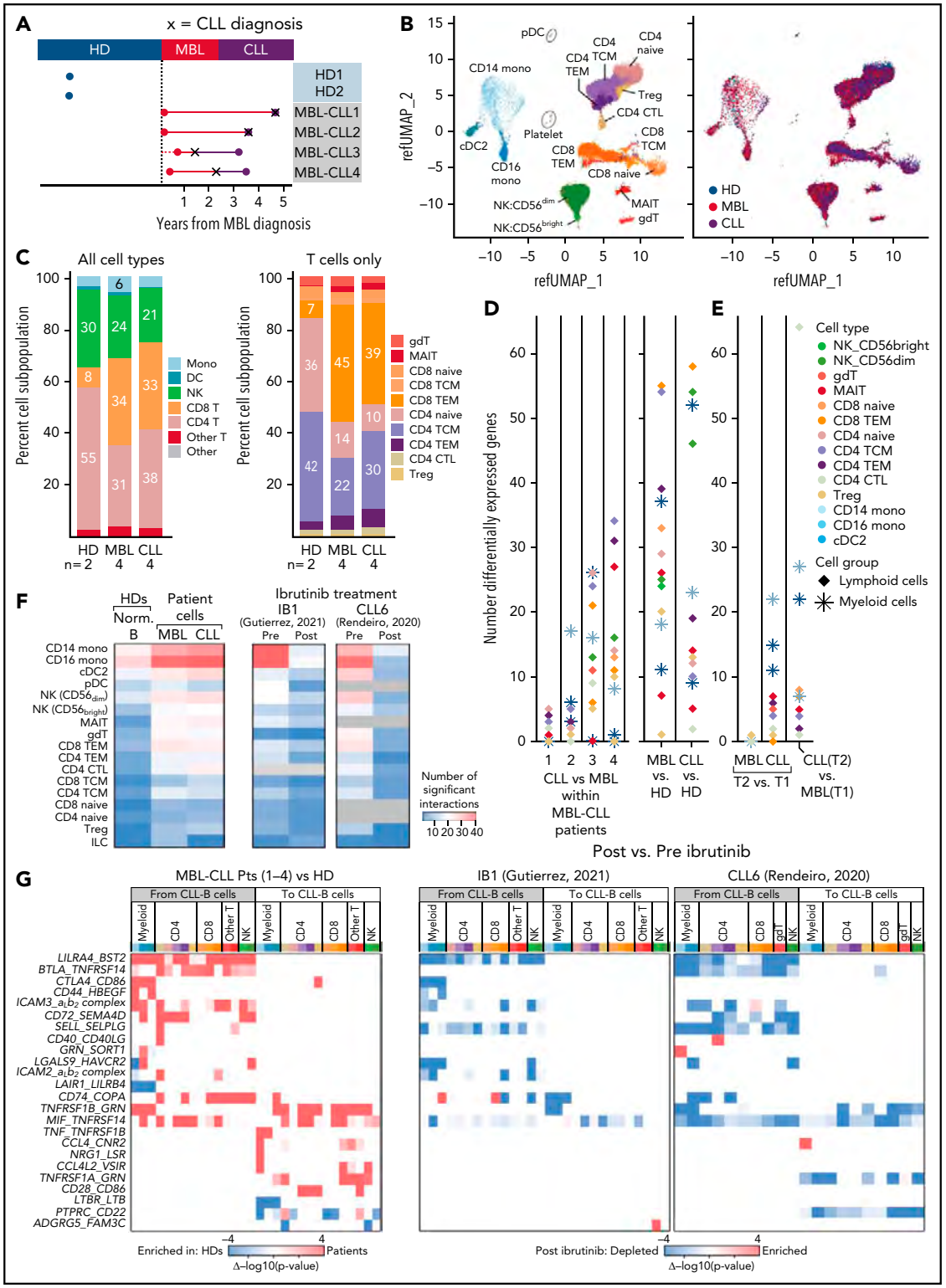


Figure 2.

Figure 2. scRNA-seq analysis of immune cells from healthy donors and disease progression from MBL to CLL. (A) scRNA-seq was performed on PBMCs collected from 4 patients with MBL (red dots) that progressed to CLL (purple dots), and from 2 HDs (blue dots). X, the time of diagnosis of CLL. (B) UMAP visualization of all immune cells colored by immune cell types (left) and by sample types (right). (C) Proportion of immune cell types (left) and T-cell subtypes (right). (D) Number of significant differentially expressed genes for each cell type by performing a comparison of paired samples within patients (left) or comparison between MBL samples or CLL samples vs healthy donors (right). Cells were categorized based on lymphoid and myeloid cells. (E) Same analysis for significant differentially expressed genes was performed on 3 independent patients with nonprogressive MBL and 7 with CLL (Figure 1). (F) Heat maps with the number of the significant ligand-receptor interactions for each cell type under different conditions using CellPhoneDB v2.1.7. Heat map comparing the number of significant interactions between healthy donors and patient samples from either MBL stage or CLL stage (left). Heat maps including samples before and after ibrutinib for 2 additional patients (right).[20,21] Gray boxes indicate an insufficient number of cells to perform interactome analysis. (G) Heat maps representing the difference of *P* values for each ligand-receptor pair regarding specific cell types (x-axis). Interactions that are enriched in patients (red) or enriched in healthy donors (blue) were calculated by subtracting $-\log_{10}$ (*P* value) in healthy donors from $-\log_{10}$ (*P* value) in patients (left). The same interactions that are either enriched (red) or depleted (blue) after ibrutinib (right)[20,21] are calculated by subtracting $-\log_{10}$ (*P* value) in preibrutinib from $-\log_{10}$ (*P* value) in postibrutinib. Pts, patients; cell type abbreviations are the same as in Figure 1.

differences between HD and MBL/CLL samples, but we minimized that risk by selecting age-matched HDs and applied uniform processing to all samples.

To investigate which dysfunctional immune mechanisms may impact CLL biology, we interrogated major molecular interactions between immune and normal B or CLL-B cells in HDs or patients, respectively, using CellPhoneDB v2.1.7, which predicts potential interactions between ligand-receptor pairs based on elevated expression in the corresponding cell types.[19] In so doing, we observed an increased total number of potential interactions in subjects with MBL compared with those in HDs. This increase remained stable with progression to CLL and was evident across diverse immune cell types but was most distinctly observed in monocytes (Figure 2F, left heat map). To examine the effects of B-cell receptor signaling inhibition with ibrutinib on the cellular interactions between immune and leukemia cells, we reanalyzed 4 additional scRNA-seq samples previously generated from PBMCs before and during ibrutinib treatment (cells collected 30-240 days after treatment) from 2 patients with CLL.[20,21] We again observed that the number of cellular interactions in pretreatment CLL samples was higher across immune cell types, especially in monocytes in both patients. Consistently, the number of interactions decreased after ibrutinib treatment to levels similarly observed in HDs (Figure 2F, right heat maps). Most of the interactions upregulated in patients with MBL/CLL involved inhibitory signals of immune cell function proceeding from CLL cells across to various immune cell types, such as *BTLA/MIF-TNFRSF14* (HVEM, observed in MBL-CLL1, -3, and -4), *CTLA4-CD86* (observed in MBL-CLL-4), and *LGALS9-HAVCR2* (TIM3, observed in MBL-CLL1-4; Figure 2G, left; supplemental Figure 3). Notably, only a proportion of cancer cells express these inhibitory signals: *BTLA* (17.4%), *MIF* (41.6%), *LGALS9* (18.2%), and *CTLA4* (10.4%) (supplemental Figure 4). We observed that all these interactions were downregulated after ibrutinib treatment (Figure 2G, right).

Altogether, we observed that the composition and state of immune cells was markedly different between HDs and patients with MBL, whereas no major additional transcriptional changes manifested during natural progression from MBL to CLL. These observations suggest that the key drivers of transcriptional immune dysfunction in CLL may be present early during the course of the disease and are in keeping with the early transcriptomic, genomic, and epigenetic changes already present in MBL, as well as the known increased risk of infections, even at the earliest stages of the disease.[22] Among the features that distinguished immune and leukemia cells interactions in patients with CLL were an increased number of cellular interactions compared with HDs, especially within myeloid cells, that predominantly involved multiple inhibitory immune signals and that were no longer detected after ibrutinib treatment. Thus, although T-cell deficits in CLL have been well investigated,[23,24] the contribution of myeloid cells to inhibitory signals has been far less well characterized and warrants further assessment.

Acknowledgments

The authors thank Jerome Ritz and the DFCI Pasquarello Tissue Bank in Hematologic Malignancies for prospective collection and processing of blood samples from healthy donors.

This work was supported, in part, by National Institutes of Health (NIH), National Cancer Institute (NCI) grants 5P01CA081534-14, P01CA206978, R01CA216273, and UG1CA233338. C.J.W. acknowledges support from the CLL Global Research Foundation. S.H.G. was supported by a Kay Kendall Leukaemia Fund Fellowship. E.M.P. acknowledges research funding from the Doris Duke Charitable Foundation (Physician-Scientist Fellowship), a Conquer Cancer (The ASCO Foundation) Young Investigator Award, and Dana-Farber Flames FLAIR. J.F. received support from the NIH, National Institute of General Medical Science under award R35GM142889. P.G. was supported by Associazione Italiana per la Ricerca sul Cancro (AIRC), Milan, Italy (Special Program on Metastatic Disease; 5 per mille #21198), and ERA NET TRANSCAN-2 Joint Transnational Call for Proposals: JTC 2016 (project #179 NOVEL), project code (MIS) 5041673. S.L. is supported by an NCI Research Specialist Award (R50CA251956).

Authorship

Contribution: N.P., J.F., S.H.G., and C.J.W. designed and conceived the study; N.P., L.Z.R., T.J.K., S.L.S., N.E.K., C.L., T.D.S., P.G., and L.S. collected samples and clinical annotations; S.L. generated the scRNA-seq libraries and processed the raw sequencing data; N.P., Y.E.T., C.K.L., N.C., E.M.P., W.Z., K.J.L., P.V.K., D.S.N., L.R.O., J.F., and S.H.G. analyzed and interpreted data; J.F., S.H.G., and C.J.W. supervised the project; and N.P., Y.E.T., S.H.G., and C.J.W. wrote the paper with assistance from all other authors.

Conflict-of-interest disclosure: N.P. is currently an employee of AstraZeneca; C.J.W. holds equity in BioNTech, Inc. and receives research funding from Pharmacyclics. S.H.G. has received speaker fees from Janssen UK and travel and honoraria from AbbVie and provides research consultancy for Novalgen Limited. P.G. has received honoraria from AbbVie, AstraZeneca, ArQule MSD, BeiGene, Celgene/Juno/BMS, Janssen, Loxo/Lilly, and Roche and research funding from AbbVie, AstraZeneca, Janssen, and Sunesis. P.V.K serves on the scientific advisory boards of Celsius Therapeutics Inc. and Biomage Inc. N.E.K. serves on the advisory boards of AbbVie, Astra Zeneca, Behring, Cytomx Therapy, Dava Oncology, Janssen, Juno Therapeutics, Oncotracker, Pharmacyclics and Targeted Oncology and on the data safety monitoring committees of Agios Pharm, AstraZeneca, BMS-Celgene, Cytomx Therapeutics, Janssen, Morpho-sys, and Rigel and research funding from AbbVie, Acerta Pharma, Bristol Meyer Squib, Celgene, Genentech, MEI Pharma, Pharmacyclics, Sunesis, TG Therapeutics, and Tolero Pharmaceuticals. L.S. has received honoraria from AbbVie, AstraZeneca, Janssen and travel funding from Janssen. T.D.S. received institutional

research support from Genentech and Phamacyclics. The remaining authors declare no competing financial interests.

ORCID profiles: C.K.L., 0000-0002-5772-8743; N.C., 0000-0003-1340-6272; E.M.P., 0000-0002-3382-9769; N.E.K., 0000-0002-5951-5055; P.G., 0000-0003-3750-7342; K.J.L., 0000-0001-9105-5856; L.R.O., 0000-0002-6725-7850; J.F., 0000-0002-0212-5451; C.J.W., 0000-0002-3348-5054.

Correspondence: Catherine J. Wu, Dana-Farber Cancer Institute, Dana 520C, 450 Brookline Ave, Boston, MA 02215; e-mail: cwu@partners.org.

Footnotes

Submitted 5 September 2021; accepted 22 December 2021; prepublished online on *Blood* First Edition 12 January 2022.

*N.P., Y.E.T., and C.K.L. contributed equally to this study.

All single-cell data reported in this article have been deposited in the dbGAP repository (accession number Pha002705.vi). Questions regarding methods and protocols will be answered in response to e-mail request to the corresponding author.

The online version of this article contains a data supplement.

There is a *Blood* Commentary on this article in this issue.

REFERENCES

1. Burger JA. The CLL cell microenvironment. *Adv Exp Med Biol*. 2013; 792:25-45.
2. Dagklis A, Fazi C, Scarfo L, Apollonio B, Ghia P. Monoclonal B lymphocytosis in the general population. *Leuk Lymphoma*. 2009;50(3): 490-492.
3. Rawstron AC, Bennett FL, O'Connor SJM, et al. Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. *N Engl J Med*. 2008; 359(6):575-583.
4. Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015;526(7574):519-524.
5. Agathangelidis A, Ljungström V, Scarfò L, et al. Highly similar genomic landscapes in monoclonal B-cell lymphocytosis and ultra-stable chronic lymphocytic leukemia with low frequency of driver mutations. *Haematologica*. 2018;103(5):865-873.
6. Kretzmer H, Biran A, Purroy N, et al. Preneoplastic alterations define CLL DNA methylome and persist through disease progression and therapy. *Blood Cancer Discov*. 2021;2(1):54-69.
7. Purroy N, Wu CJ. Coevolution of leukemia and host immune cells in chronic lymphocytic leukemia. *Cold Spring Harb Perspect Med*. 2017; 7(4):a026740.
8. Plass M., Solana J., Wolf F.A., et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*. 2018; 360(6391):eaaq1723.
9. Villani AC, Satija R, Reynolds G, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*. 2017;356(6335):eaah4573.
10. Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. *Nature*. 2011;472(7341):90-94.
11. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*. 2014; 344(6190):1396-1401.
12. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016;352(6282):189-196.
13. Gohil SH, Iorgulescu JB, Braun DA, Keskin DB, Livak KJ. Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. *Nat Rev Clin Oncol*. 2021;18(4):244-256.
14. Roerink SF, Sasaki N, Lee-Six H, et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature*. 2018;556(7702): 457-462.
15. Gruber M, Bozic I, Leshchiner I, et al. Growth dynamics in naturally progressing chronic lymphocytic leukaemia. *Nature*. 2019;570(7762): 474-479.
16. Zheng GX, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8(1):14049.
17. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29.
18. Massoni-Badosa R, Iacono G, Moutinho C, et al. Sampling time-dependent artifacts in single-cell genomics studies. *Genome Biol*. 2020;21(1):112.
19. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc*. 2020; 15(4):1484-1506.
20. Rendeiro AF, Krausgruber T, Fortelny N, et al. Chromatin mapping and single-cell immune profiling define the temporal dynamics of ibrutinib response in CLL. *Nat Commun*. 2020;11(1):577.
21. Gutierrez C, Al'Khafaji AM, Brenner E, et al. Multifunctional barcoding with ClonMapper enables high-resolution study of clonal dynamics during tumor evolution and treatment. *Nat Can*. 2021;2(7):758-772.
22. Moreira J, Rabe KG, Cerhan JR, et al. Infectious complications among individuals with clinical monoclonal B-cell lymphocytosis (MBL): a cohort study of newly diagnosed cases compared to controls. *Leukemia*. 2013; 27(1):136-141.
23. Ramsay AG, Johnson AJ, Lee AM, et al. Chronic lymphocytic leukemia T cells show impaired immunological synapse formation that can be reversed with an immunomodulating drug. *J Clin Invest*. 2008;118(7): 2427-2437.
24. Long M, Beckwith K, Do P, et al. Ibrutinib treatment improves T cell number and function in CLL patients. *J Clin Invest*. 2017;127(8):3052-3064.

DOI 10.1182/blood.2021013926

PAPER II

ZNF683 marks a CD8+ T cell population associated with anti-tumor immunity following anti-PD-1 therapy for Richter syndrome

Parry, E. M.*, **Lemvigh, C. K.***, Deng, S., Dangle, N., Ruthen, N., Knisbacher, B. A., Broséus, J., Hergalant, S., Guièze, R., Li, S., Zhang, W., Long, J., Yin, S., Werner, L., Anandappa, A., Purroy, N. Z., Gohil, S. H., Oliveira, G., Bachireddy, P., Shukla, S. A., Huang, T., Livak, K. J., Gad Getz, Neuberg, D., Feugier, P., Kharchenko, P., Wierda, W., Olsen, L. R., Jain, N., Wu, C. J.

Submitted, Cancer Cell, 2022

* = equal contribution

ZNF683 marks a CD8+ T cell population associated with anti-tumor immunity following anti-PD-1 therapy for Richter syndrome

Erin M Parry*[1,2], Camilla K Lemvigh*[1,3], Stephanie Deng[1], Nathan Dangle[1], Neil Ruthen[1], Binyamin A Knisbacher[4], Julien Broséus[5,6], Sébastien Hergalant[5], Romain Guièze[7,8], Shuqiang Li[1,4,9], Wandi Zhang[1], Jaclyn Long[9,10,11], Shanye Yin[1], Lillian Werner[12], Annabelle Anandappa[1,] Noelia Purroy[1], Satyen Gohil[1], Giacomo Oliveira[1,2], Pavan Bachireddy[1], Sachet A Shukla[12], Teddy Huang[1,9], Kenneth J Livak[1,9], Gad Getz[4], Donna Neuberg[13], Pierre Feugier[5,6], Peter Kharchenko[14], William Wierda[15], Lars Rønn Olsen[3], Nitin Jain[15], Catherine J Wu*[1,2,4]

[1] Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, United States
[2] Harvard Medical School, Boston, MA 02215, United States
[3] Department of Health Technology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
[4] Broad Institute of MIT and Harvard, Cambridge, MA 02142, United States
[5] Inserm UMRS1256 Nutrition-Génétique et Exposition aux Risques Environnementaux (N-GERE), Université de Lorraine, 54000 Nancy, France
[6] Université de Lorraine, CHRU-Nancy, service d'hématologie biologique, pôle laboratoires, 54000 Nancy, France
[7] CHU Clermont-Ferrand, 63000 Clermont-Ferrand, France
[8] EA 7453 (CHELTER), Université Clermont Auvergne, 63001 Clermont-Ferrand, France
[9] Translational Immunogenomics Lab, Dana-Farber Cancer Institute, Boston, MA 02215, United States
Department of Immunology, Blavatnik Institute, Harvard Medical School, Boston, MA 02215, United States
[10] Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, MA 02215, United States
[11] Division of Gastroenterology, Hepatology, and Nutrition, Boston Children's Hospital, Boston, MA 02115, United States
[12] Department of Hematopoietic Biology and Malignancy, University of Texas MD Anderson Cancer Center, Houston, TX 77030, United States
[13] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, United States
[14] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, United States
[15] Department of Leukemia, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, United States

*Correspondence: cwu@partners.org

Summary

Unlike many other B cell malignancies, Richter syndrome (RS), an aggressive B cell lymphoma originating from indolent chronic lymphocytic leukemia, is responsive to PD-1 blockade. To discover the determinants of response, we analyzed single-cell transcriptome data generated from 17 bone marrow samples longitudinally collected from 6 RS patients. Response was associated with CD8 effector/effector memory T cells marked by high expression of the transcription factor *ZNF683*, determined to be an intermediate exhausted population evolving from stem-like memory cells and divergent from terminally exhausted cells. Analysis of pre-treatment peripheral blood from 7 independent RS PD-1 responders validated this association of *ZNF683*^{high} T cells with response. This signature overlapped with that of tumor-infiltrating populations from solid tumors and peripheral blood CD8 T cells from melanoma checkpoint blockade responders. ZNF683 was found to directly target key T cell genes (*TCF7, LMO2, CD69*) and pathways regulating T cell cytotoxicity and activation.

Keywords: Richter's transformation, chronic lymphocytic leukemia, immunotherapy, PD-1, checkpoint blockade, T cells, ZNF683, Hobit, Tox, single-cell RNA sequencing

Introduction

Although immune checkpoint blockade (CPB) such as anti-PD1 therapy has transformed the clinical practice of oncology, responses to anti-PD1 antibody across hematologic malignancies have varied considerably, with exceptional response rates in Hodgkin lymphoma but generally disappointing activity for B cell cancers such as myeloma and chronic lymphocytic leukemia (CLL). In this context, the responses of Richter syndrome (RS) to PD-1 blockade, recently reported across several clinical trials at 43-65%, have been unexpected (Ding et al., 2017; Jain et al., 2018; Younes et al., 2019). Described as a transformation of indolent CLL into an aggressive lymphoma, RS has historically resulted in dismal overall survival and poor responsiveness to chemotherapy (Parikh et al., 2014). RS thus provides a unique setting to gain understanding of both tumor and immune determinants of response to PD-1 CPB in hematologic malignancies. Great strides in our understanding of the mechanisms underlying treatment responses and the tumor microenvironment have been provided by the rapid adoption of droplet-based single-cell RNA-sequencing (scRNA-seq) to deeply characterize heterogeneous tumor-infiltrating immune cell populations (Zhang and Zhang, 2020). Here, through scRNA-seq based examination of serial marrow samples from 6 RS and 2 CLL patients treated with nivolumab (anti-PD-1 antibody) combined with ibrutinib (Jain et al., 2018), we found the transcription factor *ZNF683*, encoding the Blimp1 homolog known as Hobit (Mackay et al., 2016), to mark a population enriched in RS responders, an association confirmed in peripheral blood from independent anti-PD1 treated RS patients. Characterization of *ZNF683* overexpression revealed *ZNF683* to directly regulate the key pathways of T cell cytotoxicity and activation, suggesting an important role in governing anti-tumor response.

RESULTS

RS as an opportune setting in which to study the response determinants to PD-1 therapy

We examined serial BM samples collected from patients with RS or relapsed/refractory (R/R) CLL on an investigator-initiated phase II trial, treated with PD-1 blockade (nivolumab; 3 mg/kg every 2 weeks in 28-day cycles) combined with ibrutinib (420 mg daily, continuous, starting day 1 of cycle 2) (NCT02420912) (Jain et al., 2018). Of 24 enrolled patients with RS, 10 demonstrated response (ORR 42%), with median time to response of 28 days and all responses achieved by 3 months. In contrast, none of 10 enrolled R/R CLL subjects demonstrated benefit to nivolumab beyond single-agent ibrutinib. We focused our discovery efforts on a cohort of 6 patients with RS (4 responders, 2 non-responders) and two patients with R/R CLL, treated on a separate study arm with the same protocol, for which matched on-therapy marrow samples were available at the time of the 3-month response assessment (**Figure 1A**, **Table S1**). All RS responders achieved best response by 3 months after initiation of therapy, with complete response (CR) observed in all, aside from RS-R3, who had a delayed partial response at that time. Only RS-R2 had pathologically confirmed RS marrow involvement at time of study initiation while the remaining cohort participants had detectable nodal RS with confirmed CLL marrow involvement. All 4 RS responders were previously treated with CLL-directed chemotherapy (median 2.5 lines, range 1-4), while RS-NR1 had novel agent exposure and RS-NR2 received RS-directed chemo-immunotherapies (**Figure 1A, Table S1**).

Per patient, viable malignant (CD5+CD19+) and non-malignant (CD5-CD19-) cell fractions were isolated from marrow specimens by fluorescence activated cell sorting (FACS). For the RS-R2^{baseline} and RS-R1^{progression} marrow samples, which had pathologically confirmed RS, the malignant fraction was further separated based on size (FSC/SSC) to identify RS (large) and CLL (small) cells (**Figure 1B, Figure S1A-B).** To examine these heterogeneous sorted cell

populations at high resolution, single-cell RNA-sequencing (scRNA-seq) was performed. After initial filtering (**Methods)** joint clustering with *Conos* (Barkas et al., 2019) was performed on 78,488 non-tumor and 117,703 malignant cells from a total of 19 marrow samples (17 serial trial samples and 2 age-matched healthy marrow donors) (**Figure 1C**).

Single-cell evaluation of both RS and CLL cells revealed that RS cells displayed higher mean reads/cell (6,094 UMI/cell, 1,964 genes/cell) compared to CLL (2,963 UMI/cell, 1,070 genes/cell) (p value < $2.2x10^{-16}$, Wilcoxon test), in line with other aggressive hematologic malignancies (Zheng et al., 2017) (**Figure S2A-H, Table S2**). Four of 11 malignant B cell clusters appeared to be patient-specific (e.g. clusters C, G, H and K), highlighting their unique gene expression patterns likely attributable to their individual tumor genetics (Penter et al., 2021a; Tirosh et al., 2016) (**Figure S3A-C**), and cluster E was predominant in the RS samples (**Table S3, Figure S3C**). By unsupervised joint clustering with *Conos* (Barkas et al., 2019), RS clustered separately from CLL (**Figure S3D**) and displayed relative upregulation of pathways of cell cycle (E2F targets, G2M checkpoint, mitotic chromatid segregation), oxidative phosphorylation and MYC targets as compared to CLL (**Figure 1D, Table S3**). In comparing the aggregate CLL cells of RS-Rs and RS-NRs, no significant pathways were found that distinguished responding from non-responding patients. Across samples, HLA class I and II expression were maintained (**Figure S3E**). Of note, we observed the entire set of malignant B cells to express a wide variety of inhibitory and immune modulatory molecules at baseline, including *CTLA4, LAG3, TNFRSF8, TGFBI/TGFB1*, suggesting complex microenvironment interactions prior to CPB initiation (**Figure S3F**). Previously, no difference in response based upon PD-1 and PD-L1 staining of tumor was demonstrated in 10 study patients (Jain et al., 2018).

Distinct infiltrating lymphocyte populations in RS marrow

Since bone marrow is an important reservoir of lymphoid cell populations, (Mazo et al., 2005; Mercier et al., 2011) we focused on the characterization of the diverse captured marrow-resident T and NK cells (60,727 (77% of non-tumor cells)) (**Figure S4A-D, Figure 2A**). Following sub-clustering (**Methods**, **Figure 2A**), we identified 11 transcriptionally distinct clusters, consisting of 2 NK cell clusters, 4 CD8 T cell clusters and 5 CD4-predominant clusters (**Figure 2A-B**). Each transcriptional cluster was defined based on a combination of cluster marker genes (**Table S4**) and lineage markers (**Figure S5A**; defining genes highlighted in **Figure 2B**).

Four clusters (clusters 1, 3, 4 and 8) were composed of $CD8^+$ T cells. Cluster 1 was a large effector/effector memory (E/EM) cell population marked by the expression of the transcription factor *ZNF683*, and displayed co-expression of cytolytic machinery genes (*GZMA, GZMB, PRF1, NKG7*) and the activation marker *CD226* and intermediate expression of exhaustion markers (*LAG3, ENTPD1, CD160*). Cluster 3 was marked by expression of *GZMK* and *CD27* and genes suggestive of marrow-resident memory (*CD69, CXCR4*). A subset of cluster 3 contained $TCF7^+$cells, as has been observed in stem-like memory populations (Miller et al., 2019; Sade-Feldman et al., 2018) (**Figure 2B**, **S5A**). Cluster 4 displayed a higher relative mitochondrial content and intermediate features between clusters 1 and 3, such as lower *ZNF683* and higher *IL7R* and *GZMK* expression compared to cluster 1. Cluster 8 CD8 T cells had high expression of multiple exhaustion markers (*TIGIT, PDCD1, LAG3, HAVCR2*) and high *TOX* expression. Other non-CD8 cytolytic clusters included cluster 9, which comprised CD4 T cells with cytotoxic gene expression, and 2 populations of NK cells (clusters 6 and 10). Of the CD4 T cells, cluster 7 showed features consistent with a T regulatory phenotype while cluster 5 was comprised of a naïve-like T cell population. In repeat scRNA-seq characterizations of pre- and post-therapy marrow samples from RS-R2 and RS-NR1, we

confirmed the cluster identities based on linked surface protein marker expression (CITE-seq, **Methods**) (**Figure 2C**, **Figure S5B)**.

We observed the CD8 T cell clusters found in RS marrow displayed high cytotoxicity and exhaustion (**Figure 2D)**. Exhaustion levels were highest in clusters 3 and 8, while cluster 1 reflected an intermediate exhaustion expression score with maintained cytotoxicity. Compared to marrow-infiltrating T cells from the RS samples to that across 30 healthy adult volunteers (Oetjen et al., 2018) (**Table S4**), RS and CLL marrow were enriched in cytotoxic populations (clusters 1, 4, 6, 8, and 9, all p values < 0.05, 2-sample t test) and T regulatory cells (cluster 7, p = 0.001, 2-sample t test). In contrast, normal marrow was enriched in cells with a naïve-like signature (cluster 5, p = 0.06, 2-sample t test) (**Figure 2E-F, Figure S5C-E**).

Increased *ZNF683*+ effector/effector memory CD8+ T cells in anti-PD1 responders

Several observations led us to implicate *ZNF683*-expressing CD8 T cells in the RS response to PD-1 blockade therapy. First, across the samples, the marrow-infiltrating T cells enriched in RS responders predominantly corresponded to the *ZNF683*^{high} CD8 cells of cluster 1 (**Figure 3A, Figure S6A-B**). This quantitative difference in the cluster 1 *ZNF683*-expressing E/EM cells between clinical outcome groups was already evident at baseline, where the pre-treatment samples from RS responders demonstrated a larger proportion of cells within cluster 1 (p = 0.04, t-test), with correspondingly fewer T regulatory cells (cluster 7, p = 0.006, t-test) than RS non-responders (**Figure 3B-**bar graphs).

Second, within assigned clusters, comparison of gene expression between RS-Rs and RS-NRs again identified higher *ZNF683* expression in RS-Rs, including in clusters 1 and 8 (**Figure 3B-C, Figure S6C**). Across all CD8 T cell clusters, *ZNF683* was one of the top upregulated

transcription factors in RS-Rs, while *TOX* was highly upregulated in RS-NRs (**Figure 3D, Figure S6G, Table S5**). Third, the kinetics of changes in *ZNF683*-expression was highly correlated with response. RS-R3, who had a delayed response to therapy, displayed an expansion of the CD8 clusters 1 and 4 at the time of response (**Figure 3E**, **Figure S6C).** With RS progression, relative *ZNF683* expression/cell decreased (**Figure 3F, Figure S6C**).

To identify co-regulated genes, we performed differential gene expression analysis between RS-R and RS-NR cells on a per-cluster basis (**Figure 3B-heatmaps, Figure S6D-F)**. In the *ZNF683*^{high} cluster 1, this notably included several T cell transcription factors (*RORA, SATB1, TCF25* and *BATF* in Rs; *TOX* and *ARID5A* in NRs) and immune signaling molecules (*CD226* in Rs; *KLRC1, KLRB1, CD27, CD69* and *GZMK* in NRs) (**Figure 6D**, **Table S5**). In the cluster 8 (exhausted), *ZNF683* was one of the top genes distinguishing the terminal exhausted cells of RS-R and RS-NR, along with *CD226*, while higher expression of *CD27, KLRC1*, and *SIRPG* was found in RS-NR *(***Figure 3C)**. We also identified gene expression changes within non-*ZNF683* expressing clusters that distinguished RS-Rs. For example, in the CD4 clusters 2 and 9, expression of cytolytic machinery (*NKG7, CST7, PRF1, GZMH, GZMM, GZMK, GZMA, FGFBP2, GNLY*) and cytokines (*CCL4, CCL5*) was increased in responders relative to non-responders (**Figure S6E-F, Table S5**), supporting cytotoxic CD4 T cells as another hallmark of response to PD-1 CPB in RS. Expression of *TCF7* did not vary with response status (**Figure S6H**).

Evaluation of the T cell receptor (TCR) repertoire of individual marrow-infiltrating T cells from a representative RS-R and RS-NR (RS-R2 and RS-NR1) revealed prominent differences in stability and phenotype among the expanded clonotypes before and following PD-1 blockade (**Figure S6I, Table S6**). RS-R2 demonstrated a single dominant clonotype (count 723 of 1,683) and multiple other highly expanded clones (counts range 16-114) prior to PD-1 therapy (**Figure**

3G-pie charts), and subsequent stable persistence of these expanded clones following PD1 therapy (**Figure 3G-left**). These expanded clonotypes predominantly resided within the *ZNF683*^{high} clusters 1 and 4 and even included a cytotoxic CD4 clonotype. In contrast, RS-NR1 had fewer hyperexpanded and large clones pre-treatment and exhibited more clonotype expansion with PD-1 therapy (**Figure 3G-right**). The top clonotypes of RS-NR1 showed higher exhaustion compared to RS-R2 (cluster 8) ($p < 0.001$, Poisson test). Neither patient was observed to have expanded clonotypes in the T regulatory compartment (**Table S6**). Bulk TCR analysis of paired baseline and 3 month post peripheral blood from 4 RS-R [R2, R4, R5, R6] and 4 RS-NR [NR1, NR2, NR4, NR6] (**Table S6**) demonstrated that the clonal repertoire in responding patients was relatively stable while clonal expansion and contraction were observed in the RS-NR (**Figure 3H**) (**Methods** (Penter et al., 2021b)). At the time of progression, we did not observe further clonal shifts in 2 RS-R patients who subsequently lost response (**Figure 3I**).

Overall, the distribution of clonotypes across diverse CD8 phenotypes suggested a continuum of malignancy-associated T cell states. Trajectory analysis (**Methods**) supported a T cell differentiation path from GZMK+ memory to *ZNF683*^{intermediate} and then a branching towards either *ZNF683*^{High} or terminal exhaustion (**Figure 3J**).

ZNF683 marks populations of tumor-infiltrating lymphocytes in RS and across cancers

To establish the extent to which *ZNF683*^{high} T cell signatures are detectable in a broader population of RS patients, we examined bulk RNA-seq data generated from 35 independent RS biopsy specimens of lymph node or spleen origin. While 28 of 35 (80%) expressed minimal normalized *ZNF683* levels in the computationally deconvoluted T cell compartment (**Methods)**, 7 patients (20%) were confirmed to have higher *ZNF683* expression and higher

cluster 1 signatures (**Figure 4A)**. A similar distribution of samples with *ZNF683*^{high} expression, when corrected for T cell fraction, were observed in 18 of 81 (20%) CLL for which non-CD19 selected transcriptomic data was available (Yin et al., 2019) (and thus allowing for the sampling of non-CLL immune populations) (**Figure 4B**). The subset expressing *ZNF683*^{high} T cells displayed a trend towards improved overall survival (p = 0.11, **Figure 4B**). Thus, even in unselected CLL patients, *ZNF683* expression appears to mark a T cell population associated with improved clinical outcome.

We evaluated the extent of overlap by our *ZNF683*^{high}-cluster 1 signature with well-annotated transcriptionally defined populations from a large dataset of single-cell tumor infiltrating lymphocytes (TIL) (Zheng et al., 2021). Indeed, we found high similarity of our signature to the two CD8+ pan-cancer derived *ZNF683*-expressing clusters (**Figure 4C)**, including one which was previously computationally identified as a transitional intermediate between naïve-like and exhausted CD8+ T cells – not dissimilar to our trajectory analysis (**Figure 3J**). Moreover, our exhausted cluster 8 showed overlap with the terminal exhausted and NK-like populations described from the pan cancer analysis, while our cluster 3 demonstrated broad overlap across several previously defined pan-cancer clusters, including with naïve and TCF7+ exhausted populations. Given the presence of such similar *ZNF683*^{high} populations in these disease settings, we examined 13 cancers across TCGA data and observed an association with survival in melanoma cases with high T cell expression of *ZNF683*, as in CLL and high cluster 1 signature **(Figure 4D)**. The observed similarity in *ZNF683*^{high} gene signatures in melanoma support the notion that *ZNF683* may function to mediate T cell gene expression states across certain solid tumor malignancy contexts.

***ZNF683* regulates key genes in T cell differentiation and activation**

To test the hypothesis that *ZNF683* governs key regulators of T cell state and activation, we examined its cellular impact *in vitro* in Jurkat cells, which lack endogenous *ZNF683*, unlike cultured primary T cells (**Figure S7A)**. A Flag-tagged L isoform of *ZNF683*, the dominant transcript associated with response and active isoform (Vieira Braga et al., 2015), was introduced into the Jurkat cells under a doxycycline-inducible promoter (**Figure 5A**; **Figure S7B-C**).

ZNF683 overexpression (+ doxycycline) revealed upregulation of *PRF1*, *ITGA1*, *CD244*, *CD226* and *IL2RB* as well as downregulation of *PRDM1* and *IL2* (**Figure 5B)**. Likewise, analysis of RNA-seq data generated from our stable ZNF683-FLAG cell lines compared to a mock vector in the presence of doxycycline produced similar findings (**Figure S7D, Table S8**). To probe the binding sites of ZNF683 and identify its direct targets, we performed CUT&RUN (Skene and Henikoff, 2017) using anti-FLAG antibodies alongside controls (H3K4Me3, isotype controls) (**Table S8**). Twenty-four peaks were differentially regulated in doxycycline-exposed Jurkat cells compared to doxycycline-absent or isotype controls (FDR <0.1, **Methods**). These peaks localized preferentially to ENCODE-identified regulatory regions adjacent to several genes, rather than promoters, including elements adjacent to key immune genes (*TNFAIP8, BTN3A2, CD69/KLRF1, ARHGAP2, CAMK4, LMO2, IL2, TCF7*; **Table S8**). Included amongst these targets were previously identified and functionally validated genes (Vieira Braga et al., 2015) (e.g. *BTN3A3*), and peaks shared with previously generated PRDM1 (Blimp1) CHIP-seq data from human cells (Davis et al., 2018) (**Figure 5C, Figure S7E, Table S8**).

In support of the notion that ZNF683 regulates a T cell differentiation path, we observed a binding peak for ZNF683 upstream of *TCF7*, which helps maintain a stem-like or progenitor

exhausted cell state (Chen et al., 2019; Miller et al., 2019; Sade-Feldman et al., 2018; Siddiqui et al., 2019). Evaluation of external ATAC-seq data generated from exhausted T cell subsets (Philip et al., 2017) revealed PD-1^{high} TILs have reduced accessible chromatin at this TCF7 peak and other identified putative ZNF683 binding sites. Thus, terminal exhausted cells not only lose *ZNF683* expression but also display chromatin remodeling-that likely abrogates some of its downstream effects (**Figure 5C**-track PD-1^{high} [blue]). Other ZNF683 targets included a regulatory element near *CD69*, a marker of tissue residency that is differentially expressed on exhausted tumor-specific T cell subsets (Beltra et al., 2020), and *KLRF1*, an NK-like marker, that also corresponded to decreased accessible chromatin region in exhausted TIL subsets. By applying CISTROME-GO (Li et al., 2019a), which integrates CUT&RUN and RNA-seq data layers, we observed enrichment in pathways corresponding to antigen binding and presentation, transcription factor activity, and T cell mediated cytotoxicity and cell killing upon *ZNF683* expression (**Figure 5D**). From the CUT&RUN data, we detected that the top binding motif of ZNF683 was predicted to overlap with the motif of *LEF1* (**Figure 5E-top, Methods**). This motif displayed notable similarity to ZNF683-homolog PRDM1 (**Figure 5E-bottom,** $p = 0.002$). Altogether, ZNF683 appears to bind to and regulate key T cell loci involved in T cell function and immune response with similarity in sequence recognition to its homolog PRDM1, and is predicted to function both as a transcriptional repressor and activator.

ZNF683 expression in peripheral blood is associated with response to immune checkpoint blockade

To determine if the association between *ZNF683* expression and response to PD-1 blockade in RS could be validated in easily sampled peripheral blood, we analyzed RNA-seq data generated from T cells isolated from pre-treatment blood samples from 7 independent RS patients treated with PD-1 therapy. Differential gene expression analysis between transcriptomes from 2 RS responders and 5 RS non-responders at baseline identified 1,528 genes ($p < 0.05$, absolute Log_2

fold change >1.5 and base mean expression value of >50, **Methods**). Among the top differentially regulated genes was *ZNF683* (Log_2 fold change = 2.13, p = 0.037, **Figure 5F**) as well as several other genes enriched in cluster 1 (*BATF, CORO1A, CD38, ITGB2, GZMM*). Baseline *PDCD1* and *ENTPD1* were also associated with subsequent response in the peripheral blood, which was not observed in the marrow-derived discovery data. Non-responders appeared to be enriched in NK/T genes, including *KLRC1, FCGR3A, KLRG1, KLRF1* and *FCER1G*, echoing findings from prior studies, where this NK-like skewing was seen in more exhausted CD8 marrow populations. Supporting our detected association of cluster 1 *ZNF683*^{high} and CPB response, our single-cell cluster 1 signature displayed high overlap with a previously identified peripheral blood signature in melanoma PD-1 CPB responders (Fairfax et al., 2020) ($p < 0.001$ hypergeometric test, **Figure 5G**).

DISCUSSION

While the functional exhaustion of anti-tumor antigen-specific CD8+ T cells is well-established in human cancers, it has been increasingly recognized that this T cell population is heterogenous, and its composition may vary per distinct cancer context. T cell transcription factors define and regulate many of these exhaustion subsets (McLane et al., 2019). Mouse studies have provided key insights into the path from stem-like exhaustion (marked by expression of *TCF7*) to terminal exhaustion (delineated by expression of *TOX*, *BLIMP1*, *EOMES* and others), and have highlighted cell populations intermediate between these two axes (McLane et al., 2019). However, our understanding of those T cell subpopulations capable of reprogramming, replication, and generating effective human anti-tumor immunity in the setting of PD-1 blockade remains incomplete.

Here, we uncover a role for the transcription factor *ZNF683*, or Hobit, in marking an intermediate exhausted population in RS bone marrow that is absent from healthy controls and is associated with CPB response. Trajectory, differential gene expression and clonotype analyses demonstrated this population as an alternate path from stem-like exhaustion memory (cluster 3) towards exhausted effector function (cluster 1, *ZNF683*^{high}), and divergent from terminal CD8 T cell exhaustion (cluster 8, *TOX*^{high}). Concordant with this work, recent investigations have similarly associated ZNF683^{high} or tissue-resident memory (TRM) populations with pre-terminal exhaustion in other malignancies (Anadon et al., 2022; Zheng et al., 2021). Our cluster 1 *ZNF683*^{high} population bears similarity to the recently reported CD69- intermediate exhausted tumor-specific populations in mouse models that reside between CD69+ progenitor exhausted and terminal exhausted populations, and maintain replicative potential (Beltra et al., 2020). Altogether, these lines of evidence, along with our discovery marrow-based analysis in RS and extension studies across TCGA-characterized solid tumors, highlight *ZNF683* as functioning as a potential important mediator of intermediate exhaustion states across several malignancies.

Our study is consistent with the notion that transcription factors such as ZNF683 and TOX govern the choice of differentiation path taken by anti-tumor T cells. In support of this model, TOX knockdown mouse models have been shown to exhibit increased *ZNF683* expression in virus-specific T cells (Alfei et al., 2019). Our results reported herein substantially advance our current understanding of how ZNF683 functions to directly and indirectly mediate its effects on T cell state. First, through analysis of a cell line model with enforced expression of *ZNF683*, we demonstrate that ZNF683 has putative transcriptional enhancer and repressor activity and targets T cell genes and pathways involved in differentiation, effector function, activation and cytotoxicity. Second, we show that its binding sites reside predominantly in enhancer elements

rather than promoters and that its binding motif displays similarity to the binding sites of its homolog BLIMP1. Notably, several of these binding sites localize within epigenetically scarred regions of exhausted T cells (Philip et al., 2017), pointing to how the differential response to this transcription factor may occur in terminal exhausted cells. We emphasize that these human-based studies of *ZNF683* are critical to undertake, despite the current challenges of lack of availability of critical tools such as robust antibodies against human ZNF683, since its function in human cells may be distinct from that in murine models. For example, T cell expression of *ZNF683* in mouse is thought to be restricted to the resident memory program (Mackay et al., 2016), unlike *ZNF683* expression in human cells, where it marks the T_{RM} populations but additionally regulates long-lived effector cells in CMV infection (Vieira Braga et al., 2015) that retain immediate effector functions upon stimulation.

With the success of clinical studies of PD-1 CPB across human cancers, a priority has been focused on identifying the underlying T cell populations capable (or not) of responding to immune checkpoint blockade. Our data provides insight into how terminally exhausted T cells are less able to respond to re-stimulation or revert to an earlier more activated state. A potentially valuable translational corollary is that the $ZNF683^{high}$ T cell gene signature detected in the marrow of our RS subjects was not only associated with CPB response but that this same signature could be detected in the pre-treatment peripheral blood of independent RS PD-1 responders and for PD-1 treated melanoma patients as well. Certainly, we have more to learn about why ZNF683 appears to hold prognostic and predictive significance in certain disease contexts (RS, SKCM) but not others. But for now, our data strongly suggest that this ZNF683 CD8+ T cell population is key in anti-tumor immunity and the ability of patients to respond to CPB. A fertile area of future investigation is the testing of ZNF683 expression as a potential

predictive marker in studies of immunomodulatory therapy across the relevant malignancies because it can be detected as a circulating cell population and its expression is not TIL restricted.

ACKNOWLEDGMENTS

We are grateful to Lucas Pomerance, Catherine Gutierrez, and David Braun for constructive and valuable discussion. We acknowledge Samantha Hoffman for helpful input. -We appreciate Elizabeth Witten for expert lab management. E.M.P. acknowledges support from the Doris Duke Charitable Foundation (Physician-Scientist Fellowship), the American Society of Oncology Conquer Cancer Foundation Young Investigator Award and the Dana-Farber FLAMES Lymphoma Fellowship. C.K.L is supported in part by the Fishman Family Fund. C.J.W. acknowledges support from the Lavine Family Foundation. S.L. is supported by the NCI Research Specialist Award (R50CA251956). The authors acknowledge the Centre de Ressources Biologiques (CRB) Lorrain of Nancy BB-0033-00035 for patient sample management and the French Innovative Leukemia Organization (FILO). This study was supported by NIH/NCI P01 CA206978 (to C.J.W. and G.G.) and NCI (1U10CA180861 and 1R01CA155010) (to C.J.W). S.A.S. is supported by CPRIT RR220009.

AUTHOR CONTRIBUTIONS

E.M.P, C.K.L. and C.J.W. designed and performed the experiments, analyzed data, and wrote the manuscript; E.M.P, S.L.D, N.D., N.P., S.L, J.L, S.G, G.O., T.H., W.Z. and A.A. performed experiments; J.B., R.G., P.F. collected RS samples and Se.H and C.K.L. performed RNAseq analysis; N.R, B.A.K., and C.K.L analyzed data with supervision and guidance from L.R.O.,

S.Y, S.A.S., P.K., K.L. and G.G. L.W. performed and D.N. supervised statistical analyses. W.W. and N.J designed trial protocol and created sample banking with P.B. All authors discussed and interpreted results.

DECLARATION OF INTERESTS

C.J.W. receives funding support from: Pharmacyclics; holds equity in: BioNTech, Inc; G.G. is a founder, consultant and holds privately held equity in Scorpion Therapeutics, receives funding support from: IBM and Pharmacyclics, is an inventor on patent applications related to: *MuTect*, *ABSOLUTE, MutSig, MSMuTect, MSMutSig, MSIDetect, POLYSOLVER,* and *TensorQTL*; R.G. receives funding support from: Abbvie, Janssen, Gilead, AstraZeneca, and Roche; N.J. receives research funding from: Pharmacyclics, AbbVie, Genentech, AstraZeneca, BMS, Pfizer, Servier, ADC Therapeutics, Cellectis, Precision BioSciences, Adaptive Biotechnologies, Incyte, Aprea Therapeutics, Fate Therapeutics, Mingsight, Takeda, Medisix, Loxo Oncology, Novalgen and serves on Advisory Board /Honoraria: Pharmacyclics, Janssen, AbbVie, Genentech, AstraZeneca, BMS, Adaptive Biotechnologies, Precision BioSciences, Servier, Beigene, Cellectis, TG Therapeutics, ADC Therapeutics, MEI Pharma; W.G.W. reports funding from GSK/Novartis, Abbvie, Genentech, Pharmacyclics LLC, AstraZeneca/Acerta Pharma, Gilead Sciences, Juno Therapeutics, KITE Pharma, Sunesis, Miragen, Oncternal Therapeutics, Inc., Cyclacel, Loxo Oncology, Inc., Janssen, Xencor. B.A.K, C.J.W and G.G. are inventors on patent: "Compositions, panels, and methods for characterizing chronic lymphocytic leukemia" (PCT/US21/45144); S.A.S. reports nonfinancial support from Bristol-Myers Squibb, and equity in Agenus Inc., Agios Pharmaceuticals, Breakbio Corp., Bristol-Myers Squibb and Lumos Pharma. N.P. is currently an employee of Bristol Myers Squibb. K.J.L. holds equity in Standard BioTools Inc. (formerly Fluidigm Corporation). C.J.W. and E.M.P are inventors on a patent, US Patent Application No.

17/634,465 filed on 02/10/2022 “Immune Signatures Predictive of Response to PD-1 Blockade in Richter’s transformation.”

Main Figure Titles and Legends

Figure 1. Single-cell RNA-sequencing cohort for determinants of response to PD-1 blockade

(A) Response data (pie chart) for RS arm of NCT 02420912 and Swimmer’s plot showing the 8 patients included in the discovery cohort and treatment schema (gray bars). RS responders are represented in purple, RS non-responders are represented in light blue and relapsed/refractory CLL patients in dark blue.

(B) Experimental schema showing flow cytometry sorting strategy of populations used for single-cell RNA-sequencing experiments followed by analytic strategy.

(C) LargeVis embedding displaying joint clustering of tumor and immune cell populations.

(D) GSEA analysis shows differences between tumor transcriptome of RS vs. CLL.

Figure 2. Transcriptionally identified T cell populations in RS and CLL marrow

(A) Joint graph LargeVis embedding of immune cell populations with subsequent sub-clustering of T and NK cells revealing 11 distinct transcriptional clusters.

(B) Heat map showing marker genes and cluster identities for 11 identified T and NK cell clusters. Cluster 1, red; Cluster 4, teal, Cluster 8, magenta.

(C) Heat map showing cell surface markers for 11 identified T and NK cell clusters by CITE-seq.

(D) Identified T and NK cell clusters plotted by Cytotoxicity and Exhaustion scores (Oliveira et al., 2021).

(E) Bar graphs showing T and NK population distributions across serial marrow samples and normal bone marrow samples.

(F) Distribution of T and NK cell populations in RT and CLL marrows as compared to normal bone marrow samples (2 sample t-tests).

Figure 3. Gene expression changes associated with PD-1 response

(A) Joint graph LargeVis sub-clustering embedding showing RS-R (purple) and RS-NR (light blue) cells as well as ZNF683-expressing cells (red).

(B) Cluster proportions for RS-R (purple) and RS-NR (blue) in bar graphs at top with heat maps showing representative cytotoxicity, exhaustion and expression changes within circled clusters for RS-R and RS-NR.

(C) Volcano plot showing gene expression differences comparing RS-R to RS-NR cells in Cluster 8 Exhausted.

(D) Top transcription factors in comparison of all CD8 T cells between RS-R and RS-NR.

(E) Kinetics of cluster 1 and 4 cell proportions showing expansion at time of response in RS-R3.

(F) Bubble plot showing percent expressing and average relative expression for RS patients with progression on PD-1 blockade.

(G) Single-cell TCR analysis showing captured distribution of clonotypes (pie chart) pre- and post CPB therapy (top) and cluster distribution (bottom bar plots) of expanded clonotypes for RS-R2 and RS-NR1 pre (navy) and post (light blue) CPB.

(H) Bulk-TCR sequencing of 4 RS-R and 4 RS-NR show clonotype stability or shifts with PD-1 CPB therapy from pre-therapy to time of response.

(I) RS-R showing changes in clonotypes from response to progression.

(J) Trajectory analysis of CD8 clusters showing inferred patterns of T cell differentiation across clusters.

Figure 4. ZNF683 marks a distinct population in RS and CLL with prognostic significance

(A) ZNF683 TPMs corrected for T cell content in RS bulk RNA-seq samples (N=35) (top). Top 20% of samples, blue; bottom 80%, pink. Cluster 1-ZNF683high signature and T cell normalized ZNF683 expression in bulk-RNA-seq data from 35 independent RS patients (bottom).

(B) ZNF683 TMPs corrected for T cell content in CLL bulk-RNA-seq samples (n=81) (top). Top 20% of samples, blue; bottom 80%, yellow. Kaplan-Meier curve showing ZNF683 expression in PBMCs is associated with overall survival in CLL (bottom).

(C) Heatmap showing single-cell GSEA scores for identified CD8 T cell clusters as compared to Pancancer analysis CD8 T cell clusters (Zheng et al., 2021).

(D) Survival curve showing ZNF683 expression and C1 gene expression signature with overall survival in TCGA melanoma data.

Figure 5. ZNF683 directly regulates key pathways of T cell activation and function

(A) Schema showing doxycycline-inducible expression of ZNF683 in Jurkat cells.

(B) Volcano plot showing genes differentially regulated by ZNF683 induction by RNA-seq.

(C) CUT&RUN on Jurkat cell lines (top) shows binding of ZNF683 at regions surrounding key immune genes that correspond to differential ATAC-seq peaks in T cell subsets24 and prior PRDM1 ChIP-seq data (Davis et al., 2018). N, Naïve; CM, Central memory, PD-1high; PD-1 high tumor infiltrating CD8 T cells.

(D) Top pathways predicted to be differentially regulated by ZNF683 through CISTROME-GO analysis.

(E) Top predicted motifs for ZNF683 (top) compared to reference PRDM1 motif21 (below).

(F) Heatmap showing differential gene expression results from bulk RNA-seq on peripheral blood human T cells from additional RS-R and RS-NR highlighting ZNF683 is associated with PD-1 response.

STAR METHODS TEXT

Resource availability

Lead contact

Further information and requests for resources, reagents and data should be directed to the lead contact, Catherine J. Wu (cwu@partners.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

scRNA sequencing, scTCR sequencing, and bulk TCR sequencing, bulk RNA sequencing and CUT&RUN sequencing data used in this study will be deposited in NCBI's Gene Expression Omnibus (GEO) upon acceptance. Accession numbers will be listed in the key resources table.

This study does not report any original code. Any additional information required to reanalyze the data reported in this study can be obtained from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human samples

Whole bone marrow and whole blood samples were obtained from RS and CLL patients enrolled on clinical trials of nivolumab plus ibrutinib therapy(AbbVie, 2021, 2022) and mononuclear bone marrow samples from control healthy donors were obtained through the tissue bank at Dana-Farber Cancer Institute, approved by and conducted in accordance with the principles of the Declaration of Helsinki and with the approval of the Institutional Review Boards (IRB) of the University of Texas/MD Anderson Cancer Center (MDACC) or of Dana-Farber Cancer Institute (DFCI). Their clinical characteristics are reported in **Table S1**. From clinical trial patients, blood, and marrow tissue samples were collected at baseline, response assessment, and at relapse. Trial samples were obtained in heparin green top tubes and placed on ice after collection. They were mixed at a 1:1 ratio with Freezing media (80% FCS and 20% DMSO) and cryopreserved and stored in liquid nitrogen until analysis. Healthy controls from DFCI underwent mononuclear cell isolation by Ficoll gradient prior to being preserved in freezing media (FCS with final concentration 10% DMSO) and cryopreserved using similar methods. Forty-seven independent fresh DLBCL subtype RS biopsy samples (spleen or lymph node) were obtained from the French FILO (French Innovative Leukemia Organization) cohort (ClinicalTrials.gov Identifier: NCT03619512).

METHOD DETAILS

Flow cytometry sorting

Cells were thawed by drop-wise addition of warmed media (RPMI 10% FCS 1% P/S) and stained with antibodies (Biolegend CD5 FITC, CD19 PE-Cy7, CD3 PB) and 7-AAD (Biolegend) before being resuspended in PBS-0.04% BSA (NEB/Invitrogen). Viable CD5+ CD19+ population was sorted for CLL and viable CD5-CD19- population for immune fraction. In marrow samples where RS tumor was present, RS and CLL fractions were sorted by size based on the increased forward scatter (FSC) of RS cells. The sorting strategy is shown in **Figure 1B** and **Figure S1.**

CITE-seq and TCR-seq

For CITE and TCR-seq, bone marrow samples were thawed into warmed PBS 10% FCS and washed with PBS. Cells were stained with Zombie Violet viability marker (Biolegend) followed by CD19 (Biolegend) and glycophorin (BD CD235) to mark tumor B cells and erythrocyte precursors respectively. The non-stained population (CD19-glycophorinA-) population was sorted for scRNA sequencing. Cells were then pelleted and stained with 55 antibodies (**Table S9**) from BioLegend Total-Seq-C system according to the manufacturer's protocol in PBS-0.04% BSA. After washing, the cells were sequenced using the 5' V2 kit with human T cell V(D)J enrichment (10x Genomics) according to manufacturer's instructions.

Single-cell RNA-sequencing

About 16,000 cells were loaded onto a 10× Genomics Chromium™ instrument (10× Genomics) according to the manufacturer's instructions. The scRNAseq libraries were processed using Chromium Next GEM Single Cell 5' Kit v2 kit or 3' v2 kit (10× Genomics). The sequencing libraries for scRNAseq, TCRseq and CITEseq were normalized to 4nM concentration and pooled. The pooled libraries were sequenced on NovaSeq S4 or HiSeq X platform (Illumina).

Bulk TCR sequencing

Cells were washed after thawing with PBS and RNA was isolated from whole blood after thawing by Qiagen RNeasy kit. Bulk-TCR sequencing was then performed as previously described (Li et al., 2019b). Clonotypes across timepoints were compared using Fisher's exact test (Penter et al., 2021b). Clonotypes determined to be significant with multiple test correction (Bonferroni) were then classified as expanded (increasing frequency in second timepoint), depleted (decreasing frequency in second timepoint), disappeared (no longer detected at second timepoint), or novel (new at second timepoint).

Bulk RNA-sequencing of patient samples

Ten million cells from 11 frozen RS patient samples were thawed directly into 1 mL of MACS buffer (0.5% BSA, mM EDTA in 1X PBS) with 100 uL CD3 MicroBeads (Miltenyi Biotec), and then incubated on ice for 15 minutes. The cells were than washed with 5 mL MACS buffer (1500 RPM spin for 5 minutes) and resuspended in 500 uL MACS. CD3+ T cell isolation was performed according to the CD3 Microbeads protocol and total RNA was extracted from the CD3+ T cells using the Qiagen RNeasy Micro Kit (Qiagen). RNA-sequencing libraries were made using the NEBNext single-cell/low input library Prep Kit for Illumina following the manufacturer manual. Sequencing was performed on an NovaSeq SP platform (Illumina).

***ZNF683* qPCR**

cDNA libraries were synthesized from 500 ng of RNA following the manufacturer's protocol for PrimeScript RT Master Mix for qRT-PCR (Takara Bio) using a SimpliAmp Thermal Cycler. cDNA was diluted with nuclease-free water to reach 76.9 ng/100 uL. Each qPCR reaction contained: 10 uL 2X TaqMan Gene Expression Master Mix (Life Technologies), 1 uL of either

appropriate TaqMan probe (Hs00543184_m1 for ZNF683) or Beta Actin (Life Technologies) or GAPDH (Life Technologies) and 9 uL of cDNA diluted to 76.9 ng/100 uL. A 384-well plate and cover was used to run qPCR reactions on the QuantStudio 6 Flex Real-Time PCR system following the manufacturer's thermal protocol for the TaqMan Gene Expression Master Mix.

Generation of ZNF683-expressing Jurkat cell lines

Sleeping Beauty transposon plasmids containing dox-inducible *ZNF683* were assembled by introducing a gBlock (IDT) construct of codon optimized ZNF683 by cloning. First, pSB-tet-GP (Addgene) was linearized by SfiI and ClaI digestion (NEB) performed overnight at 37^{o}C in CutSmart buffer (NEB) prior to gel electrophoresis and band excision and purification (Promega). PCR was performed using Phusion High Fidelity DNA Polymerase and the following primers were used to add a FLAG-tag to the C-terminal end of the ZNF83 protein after a flexible glycine linker: 5'-AAAGGCCTCTGAGGCCACCATGAAAGAGGAATCAGC-3', 5' -AAGCTTGGCCTGACAGGCCTTTATTTGTCGTCGTCATCCTTGTAGTCTGAACCGCC ATTATTTTGGTCTTGACCCA-3'. PCR was then performed on purified resulting fragment (Promega) using overlap primers F- 5'-CTCGAAAGGCCTCTGAAAGGCCTCTGAGGCCAC-3' and R- 5'-CATCAATGTATCTTATCATGTCTATAAGCTTGGCCTGACAGGC-3' and subsequently Gibson cloning (NEB) was performed. To generate *ZNF683*-expressing cell lines or mock-containing cells lines for overexpression experiments, early passage Jurkat cells E6.1 (ATCC) were nucleofected with 15 ug of pCMV(CAT)T7-SB100 (Addgene) and 15 ug of either pSB-tet-GP-ZNF683-flag or pSB-tet-GP (mock control) using manufacturer protocol (Lonza). Post-nucleofection, after GFP was visible (Day 3), cells containing plasmid integration were selected by puromycin (added at 0.25 ug/mL) for two weeks to obtain stable cell lines prior to addition

of doxycycline and subsequent experiments. Doxycycline (1 ug/mL) was added to cells and the cells were cultured for 48 hours in the presence/absence of doxycycline prior to RNA sequencing and CUT&RUN experiments.

Western blot confirmation of protein expression

ZNF683-expressing Jurkat cells were in culture with doxycycline for 48 hours prior to extraction of nuclear protein (Pierce™ NE-PER® Nuclear and Cytoplasmic Extraction Reagent Kit; Thermo Scientific, #78833), starting with $2x10^6$ cells. Sample preparation and gel loading was performed in accordance to the manufacturer's protocol (NuPAGE Bis-Tris Mini Gels [4-12% Polyacrylamide]; Life Technologies, #NP0321BOX). Protein was then transferred from the gel to a nitrocellulose membrane according to the iBlot 2 Transfer Stacks protocol (Life Technologies, #IB23002) on an iBlot 2 Gel Transfer Device. Ponceau staining confirmed protein transfer to the membrane, after which the membrane was cut to separate the experimental anti-Flag protein and Histone His3 control and both parts of the membrane were blocked in blocking buffer (5% BSA in 1X TBS-T) on a rocker at room temperature for 1 hour. The membranes were then blocked with Histone H3 Antibody (Cell Signaling Technology) and DYKDDDDK Tag Antibody (Cell Signaling Technology) according to the manufacturer's protocol, overnight on a rocker at 4°C. The membranes were washed with 1X TBS-T in three 5-minute intervals on a rocker at room temperature, incubated with the secondary antibody anti-rabbit IgG HRP-linked Antibody (Cell Signaling Technology) for 40 minutes at 4°C, followed by another three 1X TBS-T washes as aforementioned. The membranes were developed according to protocol using SuperSignal™ West Pico PLUS Chemiluminescent Substrate (Thermo Scientific) and exposed for imaging in a Bio-Rad ChemiDoc Imaging System.

RNA sequencing of cell lines

Cells were washed in PBS and total RNA was extracted using Qiagen RNAeasy Mini Kit (Qiagen) from at least $1x10^6$ cells. For *ZNF683* to mock comparison, 3 independent cell line replicates per condition each underwent extraction. 200ng of total RNA (RIN quality >8) was utilized for library construction. For dox-on and dox-off experiments, 100 ng of total RNA (RIN >8) was utilized for library construction. For cDNA libraries construction, total RNA was quantified using the Quant-iT RiboGreen RNA Assay Kit (Thermo Scientific) and normalized to 5 ng/µl. The cDNA libraries were prepared using NEBnext single cell/Low input RNA library prep kit for Illumina (NEB), followed by cDNA fragmentation and adaptor ligation. For *ZNF683* overexpression compared to mock, single-indexes were used as previously described (Biran et al., 2021). For dox on-off experiment, dual indexes were utilized as previously described (Li et al., 2019b). After Ampure beads cleaning, final sequencing libraries were quantified using a High Sensitivity DNA Kit on Bioanalyzer (Agilent). RNA-seq libraries were normalized to 4nM concentration and pooled before loading onto an Illumina sequencer, with the Next-seq 75 used for mock vs. dox-induced *ZNF683* and NovaSeq SP used for dox-on vs. dox-off. Data was analyzed as described above.

CUT&RUN

CUT&RUN was performed on Jurkat cells either with or without *ZNF683* expression, as per the CUTANA™ ChIC/CUT&RUN Kit (EpiCypher) protocol; conditions were optimized for a light cross-link condition (0.1% formaldehyde for 1 minute) and 0.001% Digitonin for cell permeabilization. The following antibodies were used: anti-FLAG (Fisher), The manufacturer's protocol for NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (NEB) was followed for DNA library prep, except for modifications as instructed by the CUT&RUN protocol. NEBNext Multiplex Oligos for Illumina (NEB) were used as dual index primers for the DNA

library. The Illumina MiniSeq sequencing platform and the MiniSeq High Output Kit (150 cycles) were used to sequence the DNA library (Illumina).

RNA-sequencing of RS nodal tissue

RNA was extracted with Macherey Nagel RNA extraction kit (Macherey-Nagel, Düren, Germany). Total RNA-Seq libraries were generated from 500 ng of total RNA using TruSeq Stranded Total RNA LT Sample Prep Kit with Ribo-Zero Gold (Illumina, San Diego, CA), according to manufacturer's instructions. The final cDNA libraries were checked for quality and quantified using capillary electrophoresis prior to sequencing with Hiseq 4000 sequencing using 1x50 bases protocol.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data processing of scRNA-seq libraries

scRNA-seq reads were processed, aligned to the Hg19 reference genome, and filtered using the Cell Ranger pipeline v2.0.0/v2.1.1 from 10x Genomics. Filtered feature-barcode matrices containing detected cellular barcodes were used in further analysis. Initially, each sample was processed individually with Pagoda2 (Fan et al., 2016). This included quality control and count normalization. Cells were filtered based on size with a minimum cell size of 500 UMIs and maximum of 8,000 UMIs (12,000 for RS samples due their larger cell size). Cells with more than 10% mitochondrial content and genes with low expression (detected in 10 cells or fewer) were removed. Across all samples, a total of 196,191 cells passed these filters and were subject to further analysis of separate tumor and immune compartments.

Data processing of CITE-seq and TCR libraries

CITE-seq and TCR reads were aligned to the Hg19 and GRCh38 reference genome, respectively, and transcripts quantified using the Cell Ranger pipeline v3.0.2 from 10x Genomics. Further processing of gene expression data was as described above. Protein expression data was processed per sample and normalized by the summed count of isotype controls (IgG1, IgG2a, IgG2b antibodies). Total number of cells captured with linked gene and protein data is 8,062 with an average 1365 genes/cell. TCRs were captured for 7,235 of these cells.

Joint clustering

In order to compare cell proportions and states across multiple time points and samples, we set out to combine our data using *Conos* v1.3.1 (Barkas et al., 2019). Joint clusters were subsequently detected using the Leiden community detection method (Traag et al., 2019). For analysis of the immune cells, we performed a subclustering of lymphocytes with a resolution to 1.5 and detected 11 transcriptionally distinct clusters. Marker genes for each cluster were computed and these were used along with expression of canonical markers to annotate clusters. Differential gene expression between responders and non-responders was performed using DESeq2 v1.22.2 (Love et al., 2014). As a single cell might not be a truly independent observation, we decided to concatenate by summation count data within a cluster per sample, thus generating a pseudo-bulk RNA sequencing count matrix. This naturally generated higher counts indicating a greater confidence that a given gene was observed. DESeq2 has its own normalization step that accounts for differences in library size, and therefore we used raw count data as input for all DESeq2 runs. This normalization accounts for instances with uneven sample contribution to a cluster.

Concurrently, we performed joint analysis of tumor samples, and detected 12 distinct clusters with four of them characterized as immune cells (spillover from sorting) using a resolution of 1.0. Only clusters consisting of malignant B cells were considered in subsequent analysis and subclustered. We performed differential gene expression between RS and CLL cells using DESeq2 with the same approach as described earlier. Gene set enrichment analysis (GSEA) (Subramanian et al., 2005) was run with the R package *fgsea* v1.13.2 (Korotkevich et al., 2021) on the ranked list of log fold changes using the C5 and Hallmark gene sets from the molecular signature database (MSigDb) (Liberzon et al., 2011, 2015),(Subramanian et al., 2005))

Addition of normals and CITE-seq samples

Lymphocytes from additional samples, including 28 normal marrows (Oetjen et al., 2018; Rozenblatt-Rosen et al., 2017) and 4 CITE-seq samples (Oetjen et al., 2018) were included in our lymphocyte sub-clustering by label propagation (55,016 and 7,954 cells). This was done by computing the probabilities for each cell belonging to each of the 11 clusters. Using the normalized protein-linked expression from CITE-seq samples, we further confirmed our cluster annotations. For statistical comparisons of populations, 2 sample T-tests were used to compare cell proportions between RS-R and RS-NR and between patients (pre-therapy) and healthy controls. For healthy control data, technical replicates were averaged and included, and the first (earliest) sample of biological replicates was used in comparisons.

Single-cell TCR repertoire analysis

Cell barcodes with corresponding alpha and beta chain nucleotides sequences were extracted and only cells with productive TCRs were considered. Productive TCRs were defined as having either one alpha chain and one beta chain, two alpha chains and one beta chain, or one alpha chain and two beta chains. Of all TCRs 84% and 83% were defined as productive for RT-R2

and RT-NR1, respectively. For the responder and non-responder, we summarized and evaluated the unique clonotype frequencies pre and post therapy. Clonotypes were categorized based on frequencies using *scRepertoire* v1.0.0 (Borcherding et al., 2020).

Trajectory analysis

We performed trajectory analysis on our 4 CD8 T cell clusters (cluster 1, 3, 4 and 8) using slingshot v1.8.0 (Street et al., 2018). To construct the trajectory, we set cluster 3 as the starting point, and ran the analysis with ~points = 150 to reduce compute time.

Bulk RNA-sequencing of patient samples and cell lines

Abundances of transcripts from the bulk RNA sequencing reads were quantified using kallisto v0.46.0 (Bray et al., 2016). Transcripts were summarized to gene level, and differential gene expression analysis was performed using DESeq2 v1.22.2.

CUT&RUN

We removed adapter sequences from the paired-end reads with Cutadapt v3.5 (Martin, 2011). Reads were then aligned to Hg19 reference genome with Bowtie2 v.2.4.2 (Langmead and Salzberg, 2012). Alignments were filtered and sorted using Samtools v1.2 (Danecek et al., 2021), converted to bed format and subsequently to bedgraphs using bedtools v2.30.0 (Quinlan and Hall, 2010). Peaks from bedgraphs were called using SEACR v1.3 (Meers et al., 2019). We used bamCoverage from deepTools v3.5.1 (Ramírez et al., 2016) to generate bigWig files to visualize peaks in IGV tools. We used DiffBind for differential peak analysis v3.4.3 (Stark and Brown).

CISTROME-GO

To analyze the functional enrichment of the transcription factor peaks generated by CUT&RUN results for *ZNF683* overexpressed Jurkat cells, the Cistrome-GO (Li et al., 2019a) online tool was applied to the results with the following parameters: Human (GRCh37/hg19) genome, Ensemble mode, all genes for GO, 1.0 cutoff of logFoldChange, 0.1 FDR for DE genes, top 10,000 peaks in the peak BED file generated from CUT&RUN data, 10.0 kb half-decay distance, 0.2 FDR cutoff of GO/KEGG terms to return, minimum of 10 genes to maximum of 2,000 genes in GO and KEGG gene sets.

Motif analysis

We performed a comprehensive motif analysis of CUT&RUN peaks using MEME-ChIP from the MEME Suite (Bailey et al., 2015). We used differential peaks with an FDR < 0.5 with default parameters and the human database (Kulakovskiy et al., 2018) to enable detection of recurrent peaks. To directly compare top MEME-ChIP result **with PRDM1** (Bailey et al., 2015), TomTom (Street et al., 2018) was used with default parameters.

Comparisons with existing datasets

To compare our signatures with other existing datasets, we first used singleR v1.4.0 (Aran et al., 2019) to annotate our CD8+ T cells with labels from existing melanoma single-cell data (Fairfax et al., 2020). Second, we ran single-sample GSEA (Barbie et al., 2009) (ssGSEA) using GSVA v1.30.0 (Hänzelmann et al., 2013) on a per cluster basis on our CD8+ T cells using the mean expression. We included the following gene sets for ssGSEA: cytotoxicity and exhaustion (Oliveira et al., 2021) and top 50 most significantly expressed genes per CD8+ T cell cluster (Anadon et al., 2022; Caushi et al., 2021; Fairfax et al., 2020; Zheng et al., 2021).

Analysis of ZNF683 signature in external datasets

DESeq2-normalized TPMs from the RNA-seq from 35 RS nodal tissue samples was examined for enrichment of our cluster 1 signature (top 50 genes, $p_{adj} < 0.05$) using ssGSEA with GSVA v1.30.0. Additionally, we examined expression of ZNF683 normalized by estimated total T cell fraction of each sample using CIBERSORT v1.05 (Newman et al., 2015). RNA-seq data from 81 treatment-naïve non-CD19 selected RNA-seq samples from the CLLmap cohort (phs00435) was examined for relative ZNF683 expression. TPMs from RNA-SeQC v2.3.6 (Graubert et al., 2021) were DESeq2-normalized and further normalized by total T-cell content (per CIBERSORT v1.0.5). Expression analysis showed a bimodal ZNF683 expression distribution with 20% (n=18) in the higher group and 80% (n=63) in the lower. Kaplan-Meier curves were used to display overall survival and assessed with the log rank test. TCGA datasets was then analyzed with similar normalization by total T-cell content (per CIBERSORT v1.0.5) followed by Kaplan-Meier curves for the top and bottom 50%.

Hypergeometric test

To examine whether our cluster 1 signature overlapped with a parallel T-cell cluster (Fairfax et al melanoma citation) derived from melanoma PD-1 CPB responders (n=8), we performed a hypergeometric test on the overlap between cluster 1 signature and the set of significantly upregulated genes in the parallel cluster (parameters: q=103, overlapping genes -1, m=163 genes in cluster 1 signature, n=7128 genes across clusters-163, k=279 genes in parallel cluster).

SUPPLEMENTARY ITEM TITLES

Supplementary Figure Legends

Supplementary Figure 1. Sorting strategy for FACS of RS and CLL

(A) Experimental schema showing flow cytometry sorting strategy for bone marrows b, Sorting strategy for CLL malignant B cell fraction of patient bone marrow samples.

(B) Sorting strategy for immune (non-malignant) cell fraction for RS and CLL cells.

Supplementary Figure 2. Single-cell RNA-sequencing metrics

(A) – (D) Malignant B cell QC metrics per sample before filtering:

(A) cells/sample

(B) percent mitochondrial content/cell per sample

(C) gene/cell per sample

(D) UMI/cell (3rd row)

(E) – (H) Immune fraction QC metrics per sample:

(E) cells/sample

(F) percent mitochondrial content per cell per sample

(G) gene/cell per sample

(H) UMI/cell per sample

Supplementary Figure 3. Joint clustering of tumor and immune cells

(A) Conos joint graph using largeVis embedding of sorted tumor fraction clusters A-O.

(B) Individual graph embeddings of each sample showing clusters A-O.

(C) Malignant B cell cluster cell proportions across sorted tumor fractions.

(D) Conos joint graph embedding showing RS (purple), CLL (blue), and immune (gray) populations.

(E) Conos joint graph embeddings showing expression of Class I and representative Class II and non-classical HLA genes in CLL and RS. Cells expressing, red.

(F) Conos joint graph embeddings showing immune marker genes expressed by malignant B cells. Cells expressing, red.

Supplementary Figure 4. Clustering of immune cells

(A) Conos joint graph embedding of immune clusters.

(B) Joint graph of immune clusters by sample.

(C) Joint graph colored by sample.

(D) Marker gene panels highlight major cell types within immune clusters. Red highlights cells expressing labelled marker gene.

Supplementary Figure 5. Lymphocyte sub-clustering

(A) Marker gene panels highlight major cell types within lymphocyte sub-clustering on Conos largeVis embedding. Red highlights cells expressing marker gene.

(B) Flow sorting strategy for CITE-seq experiment. FSC-A, forward scatter.

(C) Bar graphs showing population distributions across serial CLL and RS marrow samples and normal bone marrow samples.

(D) Individual panel graphs of patient marrow samples and normal marrow colored by cluster.

(E) Bar graphs showing population distributions for 5' single-cell RNA-seq performed in conjunction with CITE-seq for marrow samples of RS-R2 (purple) and RS-NR1 (blue) pre- and post- treatment.

Supplementary Figure 6. T cell populations associated with CPB response
(A) Conos joint graph of immune sub-cluster by sample.

(B) LargeVis of Conos joint graph of immune sub-cluster by sample and sample type.

(C) Bubble plot displaying ZNF683 and TOX expression in RS-R vs RS-NR.

(D) – (F) Volcano plots for CD8 T cell clusters (1, 2, 9) displaying gene expression differences within each cluster between non-responder (left) and responder (right).

(G) Heatmap displaying genes differentially regulated between RS-R and RS-NR in across all CD8 clusters. *ZNF683* and *TOX* are labelled among top differentially expressed genes**.**

(H) *TCF7* expression by violin plot is stable across all samples in memory cluster 3.

(I) Panel of cluster distribution for individual samples in 5' scRNA-seq performed with CITE-seq and TCR-seq (top). Top 2 expanded TCR clones in each sample are depicted on the bottom panel (bottom).

Supplementary Figure 7. ZNF683 expression in primary T cells and overexpression in Jurkat cell lines

(A) Line graph shows endogenous ZNF683 in primary T cells from 3 healthy donors expanded in culture following CD3-CD28 bead stimulation.

(B) Flow cytometry analysis of ZNF683-expressing Jurkats after puromycin selection shows stable cell lines express GFP.

(C) Western blot confirms dox-induced ZNF683 expression in Jurkat cells, as measured by FLAG protein expression.

(D) Volcano plot for gene expression differences between Jurkat cells containing dox-inducible ZNF683 vector vs mock (luciferase) vector by RNA-seq.

(E) CUT&RUN on Jurkat cell lines shows binding of ZNF683 at regions surrounding key immune genes (purple) that correspond to differential ATAC-seq peaks in T cell subsets (green) (Philip et al., 2017) and prior PRDM1 ChIP-seq data (black) (Davis et al., 2018). CUT&RUN controls depicted in black in top three tracks of each panel.

Supplemental table titles

S1: Clinical data for study patients, related to Figure 1

S2: Quality control metrics for scRNA-seq data, related to Figure S2

S3: Single-cell tumor cluster proportions, gene expression and GSEA, related to figure 1

S4: Single-cell cluster markers and cell proportions, related to Figure 2

S5: Differential gene expression between RS-R and RS-NR, related to Figure 3

S6: TCR sequencing and clonotype information, related to Figure 3

S7: GSEA enrichment analysis on bulk patient RNA-seq, related to Figure 4

S8: CUT&RUN and RNA-seq data, related to Figure 5

S9: CITE-seq antibody list, related to STAR Methods

REFERENCES

Alfei, F., Kanev, K., Hofmann, M., Wu, M., Ghoneim, H.E., Roelli, P., Utzschneider, D.T., von Hoesslin, M., Cullen, J.G., Fan, Y., et al. (2019). TOX reinforces the phenotype and longevity of exhausted T cells in chronic viral infection. Nature *571*, 265–269. https://doi.org/10.1038/s41586-019-1326-9.

Anadon, C.M., Yu, X., Hänggi, K., Biswas, S., Chaurio, R.A., Martin, A., Payne, K.K., Mandal, G., Innamarato, P., Harro, C.M., et al. (2022). Ovarian cancer immunogenicity is governed by a narrow subset of progenitor tissue-resident memory T cells. Cancer Cell *40*, 545-557.e13. https://doi.org/10.1016/j.ccell.2022.03.008.

Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172. https://doi.org/10.1038/s41590-018-0276-y.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. Nucleic Acids Res. *43*, W39-49. https://doi.org/10.1093/nar/gkv416.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112. https://doi.org/10.1038/nature08460.

Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K., and Kharchenko, P.V. (2019). Joint analysis of heterogeneous single-cell RNA-seq dataset collections. Nat. Methods *16*, 695–698. https://doi.org/10.1038/s41592-019-0466-z.

Beltra, J.-C., Manne, S., Abdel-Hakeem, M.S., Kurachi, M., Giles, J.R., Chen, Z., Casella, V., Ngiow, S.F., Khan, O., Huang, Y.J., et al. (2020). Developmental Relationships of Four Exhausted CD8+ T Cell Subsets Reveals Underlying Transcriptional and Epigenetic Landscape Control Mechanisms. Immunity *52*, 825-841.e8. https://doi.org/10.1016/j.immuni.2020.04.014.

Biran, A., Yin, S., Kretzmer, H., Ten Hacken, E., Parvin, S., Lucas, F., Uduman, M., Gutierrez, C., Dangle, N., Billington, L., et al. (2021). Activation of Notch and Myc Signaling via B-cell-Restricted Depletion of Dnmt3a Generates a Consistent Murine Model of Chronic Lymphocytic Leukemia. Cancer Res. *81*, 6117–6130. https://doi.org/10.1158/0008-5472.CAN-21-1273.

Borcherding, N., Bormann, N.L., and Kraus, G. (2020). scRepertoire: An R-based toolkit for single-cell immune receptor analysis. F1000Research *9*, 47. https://doi.org/10.12688/f1000research.22139.2.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. https://doi.org/10.1038/nbt.3519.

Caushi, J.X., Zhang, J., Ji, Z., Vaghasia, A., Zhang, B., Hsiue, E.H.-C., Mog, B.J., Hou, W., Justesen, S., Blosser, R., et al. (2021). Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers. Nature *596*, 126–132. https://doi.org/10.1038/s41586-021-03752-4.

Chen, Z., Ji, Z., Ngiow, S.F., Manne, S., Cai, Z., Huang, A.C., Johnson, J., Staupe, R.P., Bengsch, B., Xu, C., et al. (2019). TCF-1-centered transcriptional network drives an effector versus exhausted CD8 T cell fate decision. Immunity *51*, 840-855.e5. https://doi.org/10.1016/j.immuni.2019.09.013.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801. https://doi.org/10.1093/nar/gkx1081.

Ding, W., LaPlant, B.R., Call, T.G., Parikh, S.A., Leis, J.F., He, R., Shanafelt, T.D., Sinha, S., Le-Rademacher, J., Feldman, A.L., et al. (2017). Pembrolizumab in patients with CLL and

Richter transformation or with relapsed CLL. Blood *129*, 3419–3427. https://doi.org/10.1182/blood-2017-02-765685.

Fairfax, B.P., Taylor, C.A., Watson, R.A., Nassiri, I., Danielli, S., Fang, H., Mahé, E.A., Cooper, R., Woodcock, V., Traill, Z., et al. (2020). Peripheral CD8+ T cell characteristics associated with durable responses to immune checkpoint blockade in patients with metastatic melanoma. Nat. Med. *26*, 193–199. https://doi.org/10.1038/s41591-019-0734-6.

Fan, J., Salathia, N., Liu, R., Kaeser, G.E., Yung, Y.C., Herman, J.L., Kaper, F., Fan, J.-B., Zhang, K., Chun, J., et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat. Methods *13*, 241–244. https://doi.org/10.1038/nmeth.3734.

Graubert, A., Aguet, F., Ravi, A., Ardlie, K.G., and Getz, G. (2021). RNA-SeQC 2: Efficient RNA-seq quality control and quantification for large cohorts. Bioinforma. Oxf. Engl. btab135. https://doi.org/10.1093/bioinformatics/btab135.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 7. https://doi.org/10.1186/1471-2105-14-7.

Jain, N., Ferrajoli, A., Basu, S., Thompson, P.A., Burger, J.A., Kadia, T.M., Estrov, Z.E., Pemmaraju, N., Lopez, W., Thakral, B., et al. (2018). A Phase II Trial of Nivolumab Combined with Ibrutinib for Patients with Richter Transformation. Blood *132*, 296. https://doi.org/10.1182/blood-2018-99-120355.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast gene set enrichment analysis. 060012. https://doi.org/10.1101/060012.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46*, D252–D259. https://doi.org/10.1093/nar/gkx1106.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

Li, S., Wan, C., Zheng, R., Fan, J., Dong, X., Meyer, C.A., and Liu, X.S. (2019a). Cistrome-GO: a web server for functional enrichment analysis of transcription factor ChIP-seq peaks. Nucleic Acids Res. *47*, W206–W211. https://doi.org/10.1093/nar/gkz332.

Li, S., Sun, J., Allesøe, R., Datta, K., Bao, Y., Oliveira, G., Forman, J., Jin, R., Olsen, L.R., Keskin, D.B., et al. (2019b). RNase H-dependent PCR-enabled T-cell receptor sequencing for highly specific and efficient targeted sequencing of T-cell receptor mRNA for single-cell and repertoire analysis. Nat. Protoc. *14*, 2571–2594. https://doi.org/10.1038/s41596-019-0195-x.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

Mackay, L.K., Minnich, M., Kragten, N.A.M., Liao, Y., Nota, B., Seillet, C., Zaid, A., Man, K., Preston, S., Freestone, D., et al. (2016). Hobit and Blimp1 instruct a universal transcriptional program of tissue residency in lymphocytes. Science *352*, 459–463. https://doi.org/10.1126/science.aad2035.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

Mazo, I.B., Honczarenko, M., Leung, H., Cavanagh, L.L., Bonasio, R., Weninger, W., Engelke, K., Xia, L., McEver, R.P., Koni, P.A., et al. (2005). Bone marrow is a major reservoir and site of recruitment for central memory CD8+ T cells. Immunity *22*, 259–270. https://doi.org/10.1016/j.immuni.2005.01.008.

McLane, L.M., Abdel-Hakeem, M.S., and Wherry, E.J. (2019). CD8 T Cell Exhaustion During Chronic Viral Infection and Cancer. Annu. Rev. Immunol. *37*, 457–495. https://doi.org/10.1146/annurev-immunol-041015-055318.

Meers, M.P., Tenenbaum, D., and Henikoff, S. (2019). Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. Epigenetics Chromatin *12*, 42. https://doi.org/10.1186/s13072-019-0287-4.

Mercier, F.E., Ragu, C., and Scadden, D.T. (2011). The bone marrow at the crossroads of blood and immunity. Nat. Rev. Immunol. *12*, 49–60. https://doi.org/10.1038/nri3132.

Miller, B.C., Sen, D.R., Al Abosy, R., Bi, K., Virkud, Y.V., LaFleur, M.W., Yates, K.B., Lako, A., Felt, K., Naik, G.S., et al. (2019). Subsets of exhausted CD8+ T cells differentially mediate tumor control and respond to checkpoint blockade. Nat. Immunol. *20*, 326–336. https://doi.org/10.1038/s41590-019-0312-6.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457. https://doi.org/10.1038/nmeth.3337.

Oetjen, K.A., Lindblad, K.E., Goswami, M., Gui, G., Dagur, P.K., Lai, C., Dillon, L.W., McCoy, J.P., and Hourigan, C.S. (2018). Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. JCI Insight *3*. https://doi.org/10.1172/jci.insight.124928.

Oliveira, G., Stromhaug, K., Klaeger, S., Kula, T., Frederick, D.T., Le, P.M., Forman, J., Huang, T., Li, S., Zhang, W., et al. (2021). Phenotype, specificity and avidity of antitumour CD8+ T cells in melanoma. Nature *596*, 119–125. https://doi.org/10.1038/s41586-021-03704-y.

Parikh, S.A., Kay, N.E., and Shanafelt, T.D. (2014). How we treat Richter syndrome. Blood *123*, 1647–1657. https://doi.org/10.1182/blood-2013-11-516229.

Penter, L., Gohil, S.H., Lareau, C., Ludwig, L.S., Parry, E.M., Huang, T., Li, S., Zhang, W., Livitz, D., Leshchiner, I., et al. (2021a). Longitudinal Single-Cell Dynamics of Chromatin Accessibility and Mitochondrial Mutations in Chronic Lymphocytic Leukemia Mirror Disease History. Cancer Discov. *11*, 3048–3063. https://doi.org/10.1158/2159-8290.cd-21-0276.

Penter, L., Zhang, Y., Savell, A., Huang, T., Cieri, N., Thrash, E.M., Kim-Schulze, S., Jhaveri, A., Fu, J., Ranasinghe, S., et al. (2021b). Molecular and cellular features of CTLA-4 blockade for relapsed myeloid malignancies after transplantation. Blood *137*, 3212–3217. https://doi.org/10.1182/blood.2021010867.

Philip, M., Fairchild, L., Sun, L., Horste, E.L., Camara, S., Shakiba, M., Scott, A.C., Viale, A., Lauer, P., Merghoub, T., et al. (2017). Chromatin states define tumour-specific T cell dysfunction and reprogramming. Nature *545*, 452–456. https://doi.org/10.1038/nature22367.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. Oxf. Engl. *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. https://doi.org/10.1093/nar/gkw257.

Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., and Teichmann, S.A. (2017). The Human Cell Atlas: from vision to reality. Nature *550*, 451–453. https://doi.org/10.1038/550451a.

Sade-Feldman, M., Yizhak, K., Bjorgaard, S.L., Ray, J.P., de Boer, C.G., Jenkins, R.W., Lieb, D.J., Chen, J.H., Frederick, D.T., Barzily-Rokni, M., et al. (2018). Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. Cell *175*, 998-1013.e20. https://doi.org/10.1016/j.cell.2018.10.038.

Siddiqui, I., Schaeuble, K., Chennupati, V., Fuertes Marraco, S.A., Calderon-Copete, S., Pais Ferreira, D., Carmona, S.J., Scarpellino, L., Gfeller, D., Pradervand, S., et al. (2019). Intratumoral Tcf1+PD-1+CD8+ T Cells with Stem-like Properties Promote Tumor Control in Response to Vaccination and Checkpoint Blockade Immunotherapy. Immunity *50*, 195-211.e10. https://doi.org/10.1016/j.immuni.2018.12.021.

Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. ELife *6*, e21856. https://doi.org/10.7554/eLife.21856.

Stark, R., and Brown, G. DiffBind: Differential binding analysis of ChIP-Seq peak data. 75. .

Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics *19*, 477. https://doi.org/10.1186/s12864-018-4772-0.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196. https://doi.org/10.1126/science.aad0501.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233. https://doi.org/10.1038/s41598-019-41695-z.

Vieira Braga, F.A., Hertoghs, K.M.L., Kragten, N.A.M., Doody, G.M., Barnes, N.A., Remmerswaal, E.B.M., Hsiao, C.-C., Moerland, P.D., Wouters, D., Derks, I.A.M., et al. (2015). Blimp-1 homolog Hobit identifies effector-type lymphocytes in humans. Eur. J. Immunol. *45*, 2945–2958. https://doi.org/10.1002/eji.201545650.

Yin, S., Gambe, R.G., Sun, J., Martinez, A.Z., Cartun, Z.J., Regis, F.F.D., Wan, Y., Fan, J., Brooks, A.N., Herman, S.E.M., et al. (2019). A Murine Model of Chronic Lymphocytic Leukemia Based on B Cell-Restricted Expression of Sf3b1 Mutation and Atm Deletion. Cancer Cell *35*, 283-296.e5. https://doi.org/10.1016/j.ccell.2018.12.013.

Younes, A., Brody, J., Carpio, C., Lopez-Guillermo, A., Ben-Yehuda, D., Ferhanoglu, B., Nagler, A., Ozcan, M., Avivi, I., Bosch, F., et al. (2019). Safety and activity of ibrutinib in combination with nivolumab in patients with relapsed non-Hodgkin lymphoma or chronic lymphocytic leukaemia: a phase 1/2a study. Lancet Haematol. *6*, e67–e78. https://doi.org/10.1016/S2352-3026(18)30217-5.

Zhang, Y., and Zhang, Z. (2020). The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. Cell. Mol. Immunol. *17*, 807–821. https://doi.org/10.1038/s41423-020-0488-6.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

Zheng, L., Qin, S., Si, W., Wang, A., Xing, B., Gao, R., Ren, X., Wang, L., Wu, X., Zhang, J., et al. (2021). Pan-cancer single-cell landscape of tumor-infiltrating T cells. Science *374*, abe6474. https://doi.org/10.1126/science.abe6474.

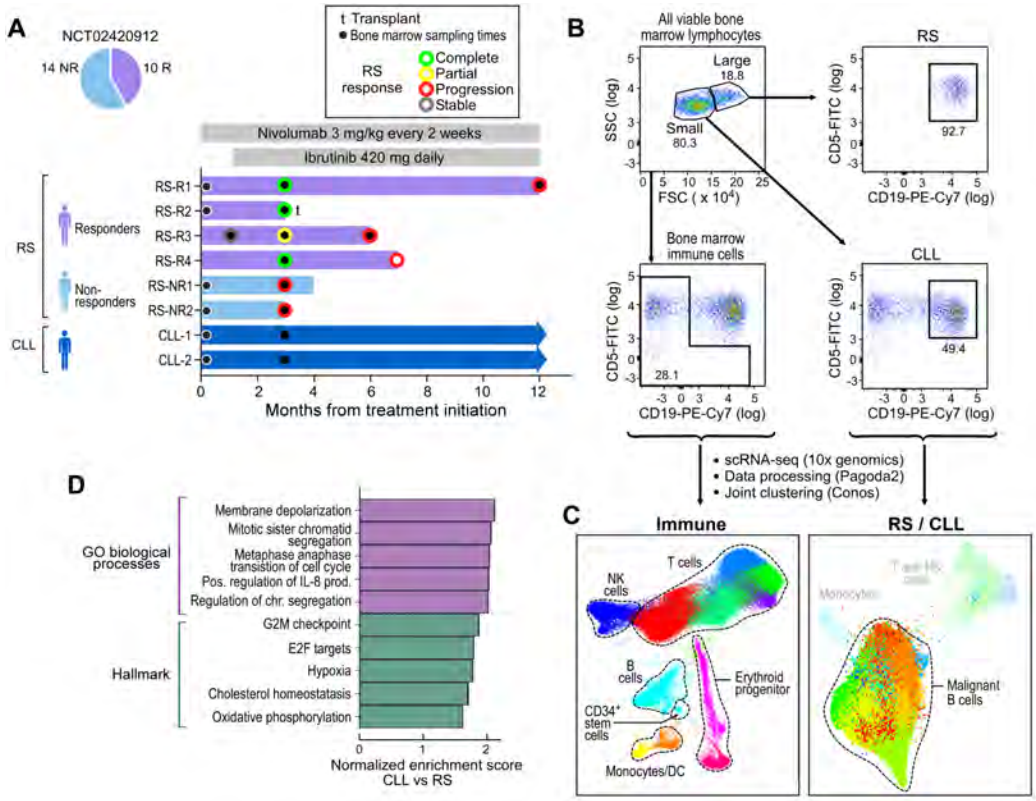
Key resources table

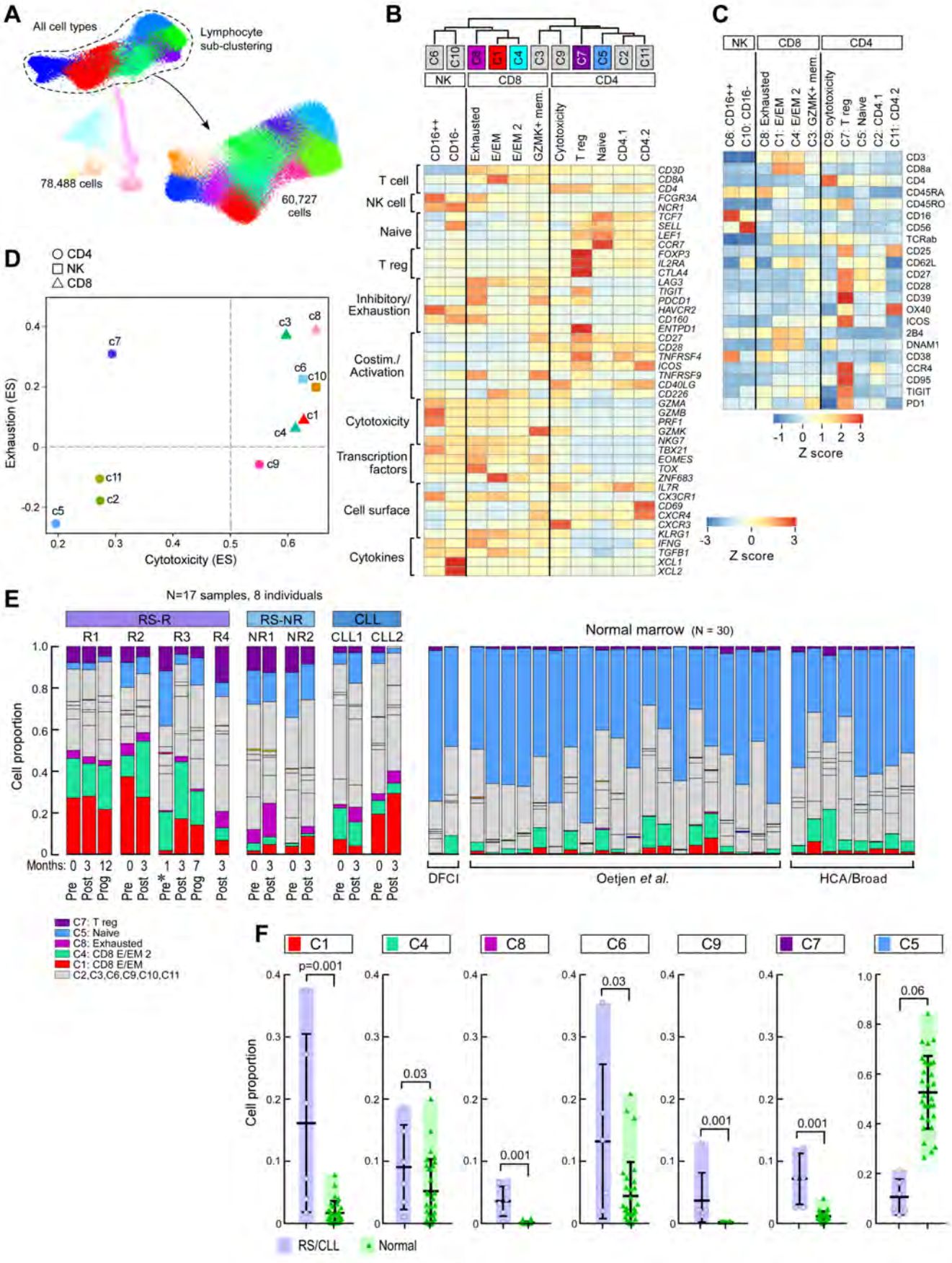
REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
anti-CD19 (PE-Cy7)	Biolegend	Cat# 302216
anti-CD5 (BV421)	Biolegend	Cat# 300626
anti-CD4 (FITC)	Biolegend	Cat# 300506
anti-CD3 (Pacific Blue)	Biolegend	Cat# 300330
anti-CD5 (FITC)	Biolegend	Cat# 364022
anti-CD19 (PE-Cy7)	Biolegend	Cat# 302216
anti-CD235	BD Biosciences	Cat# 555570
7-AAD Viability Staining Solution	Biolegend	Cat# 420404
Zombie Violet Fixable Viability Kit	Biolegend	Cat# 423114
CITE-seq antibodies in Supplementary Table 9		
Histone H3 antibody	Cell Signaling Technology	Cat# 9715S
DYKDDDDK Tag Antibody	Cell Signaling Technology	Cat# 2368S
HRP-linked Antibody	Cell Signaling Technology	Cat# 7074S
CUTANA™ ChIC/CUT&RUN Kit H3K4Me3, IgG controls	EpiCypher	Cat# 14-1048
anti-FLAG	Fisher	Cat# FG4R
CD3 MicroBeads	Miltenyi Biotec	Cat# 130-050-101
Bacterial and virus strains		
Biological samples		
Whole bone marrow and blood samples during clinical trial	Clinical trial	NCT03619512 (https://clinicaltrials.gov/ct2/show/NCT03619512)
Chemicals, peptides, and recombinant proteins		
Doxycycline	Takara Bio	631311
Puromycin	Life Technologies	A1113803
SuperSignal™ West Pico PLUS Chemiluminescent Substrate	Thermo Scientific	34580
PrimeScript RT Master Mix	Takara Bio	RR036A
TaqMan Gene Expression Master Mix	Life Technologies	4369514
Critical commercial assays		
Chromium Next GEM Single Cell 5' Kit v2 kit	10x genomics	1000244
Chromium Next GEM Single Cell 3' v2 kit	10x genomics	PN-120237
RNeasy kit	Qiagen	74106
Qiagen RNeasy Micro Kit	Qiagen	74004
NEBNext single-cell/low input library Prep Kit for Illumina	NEB	E6420L
PrimeScript RT Master Mix for qRT-PCR	Takara Bio	RR036A
Pierce™ NE-PER® Nuclear and Cytoplasmic Extraction Reagent Kit	Thermo Scientific	78833
Quant-iT RiboGreen RNA Assay Kit	Thermo Scientific	R11490
High Sensitivity DNA Kit	Agilent	5067-4626

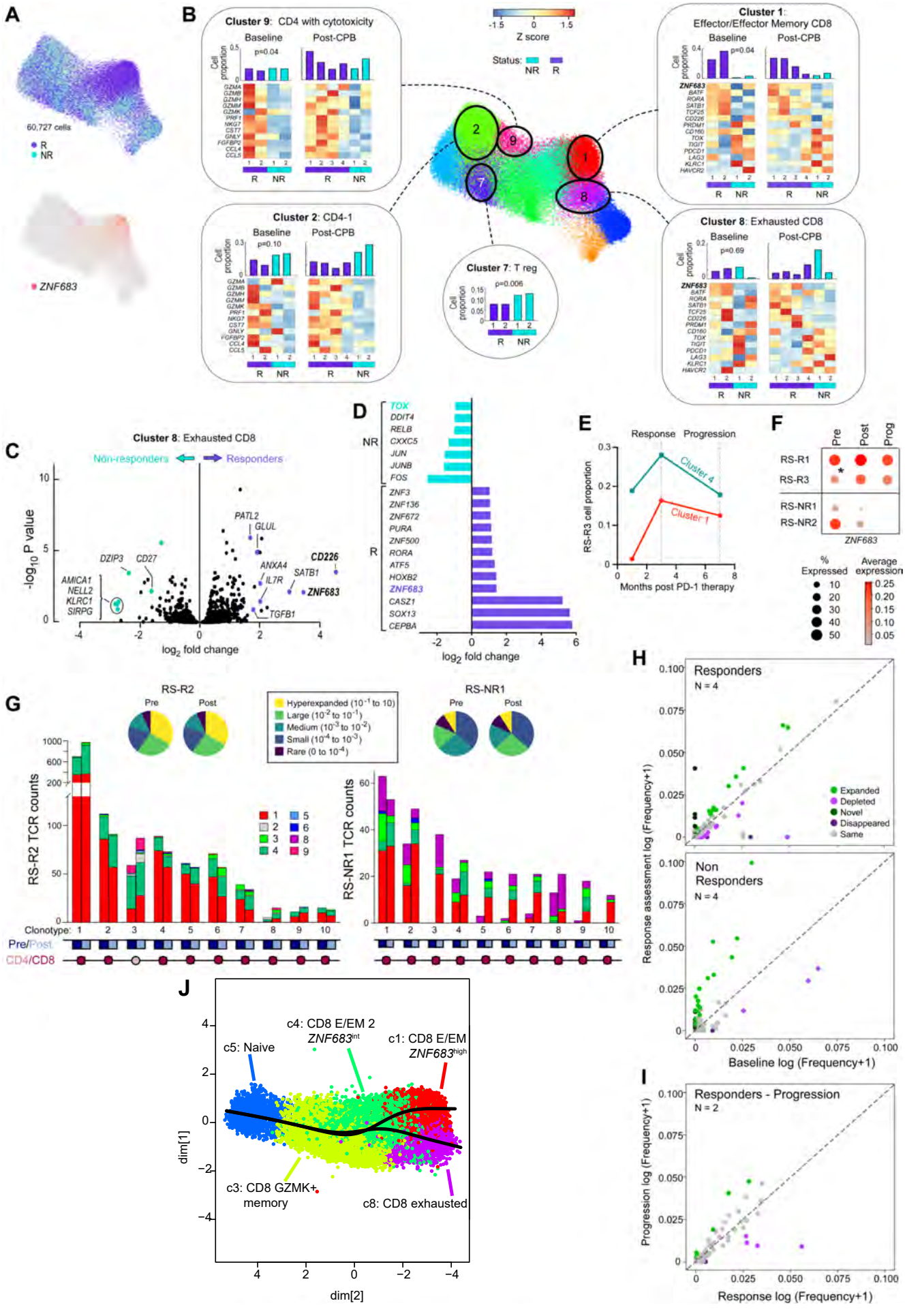
TruSeq Stranded Total RNA LT Sample Prep Kit with Ribo-Zero Gold	Illumina	20020598
Wizard SV Gel and PCR Clean-Up System	Promega	A9282
NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®	NEB	E7645S/L
MiniSeq High Output Kit (150 cycles)	Illumina	FC-420-1002
Deposited data		
Single-cell RNA sequencing data	This study	dbGAP link upon acceptance
Bulk RNA sequencing data	This study	dbGAP link upon acceptance
Single-cell TCR sequencing	This study	dbGAP link upon acceptance
Bulk TCR sequencing	This study	dbGAP link upon acceptance
CUT&RUN	This study	dbGAP link upon acceptance
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
CLL Transcriptome Data	dbGAP	phs000435
TCGA RNA-seq Sample and Clinical Data	The Cancer Genome Atlas Research Network	https://portal.gdc.cancer.gov
TCGA RNA-seq Data	Broad Institute GDAC Firehose	https://gdac.broadinstitute.org
CHiP-seq, Davis et al.	Cistrome-GO data browser	ENCSR098YLE_2
ATAC-seq, Philip et al.	Gene Expression Omnibus	GSM2365846, GSM2365852, GSM2365855, GSM2365856, GSM2365857
Experimental models: Cell lines		
Jurkat cell line, clone E6-1	ATCC	TIB-152
Experimental models: Organisms/strains		
Oligonucleotides		
5'-AAAGGCCTCTGAGGCCACCATGAAAGAGGAATCAGC-3'	This paper	FlagPCR F
5' -AAGCTTGGCCTGACAGGCCTTTATTTGTCGTCGTCATCCTTGTAGTCTGAACCGCCATTATTTTGGTCTTGACCCA-3'	This paper	FlagPCR R
5'- CTCGAAAGGCCTCTGAAAGGCCTCTGAGGCCAC-3'	This paper	Gibson overlap F
5'-CATCAATGTATCTTATCATGTCTATAAGCTTGGCCTGACAGGC-3'	This paper	Gibson overlap R
ZNF683 Taqman probe	Life Technologies	Hs00543184_m1
Beta actin Taqman probe	Life Technologies	4333762T

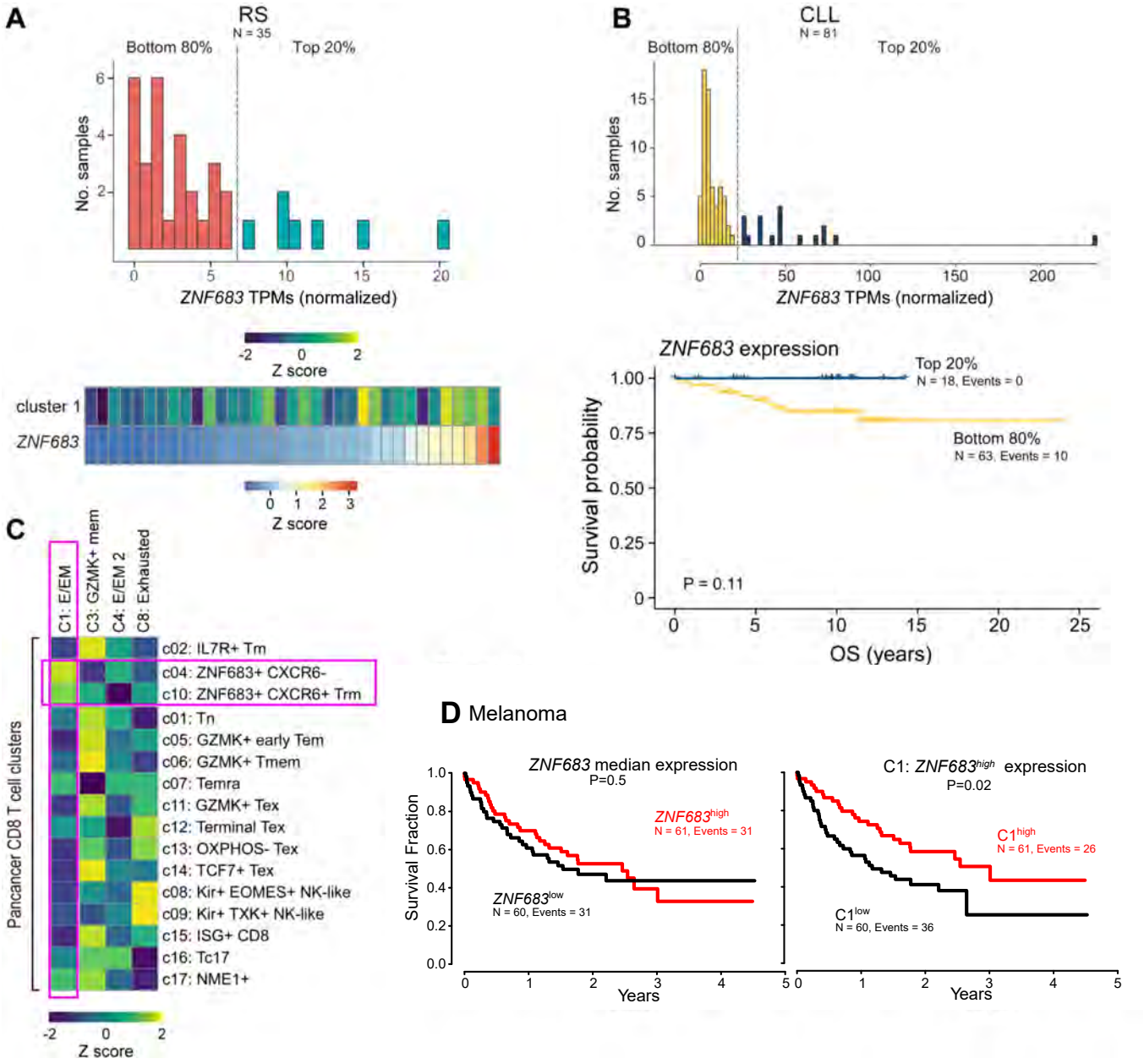
GAPDH Taqman probe	Life Technologies	4333764T
NEBNext Multiplex Oligos for Illumina, Dual Index Primers	NEB	E7600
Recombinant DNA		
ZNF683 gene block	This paper	
pSB-tet-GP	Addgene	60495
pCMV(CAT)T7-SB100	Addgene	34879
Software and algorithms		
Cell Ranger v2.0.0, v2.1.1, v3.0.2	10x Genomics	http://10xgenomics.com/
Conos v 1.3.1	Barkas et al., 2019	https://github.com/kharchenkolab/conos
DESeq2 v1.22.2	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
fgsea v1.13.2	Korotkevich et al., 2019	https://bioconductor.org/packages/release/bioc/html/fgsea.html
scRepertoire v1.0.0	Borcherding et al., 2022	https://www.bioconductor.org/packages/release/bioc/html/scRepertoire.html
slingshot v1.8.0	Street et al., 2018	https://bioconductor.org/packages/release/bioc/html/slingshot.html
kallisto v0.46.0	Bray et al., 2016	https://pachterlab.github.io/kallisto/download.html
Cutadapt v3.5	Martin et al., 2011	https://cutadapt.readthedocs.io/en/stable/
Bowtie2 v2.4.2	Langmead et al., 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
samtools v1.2	Danecek et al., 2021	http://www.htslib.org/download/
bedtools v2.30.0	Quinlan et al., 2010	https://bedtools.readthedocs.io/en/latest/index.html
SEACR v1.3	Meers et al., 2019	https://github.com/FredHutch/SEACR
deepTools v3.5.1	Ramírez et al., 2016	https://deeptools.readthedocs.io/en/develop/
DiffBind v3.4.3	Stark et al., 2011	https://bioconductor.org/packages/release/bioc/html/DiffBind.html
Cistrome-GO	Li et al., 2019	http://go.cistrome.org

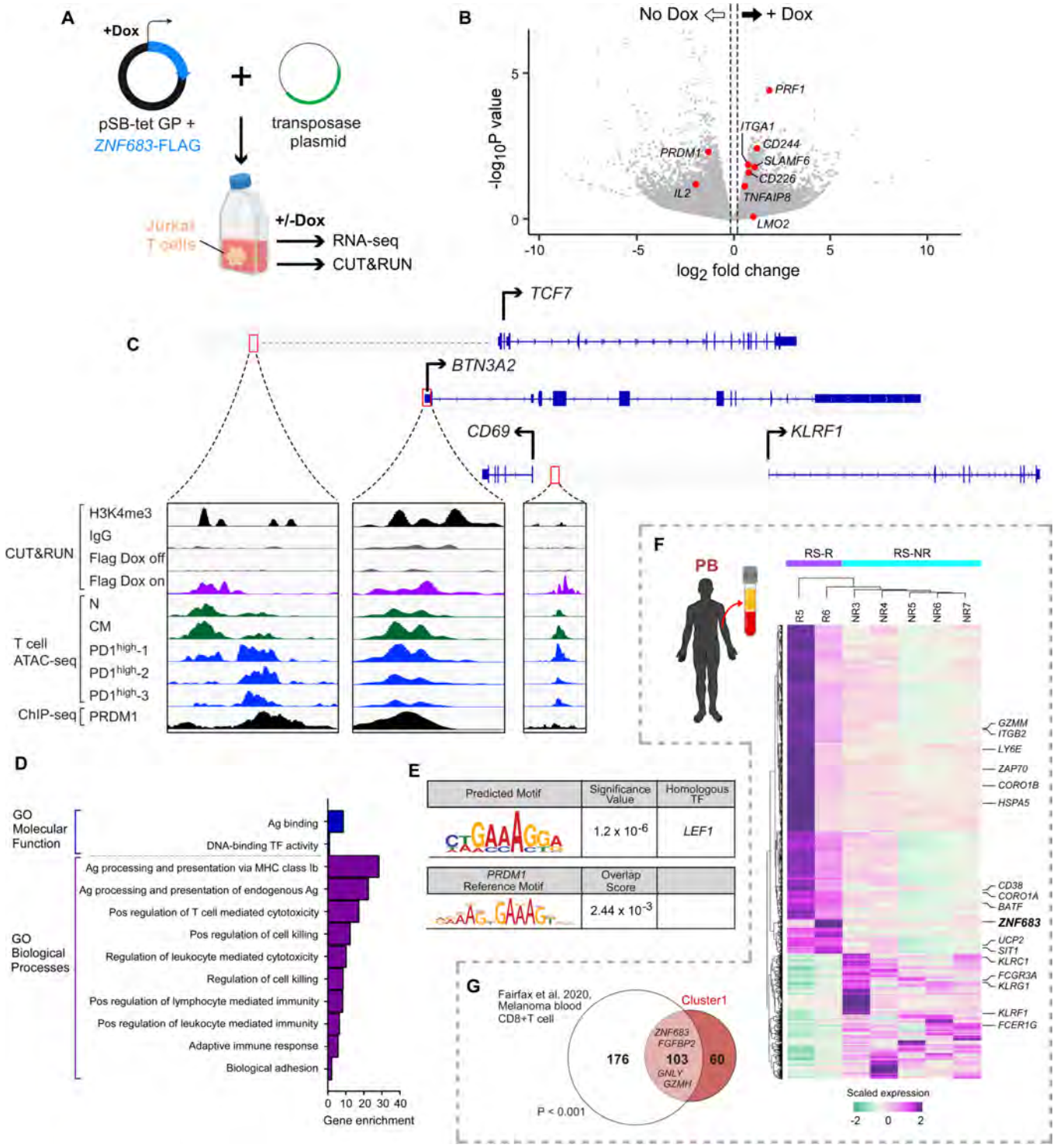
MEME-ChIP v5.4.1	Bailey et al., 2015	https://meme-suite.org/meme/tools/meme-chip
GSVA v1.30.0	Hänzelmann et al., 2013	https://bioconductor.org/packages/release/bioc/html/GSVA.html
CIBERSORT v1.05	Newman et al., 2015	https://cibersort.stanford.edu
RNA-SeQC v2.3.6	Graubert et al., 2021	https://github.com/getzlab/rnaseqc
Integrated Genomics Viewer (IGV)	Broad Institute	https://software.broadinstitute.org/software/igv/


A
NCT02420912
14 NR
10 R
t Transplant
Bone marrow sampling times
RS response
Complete
Partial
Progression
Stable
Nivolumab 3 mg/kg every 2 weeks
Ibrutinib 420 mg daily
RS
Responders
Non-responders
CLL
RS-R1
RS-R2
RS-R3
RS-R4
RS-NR1
RS-NR2
CLL-1
CLL-2
Months from treatment initiation
B
All viable bone marrow lymphocytes
Large
18.8
Small
80.3
SSC (log)
FSC (x 10^4)
RS
92.7
CD5-FITC (log)
CD19-PE-Cy7 (log)
Bone marrow immune cells
28.1
CLL
49.4
• scRNA-seq (10x genomics)
• Data processing (Pagoda2)
• Joint clustering (Conos)
C
Immune
RS / CLL
T cells
NK cells
B cells
Erythroid progenitor
CD34+ stem cells
Monocytes/DC
Malignant B cells
D
GO biological processes
Membrane depolarization
Mitotic sister chromatid segregation
Metaphase anaphase transistion of cell cycle
Pos. regulation of IL-8 prod.
Regulation of chr. segregation
Hallmark
G2M checkpoint
E2F targets
Hypoxia
Cholesterol homeostatasis
Oxidative phosphorylation
Normalized enrichment score
CLL vs RS


A
All cell types
Lymphocyte sub-clustering
78,488 cells
60,727 cells
B
C6 C10 C8 C1 C4 C3 C9 C7 C5 C2 C11
NK CD8 CD4
CD16++ CD16- Exhausted E/EM E/EM 2 GZMK+ mem. Cytotoxicity T reg Naive CD4.1 CD4.2
T cell
NK cell
Naive
T reg
Inhibitory/ Exhaustion
Costim./ Activation
Cytotoxicity
Transcription factors
Cell surface
Cytokines
CD3D CD8A CD4 FCGR3A NCR1 TCF7 SELL LEF1 CCR7 FOXP3 IL2RA CTLA4 LAG3 TIGIT PDCD1 HAVCR2 CD160 ENTPD1 CD27 CD28 TNFRSF4 ICOS TNFRSF9 CD40LG CD226 GZMA GZMB PRF1 GZMK NKG7 TBX21 EOMES TOX ZNF683 IL7R CX3CR1 CD69 CXCR4 CXCR3 KLRG1 IFNG TGFB1 XCL1 XCL2
-3 0 3
Z score
C
NK CD8 CD4
C6: CD16++ C10: CD16- C8: Exhausted C1: E/EM C4: E/EM 2 C3: GZMK+ mem. C9: cytotoxicity C7: T reg C5: Naive C2: CD4.1 C11: CD4.2
CD3 CD8a CD4 CD45RA CD45RO CD16 CD56 TCRab CD25 CD62L CD27 CD28 CD39 OX40 ICOS 2B4 DNAM1 CD38 CCR4 CD95 TIGIT PD1
-1 0 1 2 3
Z score
D
CD4
NK
CD8
Exhaustion (ES)
Cytotoxicity (ES)
c7 c3 c8 c6 c10 c1 c4 c9 c11 c2 c5
E
N=17 samples, 8 individuals
RS-R RS-NR CLL
R1 R2 R3 R4 NR1 NR2 CLL1 CLL2
Cell proportion
Months: 0 3 12 0 3 1 3 7 3 0 3 0 3 0 3 0 3
Pre Post Prog Pre Post Pre* Post Prog Post Pre Post Pre Post Pre Post Pre Post
C7: T reg
C5: Naive
C8: Exhausted
C4: CD8 E/EM 2
C1: CD8 E/EM
C2,C3,C6,C9,C10,C11
Normal marrow (N = 30)
DFCI
Oetjen et al.
HCA/Broad
F
C1 C4 C8 C6 C9 C7 C5
p=0.001 0.03 0.001 0.03 0.001 0.001 0.06
Cell proportion
RS/CLL
Normal


A
60,727 cells
R
NR
ZNF683
B
Cluster 9: CD4 with cytotoxicity
Baseline
Post-CPB
p=0.04
Z score
Status: NR R
Cluster 1: Effector/Effector Memory CD8
p=0.04
Cluster 2: CD4-1
p=0.10
Cluster 7: T reg
p=0.006
Cluster 8: Exhausted CD8
p=0.69
C
Cluster 8: Exhausted CD8
Non-responders
Responders
-log10 P value
log2 fold change
PATL2
GLUL
DZIP3
CD27
CD226
ANXA4
SATB1
AMICA1
NELL2
KLRC1
SIRPG
IL7R
ZNF683
TGFB1
D
NR
R
TOX
DDIT4
RELB
CXXC5
JUN
JUNB
FOS
ZNF3
ZNF136
ZNF672
PURA
ZNF500
RORA
ATF5
HOXB2
ZNF683
CASZ1
SOX13
CEPBA
log2 fold change
E
Response
Progression
Cluster 4
Cluster 1
RS-R3 cell proportion
Months post PD-1 therapy
F
Pre
Post
Prog
RS-R1
RS-R3
RS-NR1
RS-NR2
ZNF683
% Expressed
Average expression
G
RS-R2
RS-NR1
Pre
Post
Hyperexpanded (10-1 to 10)
Large (10-2 to 10-1)
Medium (10-3 to 10-2)
Small (10-4 to 10-3)
Rare (0 to 10-4)
RS-R2 TCR counts
RS-NR1 TCR counts
Clonotype:
Pre/Post
CD4/CD8
H
Responders
N = 4
Non Responders
N = 4
Expanded
Depleted
Novel
Disappeared
Same
Response assessment log (Frequency+1)
Baseline log (Frequency+1)
I
Responders - Progression
N = 2
Progression log (Frequency+1)
Response log (Frequency+1)
J
c4: CD8 E/EM 2
ZNF683int
c5: Naive
c1: CD8 E/EM
ZNF683high
c3: CD8 GZMK+ memory
c8: CD8 exhausted
dim[1]
dim[2]


A
RS
N = 35
Bottom 80%
Top 20%
No. samples
ZNF683 TPMs (normalized)
Z score
cluster 1
ZNF683
B
CLL
N = 81
Bottom 80%
Top 20%
No. samples
ZNF683 TPMs (normalized)
ZNF683 expression
Top 20%
N = 18, Events = 0
Bottom 80%
N = 63, Events = 10
Survival probability
P = 0.11
OS (years)
C
C1: E/EM
C3: GZMK+ mem
C4: E/EM 2
C8: Exhausted
Pancancer CD8 T cell clusters
c02: IL7R+ Tm
c04: ZNF683+ CXCR6-
c10: ZNF683+ CXCR6+ Trm
c01: Tn
c05: GZMK+ early Tem
c06: GZMK+ Tmem
c07: Temra
c11: GZMK+ Tex
c12: Terminal Tex
c13: OXPHOS- Tex
c14: TCF7+ Tex
c08: Kir+ EOMES+ NK-like
c09: Kir+ TXK+ NK-like
c15: ISG+ CD8
c16: Tc17
c17: NME1+
Z score
D Melanoma
ZNF683 median expression
P=0.5
ZNF683high
N = 61, Events = 31
ZNF683low
N = 60, Events = 31
Survival Fraction
Years
C1: ZNF683high expression
P=0.02
C1high
N = 61, Events = 26
C1low
N = 60, Events = 36
Years


A
+Dox
pSB-tet GP +
ZNF683-FLAG
transposase
plasmid
Jurkat
T cells
+/-Dox
RNA-seq
CUT&RUN
B
No Dox
+ Dox
-log10P value
log2 fold change
PRF1
ITGA1
CD244
SLAMF6
CD226
TNFAIP8
LMO2
PRDM1
IL2
C
TCF7
BTN3A2
CD69
KLRF1
CUT&RUN
H3K4me3
IgG
Flag Dox off
Flag Dox on
T cell
ATAC-seq
N
CM
PD1high-1
PD1high-2
PD1high-3
ChIP-seq
PRDM1
D
GO
Molecular
Function
Ag binding
DNA-binding TF activity
GO
Biological
Processes
Ag processing and presentation via MHC class Ib
Ag processing and presentation of endogenous Ag
Pos regulation of T cell mediated cytotoxicity
Pos regulation of cell killing
Regulation of leukocyte mediated cytotoxicity
Regulation of cell killing
Pos regulation of lymphocyte mediated immunity
Pos regulation of leukocyte mediated immunity
Adaptive immune response
Biological adhesion
0 10 20 30 40
Gene enrichment
E
Predicted Motif
Significance Value
Homologous TF
1.2 x 10-6
LEF1
PRDM1 Reference Motif
Overlap Score
2.44 x 10-3
F
PB
RS-R
RS-NR
R5
R6
NR3
NR4
NR5
NR6
NR7
GZMM
ITGB2
LY6E
ZAP70
CORO1B
HSPA5
CD38
CORO1A
BATF
ZNF683
UCP2
SIT1
KLRC1
FCGR3A
KLRG1
KLRF1
FCER1G
Scaled expression
-2 0 2
G
Fairfax et al. 2020.
Melanoma blood
CD8+T cell
Cluster1
176
103
60
ZNF683
FGFBP2
GNLY
GZMH
P < 0.001

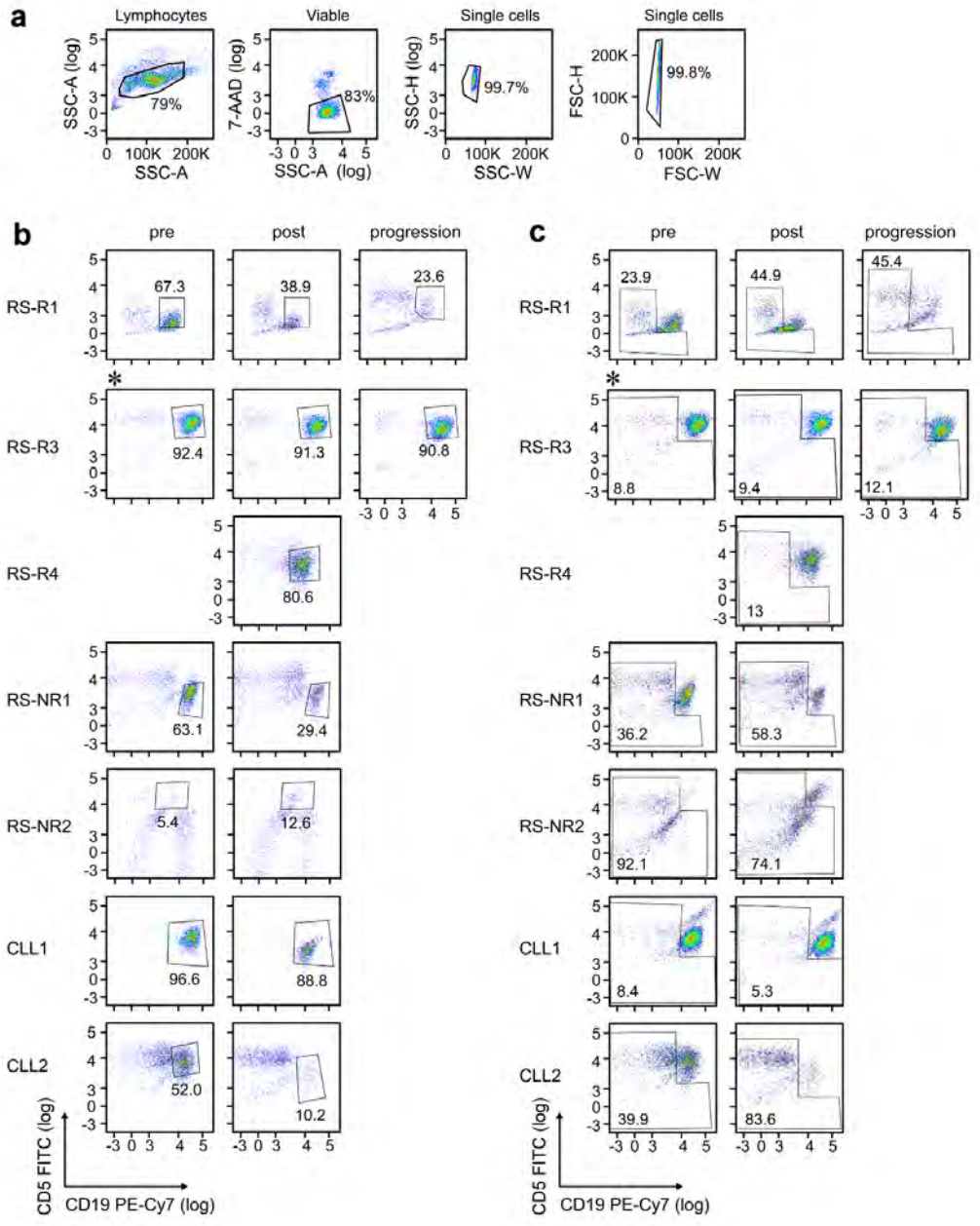


Figure S1. Sorting strategy for FACS of RS and CLL.

(A) Experimental schema showing flow cytometry sorting strategy for bone marrows b, Sorting strategy for CLL malignant B cell fraction of patient bone marrow samples.

(B) Sorting strategy for immune (non-malignant) cell fraction for RS and CLL cells.

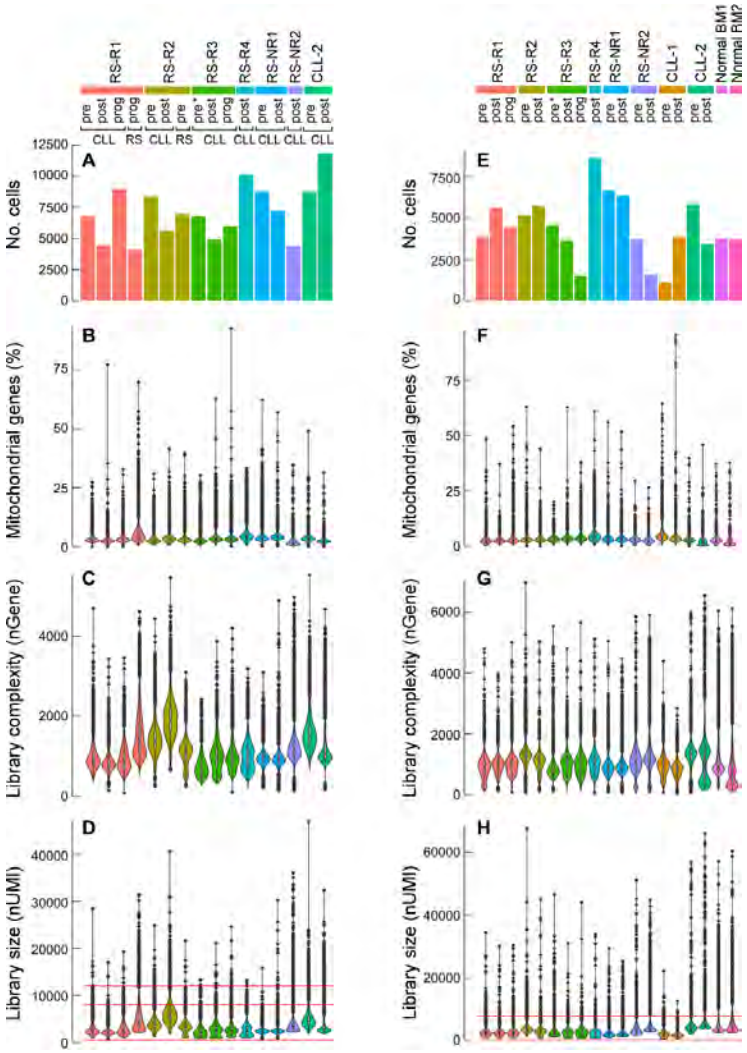


Figure S2. Single-cell RNA-sequencing metrics.

(A) – (D) Malignant B cell QC metrics per sample before filtering:

(A) cells/sample

(B) percent mitochondrial content/cell per sample

(C) gene/cell per sample

(D) UMI/cell (3rd row)

(E) – (H) Immune fraction QC metrics per sample:

(E) cells/sample

(F) percent mitochondrial content per cell per sample

(G) gene/cell per sample

(H) UMI/cell per sample

Table S2. QC metrics on non-tumor, tumor scRNA-seq post filtering, Related to Figure S2.

Subject ID	Timepoint	No. cells	mean numi/cell	mean ngene/cell	% mitochondrial reads
RS-R1	pre	3157	2684.72	957.82	2.83
RS-R2	post	5424	2601.85	966.26	2.72
RS-R1	progression	4178	2754.55	979.55	2.95
RS-R2	pre	4738	3913.49	1303.82	3.08
RS-R2	post	5230	3126.20	1077.79	3.24
RS-R3	pre	4424	2810.27	846.81	3.26
RS-R3	post	3492	2740.32	992.75	3.77
RS-R3	progression	1378	2933.43	996.26	3.82
RS-R4	post	8120	2571.84	1042.79	4.16
RS-NR1	pre	6137	2396.28	919.96	3.31
RS-NR1	post	5977	2520.34	951.49	3.31
RS-NR2	pre	3211	3622.74	1130.42	3.04
RS-NR2	post	1129	4146.55	1202.74	2.59
CLL-1	pre	907	2572.04	1005.77	4.70
CLL-1	post	3690	2055.69	851.58	3.63
CLL-2	pre	4647	4408.19	1353.41	2.92
CLL-2	post	2921	5111.08	917.75	1.65
Normals, Donor 1	NA	2835	4010.31	961.46	2.66
Normals, Donor 2	NA	2984	4417.59	606.30	1.57

Subject ID	Timepoint-tumor	No. cells	mean numi/cell	mean ngene/cell	% mitochondrial reads
RS-R1	pre-CLL	6712	2630.43	972.96	3.09
RS-R1	post-CLL	4423	2218.69	846.27	2.66
RS-R1	progression-CLL	8892	2818.00	936.88	3.10
RS-R1	progression-RT	3397	4794.60	1473.52	4.64
RS-R2	pre-CLL	8161	3972.56	1424.41	2.95
RS-R2	pre-RT	6566	5885.76	1911.27	3.34
RS-R2	post-CLL	5285	3337.94	1122.98	3.35
RS-R2	pre-CLL	3809	3725.41	1405.29	3.64
RS-R2	pre-RT	2741	6288.29	2082.81	3.92
RS-R3	pre-CLL	6670	2274.80	747.70	2.68
RS-R3	post-CLL	4462	2856.17	1024.06	3.73
RS-R3	progression-CLL	5801	2779.27	988.91	3.39
RS-R4	post-CLL	9364	2703.59	1016.53	4.38
RS-NR1	pre-CLL	8447	2396.13	951.43	3.77
RS-NR1	post-CLL	7055	2485.57	957.18	4.24
RS-NR2	post-CLL	3985	3609.75	1238.10	1.91
CLL-2	pre-CLL	8533	4933.32	1554.01	3.55
CLL-2	post-CLL	11686	3420.43	1209.31	2.36

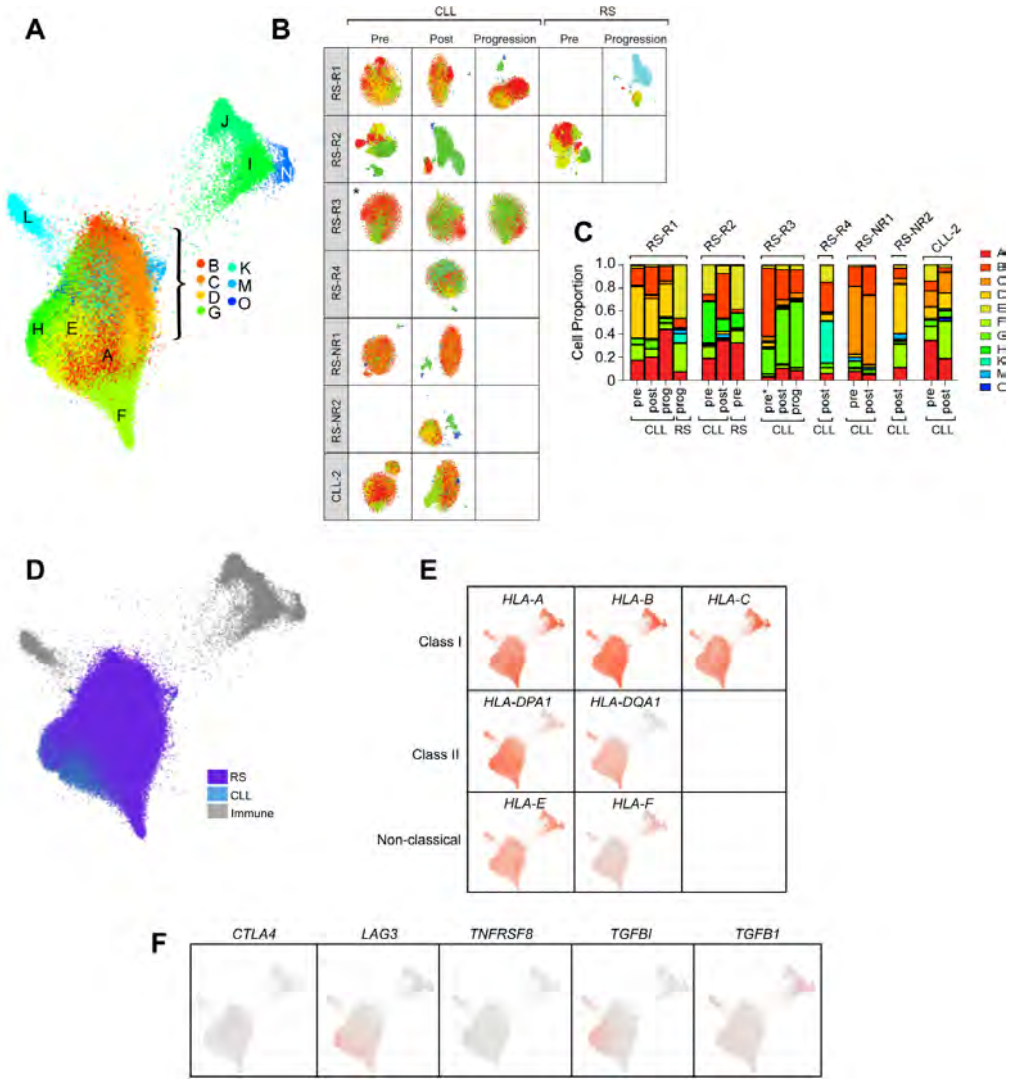


Figure S3. Joint clustering of tumor and immune cells, related to Figure 1.

(A) Conos joint graph using largeVis embedding of sorted tumor fraction clusters A-O.

(B) Individual graph embeddings of each sample showing clusters A-O.

(C) Malignant B cell cluster cell proportions across sorted tumor fractions.

(D) Conos joint graph embedding showing RS (purple), CLL (blue), and immune (gray) populations.

(E) Conos joint graph embeddings showing expression of Class I and representative Class II and non-classical HLA genes in CLL and RS. Cells expressing, red.

(F) Conos joint graph embeddings showing immune marker genes expressed by malignant B cells. Cells expressing, red.

Table S1. Clinical data for study patients, Related to Figure 1.

a. Discovery-Bone marrow										
Subject ID	**Gender**	**Prior CLL therapies**	**Cytogenetics**	**Gene Mutation (Clinical Testing)**	***IGHV* status**	**Age at RS diagnosis (years)**	**Baseline marrow findings**	**Prior RS therapy**	**Best response (months after PD-1 blockade)**	**Months to progression after PD-1 blockade**
RS-R1	F	R, FCR	del(17p) CK	*TP53*	UM	62	CLL (80%)	NA	CR (1)	12
RS-R2	F	FCR, BR, OBI-HDMP, R-Idela	Tri12 CK	*TP53*	UM	66	RS and CLL	NA	CR (1)	NA
RS-R3	M	FR, Bendamustine, IBR	Tri12	*BTK*, *NOTCH1*	UM	53	CLL (95% at C2D1)	NA	PR (3)	6
RS-R4	M	FCR	del(11q) CK	*SF3B1*, *ATM*	UM	66	CLL (70% CLL at C2D1)	NA	CR (1)	6
RS-NR1	F	IBR	del(17p) CK	*BIRC3*, *NOTCH1*	UM	64	CLL	NA	NA	1
RS-NR2	M	none	del(11q) CK	not done	UM	60	CLL	BR, R-EPOCH and RICE	NA	3
CLL-1	F	ofatumumab	del(13q)	*TP53*	MUT	NA	CLL	NA	NA	NA
CLL-2	M	VEN, FCR	del(17p)	*TP53*	UM	NA	CLL	NA	NA	NA

b. Validation-Blood										
RS-R5	M	none	not done	*MYD88*, *CXCR4*, *CD79B*	not done	66	NA	RCHOP, auto-SCT	CR	NA
RS-R6	M	PCR, BR, ibrutinib	NEG FISH	not done	not done	70	NA	none	PR (1)	NA
RS-NR3	F	lenalidomide	Tri12	*BIRC3*	MUT	78	NA	none	no response	NA

RS-NR4	F	FR, FCR, BR, ibrutinib	del(17p) CK	*TP53*	MUT	63	NA	none	no response	NA
RS-NR5	F	BR, chlorambucil, rituximab/HDMP, ibrutinib, ofatumumab	del(17p) CK	not done	UM	88	NA	none	no response	NA
RS-NR6	F	FCR, TGR1202/ublituximab	Tri12	not done	MUT	65	NA	R-EPOCH, haplo-SCT, OFAR, VEN+obin	no response	NA
RS-NR7	M	FMD, BR	not done	not done	not done	49	NA	CD19 mAb, RICR, R-DHAP, R-Hyper-CVAD, MUD-SCT, ibrutinib	no response	NA

Abbreviations: M, male; F, female; FCR = fludarabine, cyclophosphamide, rituximab; R, rituximab; BR, bendamustine, rituximab; OBI-HDMP, obinutuzumab and high dose methylprednisone, R-idela, Rituximab-idealisib; IBR, ibrutinib; VEN, venetoclax; FR, fludarabine and rituximab; FMD, Fludarabine, mitoxantrone and dexamethasone; CK, complex karyotype; UM, IGHV unmutated; M, IGHV mutated; R-EPOCH, rituximab plus etoposide, prednisone, vincristine, doxorubicin and cyclophosphamide; R-CHOP, rituximab, cyclophosphamide, doxorubicin, vincristine and prednisone; RICE, rituximab, ifosfamide, carboplatinum, etoposide; SCT, stem cell transplant; OFAR, oxaliplatin, fludarabine, cytarabine and rituximab; MUD, matched unrelated donor; DHAP, dexamethasone, cytarabine, cisplatin; CR; complete response, PR, partial response

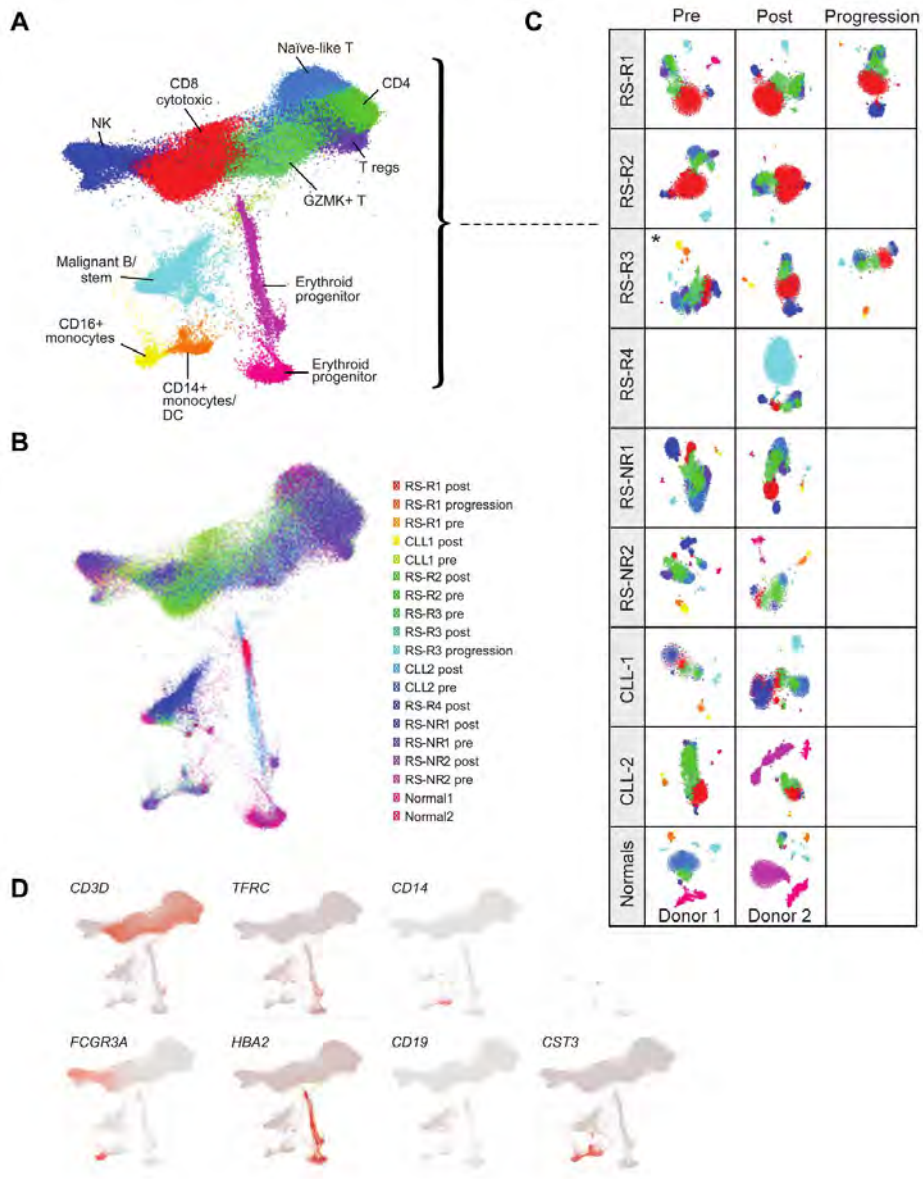


Figure S4. Clustering of immune cells, related to Figure 2.

(A) Conos joint graph embedding of immune clusters.

(B) Joint graph of immune clusters by sample.

(C) Joint graph colored by sample.

(D) Marker gene panels highlight major cell types within immune clusters. Red highlights cells expressing labelled marker gene.

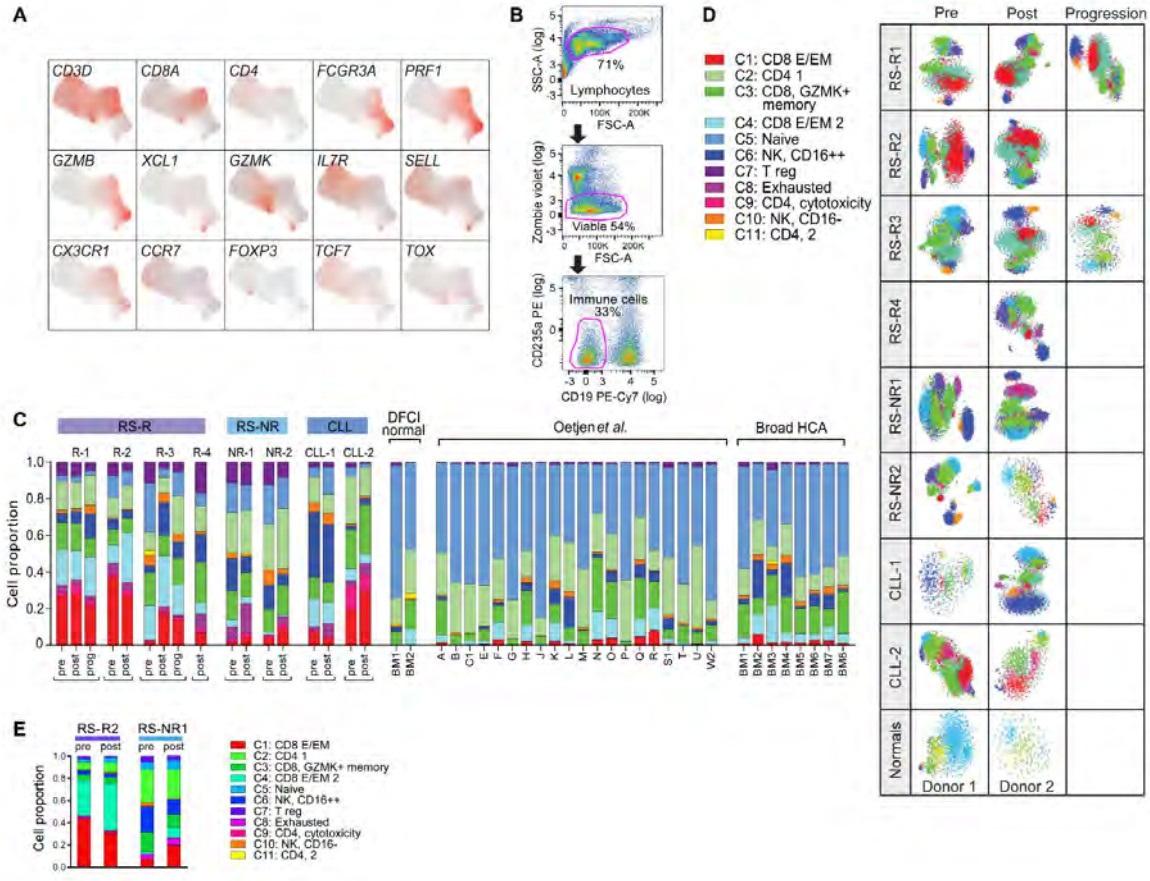


Figure S5. Lymphocyte sub-clustering, related to Figure 2.

(A) Marker gene panels highlight major cell types within lymphocyte sub-clustering on Conos largeVis embedding. Red highlights cells expressing marker gene.

(B) Flow sorting strategy for CITE-seq experiment. FSC-A, forward scatter.

(C) Bar graphs showing population distributions across serial CLL and RS marrow samples and normal bone marrow samples.

(D) Individual panel graphs of patient marrow samples and normal marrow colored by cluster.

(E) Bar graphs showing population distributions for 5' single-cell RNA-seq performed in conjunction with CITE-seq for marrow samples of RS-R2 (purple) and RS-NR1 (blue) pre- and post- treatment.

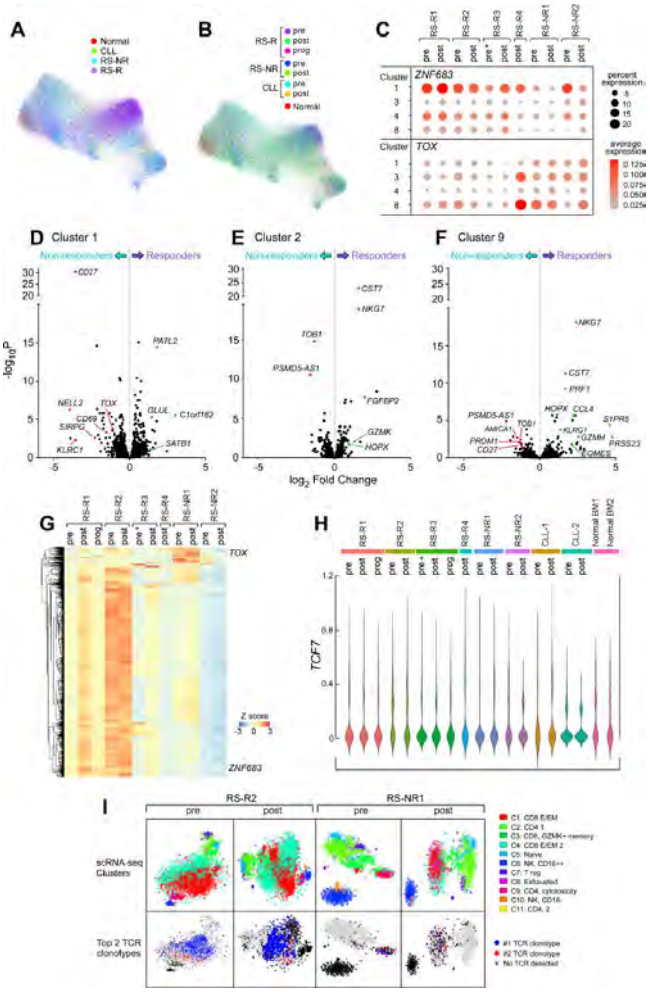


Figure S6. T cell populations associated with CPB response, related to Figure 3.

(A) Conos joint graph of immune sub-cluster by sample.

(B) LargeVis of Conos joint graph of immune sub-cluster by sample and sample type.

(C) Bubble plot displaying ZNF683 and TOX expression in RS-R vs RS-NR.

(D) – (F) Volcano plots for CD8 T cell clusters (1, 2, 9) displaying gene expression differences within each cluster between non-responder (left) and responder (right).

(G) Heatmap displaying genes differentially regulated between RS-R and RS-NR in across all CD8 clusters. *ZNF683* and *TOX* are labelled among top differentially expressed genes**.**

(H) *TCF7* expression by violin plot is stable across all samples in memory cluster 3.

(I) Panel of cluster distribution for individual samples in 5' scRNA-seq performed with CITE-seq and TCR-seq (top). Top 2 expanded TCR clones in each sample are depicted on the bottom panel (bottom).

Table S7a. ssGSEA enrichment scores for cluster 1 signature on bulk RNA seq from RS patients, Related to Figure 4.

	cluster 1 signature
RP.1917_DFCI.5486.RS.01_v1_Exome_OnPrem	6.27
RP.1917_DFCI.5488.RS.01_v1_Exome_OnPrem	6.14
RP.1917_DFCI.5491.RS.01_v1_Exome_OnPrem	6.75
RP.1917_DFCI.5494.RS.01_v1_Exome_OnPrem	6.36
RP.1917_DFCI.5495.RS.01_v1_Exome_OnPrem	6.18
RP.1917_DFCI.5496.RS.01_v1_Exome_OnPrem	6.67
RP.1917_DFCI.5497.RS.01_v1_Exome_OnPrem	5.98
RP.1917_DFCI.5498.RS.01_v1_Exome_OnPrem	6.32
RP.1917_DFCI.5500.RS.01_v1_Exome_OnPrem	6.02
RP.1917_DFCI.5502.RS.01_v2_Exome_OnPrem	6.72
RP.1917_DFCI.5503.RS.01_v1_Exome_OnPrem	6.38
RP.1917_DFCI.5504.RS.01_v1_Exome_OnPrem	6.48
RP.1917_DFCI.5505.RS.01_v1_Exome_OnPrem	6.90
RP.1917_DFCI.5506.RS.01_v1_Exome_OnPrem	6.24
RP.1917_DFCI.5507.RS.01_v1_Exome_OnPrem	6.51
RP.1917_DFCI.5508.RS.01_v1_Exome_OnPrem	6.29
RP.1917_DFCI.5509.RS.01_v1_Exome_OnPrem	6.41
RP.1917_DFCI.5510.RS.01_v1_Exome_OnPrem	6.39
RP.1917_DFCI.5511.RS.01_v1_Exome_OnPrem	6.41
RP.1917_DFCI.5514.RS.01_v1_Exome_OnPrem	6.48
RP.1917_DFCI.5515.RS.01_v1_Exome_OnPrem	6.50
RP.1917_DFCI.5516.RS.01_v1_Exome_OnPrem	6.32
RP.1917_DFCI.5517.RS.01_v1_Exome_OnPrem	6.53
RP.1917_DFCI.5522.RS.01_v1_Exome_OnPrem	6.65
RP.1917_DFCI.5522.RS.02_v1_Exome_OnPrem	6.87
RP.1917_DFCI.5523.RS.01_v1_Exome_OnPrem	6.68
RP.1917_DFCI.5524.RS.01_v1_Exome_OnPrem	6.16
RP.1917_DFCI.5525.RS.01_v1_Exome_OnPrem	6.23
RP.1917_DFCI.5526.RS.01_v1_Exome_OnPrem	6.40
RP.1917_DFCI.5527.RS.01_v1_Exome_OnPrem	6.25
RP.1917_DFCI.5528.RS.01_v1_Exome_OnPrem	6.46
RP.1917_DFCI.5529.RS.01_v1_Exome_OnPrem	5.90
RP.1917_DFCI.5530.RS.01_v1_Exome_OnPrem	6.78
RP.1917_DFCI.5531.RS.01_v1_Exome_OnPrem	6.12
RP.1917_DFCI.5532.RS.01_v1_Exome_OnPrem	6.72

Table S7b. ssGSEA enrichment scores for Pan-cancer CD8 T cell clusters on scRNA-seq CD8 clusters, Related to Figure 4.

	cluster 1	**cluster 3**	**cluster 4**	**cluster 8**
CD8.c02	0.35	0.49	0.42	0.36
CD8.c04	0.73	0.69	0.72	0.70
CD8.c10	0.76	0.76	0.75	0.76
CD8.c01	0.73	0.77	0.75	0.70
CD8.c05	0.45	0.57	0.49	0.52
CD8.c06	0.71	0.74	0.72	0.70
CD8.c07	0.68	0.55	0.68	0.68
CD8.c11	0.41	0.54	0.43	0.50
CD8.c12	0.64	0.64	0.62	0.65
CD8.c13	-0.08	0.07	-0.05	0.09
CD8.c14	0.46	0.53	0.49	0.48
CD8.c08	0.46	0.50	0.48	0.56
CD8.c09	0.37	0.37	0.41	0.49
CD8.c15	0.43	0.44	0.43	0.43
CD8.c16	0.35	0.40	0.40	0.29
CD8.c17	0.73	0.74	0.73	0.73

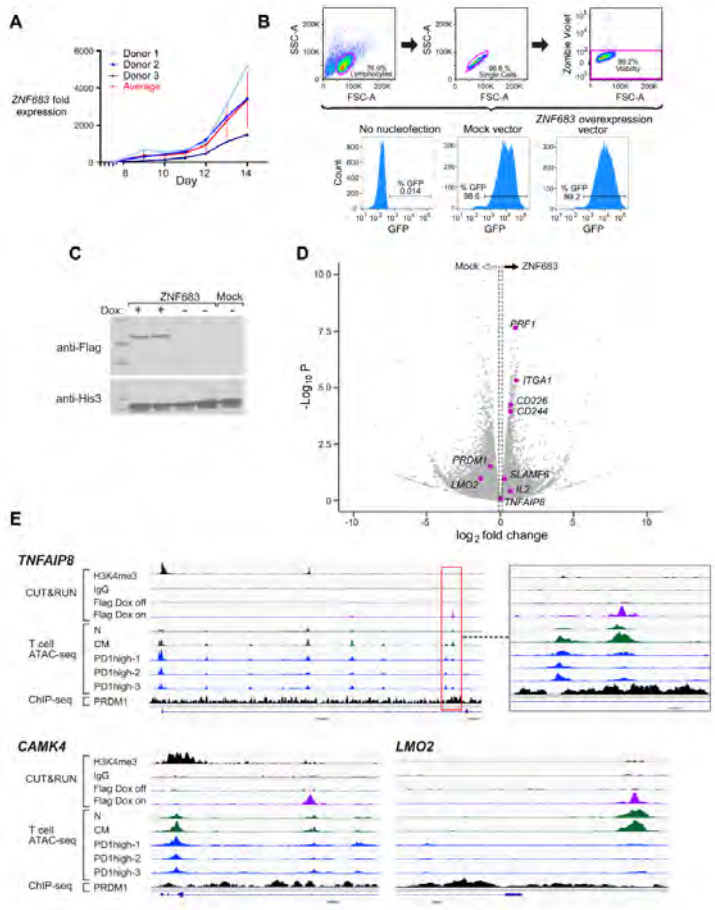


Figure S7. ZNF683 expression in primary T cells and overexpression in Jurkat cell lines, related to Figure 5.

(A) Line graph shows endogenous ZNF683 in primary T cells from 3 healthy donors expanded in culture following CD3-CD28 bead stimulation.

(B) Flow cytometry analysis of ZNF683-expressing Jurkats after puromycin selection shows stable cell lines express GFP.

(C) Western blot confirms dox-induced ZNF683 expression in Jurkat cells, as measured by FLAG protein expression.

(D) Volcano plot for gene expression differences between Jurkat cells containing dox-inducible ZNF683 vector vs mock (luciferase) vector by RNA-seq.

(E) CUT&RUN on Jurkat cell lines shows binding of ZNF683 at regions surrounding key immune genes (purple) that correspond to differential ATAC-seq peaks in T cell subsets (green)[24] and prior PRDM1 ChIP-seq data (black)[21]. CUT&RUN controls depicted in black in top three tracks of each panel.

Table S9. Antibodies used for CITE-seq, Related to STAR methods.

TotalSeq-C antibody	Biolegend catalog number
CD80	301071
CD86	305447
PDL1	328751
CD3	300479
CD19	302265
CD45RA	304163
CD4	300567
CD8a	301071
CD14	301859
CD16	302065
CD56	392425
CD25	302649
CD45RO	404259
PD1	329963
TIGIT	372729
IgG1isotype	400187
IgG2aisotype	400293
IgG2bisotype	400381
CD20	302363
NKp46	331941
CD69	310951
CD62L	304851
CCR7	353251
CD27	302853
HLADR	307663
CD11b	301359
ICOS	313553
41BB	309839
CD28	302963
IL7RA	351356
CD45	304068
CD15	323053
CD73	344031
CD70	355119
CD44	338827
TCRab	306743
KLRG1	138433
CD39	328237
NKG2D	320837

CD5	300637
CD10	312233
CTLA4	369621
CD95	305651
OX40	350035
CXCR3	353747
CCR4	359425
CXCR4	306533
2B4	329529
CD40	334348
DNAM1	338337
CD49f	313635
CD38	303543
B7H4	358116
CD11a	350617

PAPER III

Preneoplastic Alterations Define CLL DNA Methylome and Persist through Disease Progression and Therapy

Kretzmer, H., Biran, A.*, Purroy, N. Z.*, **Lemvigh, C. K.***, Clement, K*., Gruber, M.*, Gu, H., Rassenti, L., Mohammad, A. W., Lesnick, C., Slager, S. L., Braggio, E., Shanafelt, T. D., Kay, N. E., Fernandes, S. M., Brown, J. R., Wang, L., Li, S., Livak, K. J., Neuberg, D. S., ... Meissner, A.

Published, Blood cancer discovery, 2021

The work carried out in relation to this thesis included an extensive preliminary analysis of the scRNA sequencing data presented here. The work comprised investigation of data integration, initial figure generation and interpretation of results.

* = equal contribution

RESEARCH ARTICLE

Preneoplastic Alterations Define CLL DNA Methylome and Persist through Disease Progression and Therapy

Helene Kretzmer[1], Anat Biran[2], Noelia Purroy[2,3,4], Camilla K. Lemvigh[2,5], Kendell Clement[3,6], Michaela Gruber[2,7], Hongcang Gu[3], Laura Rassenti[8], Arman W. Mohammad[3], Connie Lesnick[9], Susan L. Slager[9], Esteban Braggio[10], Tait D. Shanafelt[9], Neil E. Kay[9], Stacey M. Fernandes[2], Jennifer R. Brown[2,4,11], Lili Wang[12], Shuqiang Li[13], Kenneth J. Livak[13], Donna S. Neuberg[14], Sven Klages[15], Bernd Timmermann[15], Thomas J. Kipps[8], Elias Campo[16], Andreas Gnirke[3], Catherine J. Wu[2,3,4,11], and Alexander Meissner[1,3,6]

ABSTRACT Most human cancers converge to a deregulated methylome with reduced global levels and elevated methylation at select CpG islands. To investigate the emergence and dynamics of the cancer methylome, we characterized genome-wide DNA methylation in preneoplastic monoclonal B-cell lymphocytosis (MBL) and chronic lymphocytic leukemia (CLL), including serial samples collected across disease course. We detected the aberrant tumor-associated methylation landscape at CLL diagnosis and found no significant differentially methylated regions in the high-count MBL-to-CLL transition. Patient methylomes showed remarkable stability with natural disease and posttherapy progression. Single CLL cells were consistently aberrantly methylated, indicating a homogeneous transition to the altered epigenetic state and a distinct expression profile together with MBL cells compared with normal B cells. Our longitudinal analysis reveals the cancer methylome to emerge early, which may provide a platform for subsequent genetically driven growth dynamics, and, together with its persistent presence, suggests a central role in disease onset.

SIGNIFICANCE: DNA methylation data from a large cohort of patients with MBL and CLL show that epigenetic transformation emerges early and persists throughout disease stages with limited subsequent changes. Our results indicate an early role for this aberrant landscape in the normal-to-preneoplastic transition that may reflect a pan-cancer mechanism.

See related commentary by Rossi, p. 6.

INTRODUCTION

In normal adult tissues, cell identity is associated with accurate maintenance of a distinct DNA methylation landscape (1, 2). By contrast, cells profiled from virtually every human cancer type display local hypermethylation at typically lowly methylated CpG-rich regions and simultaneously global hypomethylation at highly methylated domains (3–6).

The striking universality of this phenomenon across cancer types raises the fundamental question of whether a cell first becomes cancerous and then acquires an aberrant methylome or if the aberrant methylome is a prerequisite. Methylation dynamics of similar proportions have otherwise only been observed during early embryonic development or the germline specification. At the same time, the generation and propagation of most other benign adult cell types show relatively stable global methylation patterns (1–7). One notable exception to the epigenetic stability of adult cell types is the maturation of B cells from hematopoietic stem cells through several intermediate stages to mature B cells, which is a critical process for the establishment of a highly effective, dynamic immune system (8). This maturation process involves genetic modulation such as somatic hypermutation of the immunoglobulin heavy-chain variable (*IGHV*) region and immunoglobulin class switch recombination (9), as well as a modulation of the methylome (10, 11). Interestingly, the methylation dynamics observed in B-cell maturation share many features with the cancer methylome (10, 11).

Chronic lymphocytic leukemia (CLL) is a malignancy of aberrant clonal mature B cells in the blood, bone marrow, and lymphoid organs that provides an ideal model setting to gain insight into the emergence of the altered methylome. Its typically indolent course enables longitudinal studies within individual patients from a pretreatment "watch and wait" phase—the duration of which is highly variable among patients, lasting months to years (12)—to the posttreatment setting and even onto progression (13, 14). A precursor stage termed monoclonal B-cell lymphocytosis (MBL) has also been described, defined as elevated white blood cell (WBC) counts with clonal B cells of a CLL immunophenotype. High-count MBL on average progresses to CLL that requires treatment in

[1]Department of Genome Regulation, Max Planck Institute for Molecular Genetics, Berlin, Germany. [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts. [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts. [4]Harvard Medical School, Boston, Massachusetts. [5]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark. [6]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts. [7]Division of Haematology and Haemostaseology, Department of Internal Medicine I, Medical University of Vienna, Vienna, Austria. [8]Division of Hematology-Oncology, Department of Medicine, Moores Cancer Center, University of California, San Diego, La Jolla, California. [9]Mayo Clinic, Division of Hematology, Rochester, Minnesota. [10]Mayo Clinic Arizona, Scottsdale, Arizona. [11]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts. [12]Department of Systems Biology, Beckman Research Institute, City of Hope, Monrovia, California. [13]Translational Immunogenomics Laboratory, Dana-Farber Cancer Institute, Boston, Massachusetts. [14]Data Science, Dana-Farber Cancer Institute, Boston, Massachusetts. [15]Sequencing Core Facility, Max Planck Institute for Molecular Genetics, Berlin, Germany. [16]Lymphoid Neoplasm Program, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hematopathology Section, Hospital Clínic; Departament d'Anatomia Patològica, Universitat de Barcelona, Barcelona, Spain.

Note: Supplementary data for this article are available at Blood Cancer Discovery Online (https://bloodcancerdiscov.aacrjournals.org/).

A. Biran, N. Purroy, C. K. Lemvigh, K. Clement, and M. Gruber contributed equally to this article.

C.J. Wu and A. Meissner jointly supervised this work.

Current address for M. Gruber: CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria; and current address for T.D. Shanafelt, Stanford University School of Medicine, Stanford, California.

Corresponding Authors: Alexander Meissner, Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany. Phone: 49-30-8413-1880; E-mail: meissner@molgen.mpg.de; and Catherine J. Wu, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115. Phone: 617-632-5943; E-mail: cwu@partners.org

Blood Cancer Discov 2021;2:54–69

doi: 10.1158/2643-3230.BCD-19-0058

1% to 2% of patients per year (15). A well-established prognostic factor in CLL is the mutational status of the *IGHV* region genes, with mutated *IGHV* showing a much better prognosis than CLL with unmutated *IGHV* (16, 17). The *IGHV* mutational status has been thought to reflect differences in the cell of origin, with a similarity in methylation profiles of unmutated CLLs and pregerminal center B cells, and of mutated CLL with mature, postgerminal center memory B cells, suggesting that CLL emerges from a spectrum of B cells undergoing broad DNA methylation alterations (11, 16, 18, 19). In addition to these characteristic global changes, we previously identified a pervasive local disorder of methylation across genomic features in CLL, not present in normal tissues (20). Although general changes in methylation profiles during B-cell development and cancer have been described (6, 10, 11, 20–24), little is currently known about: (i) if and which additional methylation changes are necessary to transition from normal into a preneoplastic state and further into cancer, (ii) how this altered cancer methylome is affected by therapy, and (iii) why it is found so ubiquitously across different types and stages of cancer. Furthermore, the chronologic origin of altered methylation with respect to cancer initiation and progression is not well understood but would be of relevance for early detection and could lead to novel therapeutic strategies.

To approach these questions, we used bulk and single-cell reduced representation bisulfite sequencing (RRBS; refs. 25–27) to profile normal mature B cells, as well as cells from patients in the preneoplastic MBL phase and during CLL progression, including after treatment. We characterized the methylation status of samples collected from 53 patients supplemented with WBC counts as a measure of tumor burden, and hence the effect of treatment (average sampling period of 5.7 years). Further, we used single-cell transcriptomics to complement the DNA methylation results in the patients transitioning from MBL to CLL. Our analyses reveal that changes in methylome and transcriptome are established early on, already at the precursor stage, and remain remarkably stable throughout the disease and even after therapy.

RESULTS

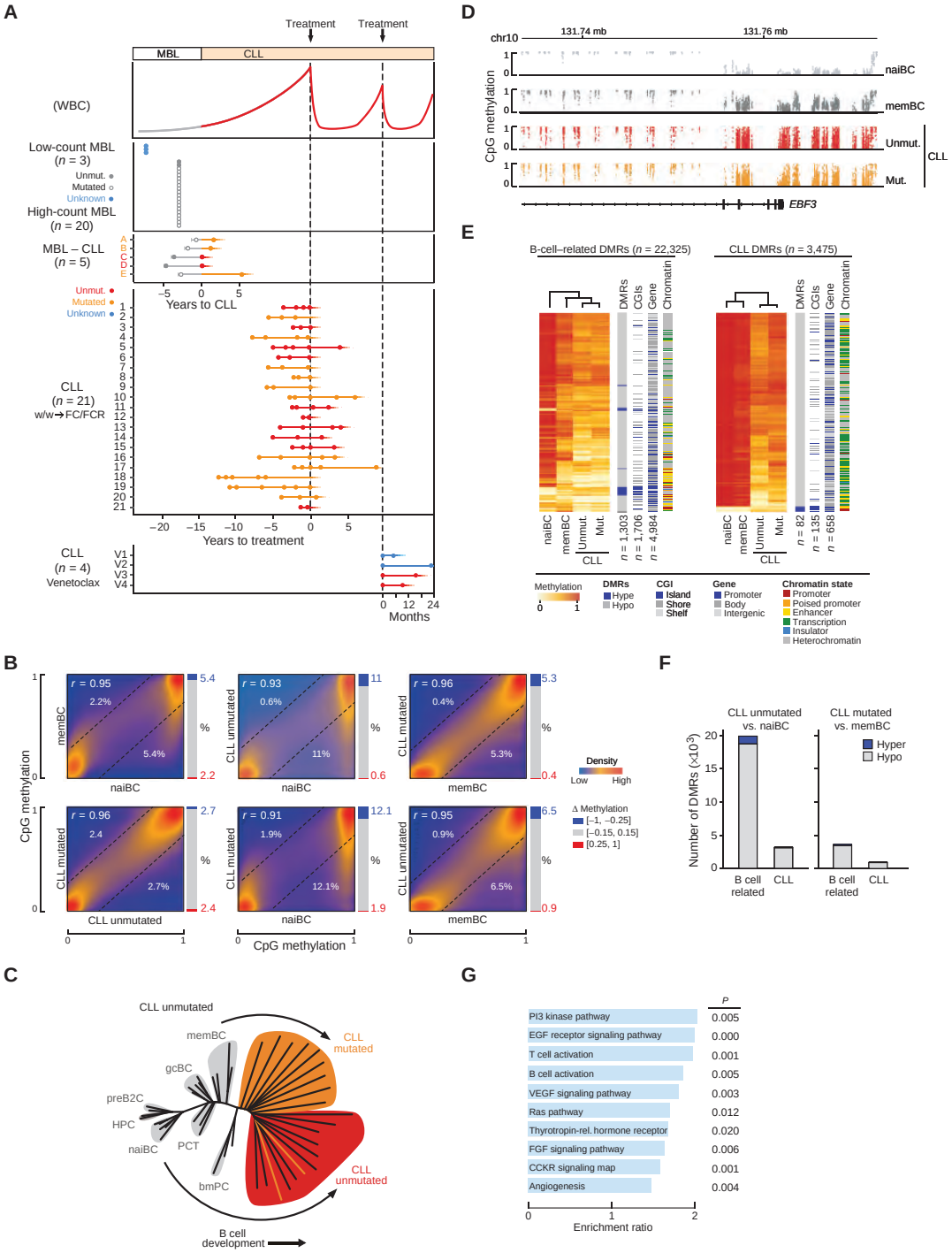
Unmutated and Mutated CLLs Converge to a Similar Methylome

To systematically study the DNA methylation dynamics across the disease course of CLL, we generated RRBS datasets from $CD19^+$ $CD5^+$ cells collected from 23 individuals with MBL, matched samples for 5 patients capturing both the MBL and their transition to CLL, and serial pre- and posttreatment samples from 25 patients collected following the diagnosis of CLL (28, 29) and compared these with published B-cell–lineage subpopulations (refs. 10, 30; Fig. 1A; Supplementary Table S1).

Genome-wide correlation of single CpG methylation showed a substantial similarity of unmutated and mutated methylation profiles (r = 0.96); however, compared with their putative cell of origin, the CLL *IGHV* subtypes showed different degrees of abnormality. Although the unmutated CLL showed more changes compared with naïve B cells, the mutated CLL exhibited a methylation landscape more similar to memory B cells than naïve to memory B cells (Fig. 1B). As noted above, CLLs originate from a range of developmental stages with pregerminal center B cells thought to give rise to unmutated CLL and mature, postgerminal center memory B cells to mutated CLLs (10, 11, 16, 18, 19, 31). Evaluation of single samples in a phylogenetic tree analysis revealed that the unmutated and mutated CLL samples are characterized by a methylome that consistently differs from normal naïve and memory B cells, suggesting a convergent disease-associated methylome, irrespective of *IGHV* mutation status (Fig. 1C; Supplementary Fig. S1A). Together, these results suggest that both *IGHV* subtypes of CLL undergo methylation changes specific to CLL. However, some of these changes also appear to be normally acquired during B-cell maturation, as observed in the example of the *EBF3* locus (Fig. 1D).

To more systematically evaluate regions that are consistently altered in CLL, we identified differentially methylated regions (DMR) between (i) unmutated CLL versus naïve B cells (n = 23,206 DMRs) and (ii) mutated CLL versus memory B cells

Figure 1. CLL methylation signatures distinguish CLL from normal B cells. **A,** Schematic representation of progression from the precursor state of MBL to CLL, depicting the extended period of "watch and wait" (w/w) until first treatment and an overview of the patient cohort from which 109 samples were collected to generate RRBS data. Combination chemoimmunotherapy (CIT; FC/FCR) typically leads to a rapid decrement in WBC counts. Our cohort included samples from 23 MBL, 5 paired samples of MBL and CLL, and 25 CLL patients. More specifically, the MBL samples are n = 20 high and n = 3 low count [2 with unmutated *IGHV* (≥98% homology with germline sequence), 18 with mutated *IGHV* (<98% homology with germline sequence), 3 with unknown status], and the CLL are n = 21 CIT treated (red: *IGHV* unmutated status; orange: *IGHV* mutated status) and n = 4 venetoclax treated, after having already progressed from first-line (fludarabine-based) regimens. Each circle indicates a sample collected. **B,** Correlation of CpG methylation levels between naïve B cell, memory B cells, and unmutated and mutated CLL at CpG-level resolution (purple, low density; orange, high density). The methylome of memory B and CLL cells, in contrast to naïve B cells, is strongly hypomethylated and shows hypermethylation in otherwise lowly methylated regions. Bar charts give fraction of hypermethylated (>0.25) and hypomethylated (<0.25) CpGs. N = 3,490,971 (mem-nai); 3,202,573 (unmut-nai); 3,034,005 (mut-nai); 3,202,573 (unmut-mem); 3,034,005 (mut-mem); 2,974,458 (mut-unmut). **C,** Phylogenetic CpG methylation tree of normal B cells (gray shading) and first time point samples of patients with CLL [*IGHV*: unmutated (red) and mutated (orange)] in the context of normal B-cell differentiation states. All CLLs cluster to the more mature end of the tree and separate by the mutational status of the *IGHV* chain genes except for two cases (orange lines). Each line represents a sample. Arrows indicate presumed cell of origin to CLL transition. bmPC, bone marrow plasma cell; gcBC, germinal center B cell; HPC, hematopoietic progenitor cell; PCT, plasma cell from tonsil; preB2C, pre-B-II cell. **D,** Average CpG methylation from bulk in naïve B cells (light gray), memory B cells (dark gray), and unmutated (red) and mutated CLL (orange) across the *EBF3* locus. Both CLL samples exhibit similar levels of modulation across the entire region. Specifically, both CLL samples reveal stronger hypermethylation in the promoter region than the normal B cells and loss of methylation in the usually highly methylated gene body. Although to a much lesser degree, this effect can also be found between naïve and memory B cells. **E,** Average CpG methylation levels for CLL DMRs of unmutated CLL versus naïve B cells and mutated versus memory B-cell comparisons. The rows represent overlapping and mutated or unmutated CLL–specific DMRs. Samples are merged into a mean methylation representation per group (columns) and DMR (rows). Rows were ordered using unsupervised hierarchical clustering. Side annotations for DMR location and chromatin state: hyper- or hypomethylated DMR; CpG density: CpG island, shore, or shelf; location: promoter, gene body, or intergenic; chromatin state: promoter, poised promoter, enhancer, transcription related, insulator, or heterochromatin. **F,** Numbers of hypomethylated (gray) and hypermethylated (blue) DMRs (minimum difference of 0.25, minimum 3 CpG in length) in unmutated CLL versus naïve B-cell and mutated versus memory B-cell comparisons. DMRs are classified as B-cell related or CLL based on their overlap with CpGs that are differentially methylated during normal B-cell development. **G,** Overrepresentation enrichment analysis for genes with CLL DMRs compared with the background, i.e., all DMRs. Enriched pathways (Panther) include PI3K, EGFR, Ras, FGF, and CCKR signaling. A common characteristic of these pathways is their implication in cell survival, gene expression regulation, growth factors, activation of proliferation, and cell invasion. Shown are the top 10 pathways based on P value.



($n = 4{,}653$ DMRs; Supplementary Table S2; ref. 32). To disentangle methylation changes associated with normal cell lineage–specific differentiation from potentially cancer-related changes, we classified the aggregate of these two sets of DMRs as B-cell related ($n = 22{,}325$) or CLL ($n = 3{,}475$), depending on whether they were classified as dynamically changing during normal B-cell development (Fig. 1E and F; Methods; ref. 30). The majority (85%) of the DMRs overlapped with developmental regions, whereas 15% were classified as CLL DMRs (Fig. 1E and F; Supplementary Fig. S1B). Based on the clustering, B-cell lineage–related DMRs showed a gradual shift, mostly toward hypomethylation, from naïve to memory and both CLL subtypes, reflecting the normal B-cell developmental changes that are retained in CLL. In contrast, as expected, the set of CLL DMRs readily distinguished normal B cells from CLL (Fig. 1E). Moreover, genes that were associated with CLL DMRs were found to be overrepresented among pathways related to cell growth and survival, proliferation, and neoplastic transformation, suggesting possible regulatory relevance (Fig. 1G).

We additionally confirmed the DMRs to be a distinctive feature between normal B and CLL cells by analyzing replicates of CD5-positive and -negative naïve and memory B cells from a set of three healthy donors. Genome-wide phylogenetic tree clustering and the correlation of methylation rates revealed two major clusters, separating the samples by naïve and memory B cells but not by CD5 status (Supplementary Fig. S2A and S2B). Based on the presence of the CLL-specific DMRs, CD5-sorted healthy donor samples were found to cluster into the group of previously published reference B cells, hence demonstrating similar methylation in DMRs independent of CD5 status of naïve or memory B-cell state (Supplementary Fig. S2C).

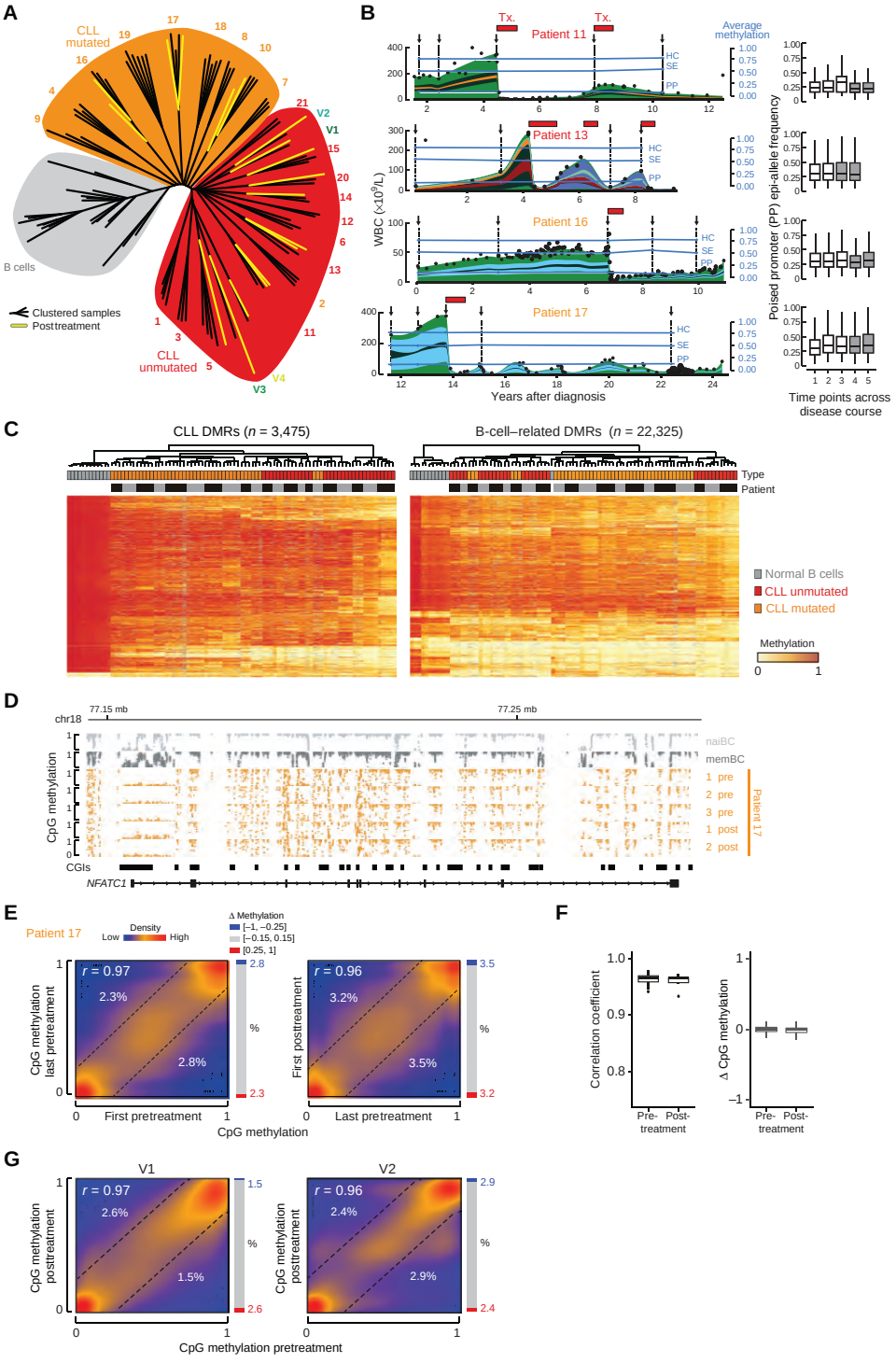
CLL Methylome Remains Mostly Unchanged after Treatment

To evaluate the stability of these DMRs and the dynamics of the CLL methylome over time, we analyzed longitudinal samples collected during natural CLL progression. CLL allows that leukemic burden can be approximately estimated by measuring the WBC count over time since it is, for many patients, primarily a circulating malignancy. To study the CLL methylome before and after the first treatment, we performed unsupervised phylogenetic clustering of the pre- and posttreatment (fludarabine, cyclophosphamide, and rituximab, FCR) samples of patients. Interestingly, we found no consistent methylation differences that separate pre- and posttreatment samples, and also no WBC-related effects could be seen in the clustering (Fig. 2A). Next, we compared methylation levels across patient time points for selected chromatin states derived from published data of the lymphoblastoid cell line GM12878 (33). Despite vastly different growth patterns and subclonal dynamics (defined by prior genetic characterization; ref. 28), global methylation levels of various genomic features, such as heterochromatin, strong enhancers, and poised promoters, for serial samples from all 21 patients remained stable and were consequently independent of the dynamic changes in WBC counts (Fig. 2B, left; Supplementary Fig. S3A and S3B).

We observed substantial posttreatment reduction in WBCs creating population bottlenecks for the nine patients following treatment with FCR (Fig. 2B; Supplementary Fig. S3A). However, this was not associated with any notable DNA methylation changes over the three representative genomic features or the number of distinct epialleles present (Fig. 2B; Supplementary Fig. S3; ref. 34). Indeed, methylation levels were mostly independent of detected subclonal genetic evolution and from patterns of growth. Similar stability was also found at the previously identified DMRs, with once acquired changes appearing to persist during CLL progression (Fig. 2C). These deregulated regions, including the example of *NFATC1* (Fig. 2D), further highlight the remarkable stability of methylation patterns.

We also compared the nine patients by focusing on the clinically most divergent time points, i.e., first and last

Figure 2. The CLL methylome remains mostly unchanged over disease progression, including after treatment. **A,** Phylogenetic tree of normal B cells and all measured time points of patients with CLL ($n = 83$) using global CpG methylation levels. Each line represents a sample; subtrees are multiple samples of the same patient. Yellow lines represent posttreatment samples (chemoimmunotherapy and venetoclax V1–V4). Mutated and unmutated are colored as before, and all patient numbers are shown next to the respective branches. All samples from the same patient clustered together, whereas normal samples are distinct from the CLL cohort. **B,** WBC counts and methylation dynamics for selected genomic features to represent global hypomethylation (HC, heterochromatin; SE, strong enhancers) and hypermethylation (PP, poised promoters) across disease progression for patients 11, 13, 16, and 17 (for all others, see Supplementary Fig. S3A). The methylation levels remain constant over time and after treatment. Black dots: WBC counts (left axis). Blue lines and dots: measurements of CpG methylation levels (right axis). Black arrows and dashed lines indicate collected time points for DNA methylation analysis. Boxplots to the right display the coverage normalized epiallele fraction in poised promoter regions. Treatment exposure at time points is indicated as shaded boxplots. In addition to the methylation level, the epiallelic fractions' distribution stays stable over time and after treatment. WBC plots are taken from the same patients studied in ref. 28 and have been overlaid with our DNA methylation data. Distinct genetically defined subclones are indicated with the different colors. Tx., treatment. **C,** Left, average methylation levels per sample (columns) for CLL DMRs (rows) of unmutated CLL versus naïve B cells and mutated versus memory B-cell comparisons. Rows were ordered using unsupervised hierarchical clustering. Right, average methylation levels per sample (columns) for B-cell–related DMRs (rows) of unmutated CLL versus naïve B cells and mutated versus memory B-cell comparisons. Rows were ordered using unsupervised hierarchical clustering. **D,** Average CpG methylation in naïve B cells (light gray), memory B cells (dark gray), and five serial samples collected from patient 17 (orange, top to bottom: three pretreatment and two posttreatment samples) across the *NFATC1* locus. Dots represent CpG-level methylation of each sample. **E,** Correlation of CpG methylation levels in the first pre- and last pretreatment sample as well as last pre- and first posttreatment of patient 17 at CpG-level resolution ($n = 1{,}912{,}382$). For all other samples, see Supplementary Table S3. Bar charts give fraction of hypermethylated (>0.25) and hypomethylated (<0.25) CpGs. Numbers are given within the scatter. **F,** Boxplot of correlation coefficients of genome-wide CpG-level correlation between the first pre- and last pretreatment samples of all patients with CLL as well as last pre- and first posttreatment samples for the posttreatment CLL (left). Corresponding boxplot of genome-wide CpG-level difference between the first pre- and last pretreatment samples as well as last pretreatment to first posttreatment samples (right). $n = 21$ pretreatment data points and $n = 9$ posttreatment data points. In the boxplots, the centerline is median; boxes, first and third quartiles; whiskers, 1.5× interquartile range; data beyond the end of the whiskers are omitted. **G,** Correlation of CpG methylation levels in the pre- and posttreatment samples of the venetoclax-treated patients V1 and V2 at CpG-level resolution. For all other samples, see Supplementary Fig. S4. Bar charts give fraction of hypermethylated (>0.25) and hypomethylated (<0.25) CpGs. Numbers are provided within the scatter.



pretreatment and last pre- and first posttreatment. No joint DMRs between all first pretreatment versus last pretreatment time points could be detected. The variability between samples of the same patient and the lack of shared events appear to be more in line with patient-specific evolution than a common path across patients. Moreover, correlation analysis on the CpG level also confirmed a largely stable methylome across CLL evolution and even after treatment (Fig. 2E and F).

To explore if the observed methylation stability is therapy specific, we next analyzed four patients treated with the BCL2 inhibitor venetoclax (35). As with the FCR chemoimmunotherapy-treated patients, these CLL samples collected before and after venetoclax exposure clustered tightly together, within the group of chemoimmunotherapy-treated patients (Fig. 2A and G). We further confirmed the stability of the B-cell–related and CLL DMRs in this treatment cohort, in which we could only detect globally on average less than 6% of CpGs to vary between pre- and post-venetoclax treatment (Supplementary Fig. S4A–S4D).

Combined with the early emergence of the altered methylation landscape, our posttreatment results highlight the striking stability of the CLL methylome, with minimal changes over disease progression, including after treatment.

Variations in the CLL Methylome Appear Stochastic among Patients

Because only a few patient-specific methylation dynamics were observed, we assessed if their occurrence exceeded random dynamics present among normal B-cell subtypes. Focusing on the chemoimmunotherapy-treated patient samples, we first compared the number of dynamic CpGs between first and last pretreatment, and last pre- and first posttreatment CLL samples with differences between biological replicates of naïve and memory B cells (Supplementary Fig. S5A; Supplementary Table S3). Although we did not detect any correlation with the time to treatment, we observed the fraction of dynamic CpGs to be slightly higher in posttreatment samples. Overall, only 1 of 21 patients stood out with higher variability ($r = 0.94$ pretreatment and $r = 0.93$ for pre- to posttreatment comparison); however, this is very similar to the variation observed at the transition of naïve to memory B cells ($r = 0.95$). Moreover, the relative number of CpGs that exhibited substantial differences was less than 5% of all considered CpGs for most CLL cases, and those CpGs were frequently located in heterochromatin regions and outside of CpG islands (Supplementary Fig. S5B). Most of the dynamic CpGs were restricted to individual patients, and 99.9% of CpGs were shared within a maximum of six and four patients for pre- and posttreatment comparisons, respectively (Supplementary Fig. S5C), thus again confirming limited variation and a high degree of stability of CLL methylome over time.

To further separate random single dynamic position events from consistently altered regions during disease progression and following treatment, we focused on patient-specific DMRs identified either between first and last pretreatment for all 21 patients (Supplementary Fig. S5D, left), or between the last pre- and first posttreatment for the nine patients (Supplementary Fig. S5D, right). We found a median of 106 pre- and 143 posttreatment DMRs per sample. Of these, the vast majority (89% and 86%) appeared in regions shared with normal B-cell development. The remaining DMRs comprised only 8% of the aforementioned dynamic CpGs. However, with 50% of the dynamic CpGs still localized to CLL-specific regions, our data suggest the presence of randomly modulated individual CpGs rather than stretches of adjacent CpGs. Furthermore, about half of the CLL DMRs were located in heterochromatin, supporting the assumption that these may represent a secondary effect (Supplementary Fig. S5E). Finally, gene set enrichment analysis revealed pathways supported by only a few genes (i.e., B-cell and T-cell pathways supported by three genes and p53 by two genes) or pathways with no apparent link to CLL (Supplementary Fig. S5E).

In sum, although a low number of patient-specific methylation changes accompany the individual tumor evolution, we observed remarkable stability and similarity of the acquired CLL methylome across patients.

The Altered Methylome Is Already Detectable at the MBL Stage

Because we observed an altered methylome already present within the first time points of the characterized CLL specimens, we next turned to specimens collected from our patients with MBL to evaluate the cancer precursor methylome. We again performed unsupervised phylogenetic clustering but now including the high-count MBL samples. Despite clinical classification as a precursor state, all of the MBL cases were found to cluster directly among the group of CLL samples and not to branch earlier from the trunk of this tree (Fig. 3A). Most strikingly, all five matched MBL–CLL cases appeared to be as similar to each other as the biological replicates of different healthy B-cell types and much more similar to each other than to the other CLL cases. Thus, patient-specific methylation signatures appeared to be stronger than any preleukemic versus leukemic methylation signature, which would have otherwise resulted in the separate clustering of the MBL samples from the CLL samples.

Strikingly, the identified CLL-associated DMRs were also present in our patients with MBL (Supplementary Fig. S6A and S6B). We extended the analysis to search for additional consistently occurring methylation changes that could potentially drive the MBL-to-CLL transition. However, no statistically significant DMRs could be detected between the methylomes of the individuals that transitioned from MBL to CLL. As a representative example of this shared landscape between MBL and CLL, we show the methylation patterns for the gene *NFATC1*, which has been reported as overexpressed in CLL due to loss of epigenetic repression (36) and is an upstream effector of BCL2, which itself is frequently deregulated due to chromosomal translocations in B-cell malignancies (Fig. 3B). Through a correlation analysis at single CpG resolution of the methylomes of the matched MBL and CLL pairs, we further observed the striking similarity between MBL and CLL methylomes. These results revealed only minor, if any, targeted remodeling of the methylome between the precursor and CLL stages within a given patient (Fig. 3C; Supplementary Fig. S6C). To appreciate this high similarity, we note that comparably high correlations are otherwise found between biological replicates of flow cytometrically isolated normal B-cell subpopulations.

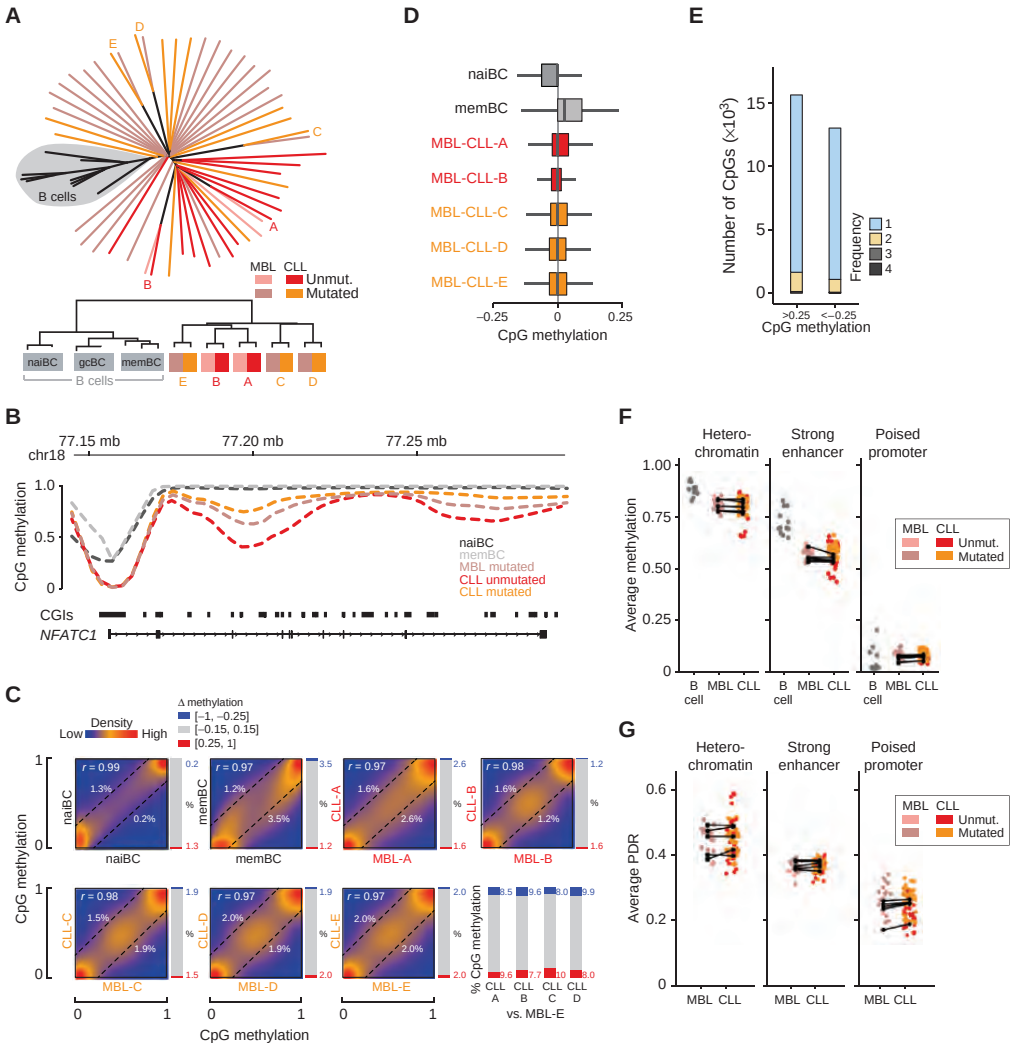


Figure 3. CLL methylation signatures are already present in MBL. **A,** Phylogenetic tree of normal B cells, MBL, CLL (first pretreatment sample per case), and matched MBL–CLL pairs using all CpGs ($n = 3.5$ million). Left, all MBL cases clustered together with CLL cases. Each line represents a sample. Right, unsupervised hierarchical clustering of normal B cells and matched MBL–CLL pairs. Matched MBL–CLL cases clustered as closely as biological replicates of normal B cells; mean joint fork distance (log_2FC) of matched versus mismatched MBL–CLL versus normal B cells: -0.03 versus -1.06. gcBC, germinal center B cell. **B,** RRBS-based genome browser tracks. Average methylation is shown as a smoothed line across the *NFATC1* locus, grouped by naïve B cells (light gray, $n = 3$ samples), memory B cells (dark gray), unmutated (red) and mutated CLL (orange), and MBL (brown). Naïve and memory B cells are hypermethylated in the gene body, with only a small drop at the transcription start site. Unmutated and mutated CLL are hypomethylated across the promoter and gene body. **C,** Correlation of CpG methylation levels in matched MBL–CLL pairs ($n = 5$), in biological replicates of normal B cells (naïve B cells and memory B cells), at CpG-level resolution ($n = 1{,}801{,}907$ CpGs). Plots show a high correlation between biological replicates of normal B cells and a similarly high correlation between MBL–CLL pairs. Numbers give quantification of hypermethylated (>0.25) and hypomethylated (<0.25) CpGs. For comparison of CpG differences with mismatched samples, the quantification of all CLL samples versus one MBL is shown. $r = 0.97$–0.98 for matched pairs, $r = 0.85$–0.9 for mismatched pairs, $r = 0.99$ for naïve B cells, and $r = 0.97$ for memory B cells. **D,** CpG-level resolution differences between matched MBL–CLL pairs and among biological replicates of normal B cells. For all MBL–CLL comparisons, a maximum of 4.2% of positions show a difference of >0.25 ($n = 1{,}801{,}907$ CpGs). Methylation differences among biological replicates were less than 2% (naïve B cells) and 3.6% to 9.5% (memory B cells) of CpGs with a difference of >0.25. Per boxplot, the median value is indicated by the centerline, with first and third quartiles as outlines of boxes, and 1.5× interquartile range as whiskers; data beyond the end of the whiskers are omitted. **E,** Number of CpGs with a difference of >0.25 in MBL–CLL comparisons and frequency of recurrent observations across the five pairs (light gray, unique for one pair, to dark grey, observed in four of five pairs, no five of five detected). The minority of CpGs are recurrently differentially methylated. **F,** Comparison of average chromatin state methylation among DNA from normal B cells, MBL, and CLLs. Black horizontal lines, matched MBL–CLL pairs ($n = 1{,}801{,}907$ CpGs). **G,** Comparison of chromatin state proportions of discordant methylation rates (proportion of discordant reads, PDR) between MBL and CLLs. Black horizontal lines, matched MBL–CLL pairs.

Also, at single CpG resolution, we found that at most 4% of all covered CpGs showed a difference of 0.25 or greater between MBL and CLL samples of the same patient (Fig. 3D). An expanded analysis to examine whether individual CpGs were conserved targets across patients revealed this is not the case (Fig. 3E). Finally, we compared methylation and read-discordance levels for different chromatin states. This showed that differences and variability compared with normal B cells affect MBL and CLL cases to the same degree, again highlighting that the similarity of MBL and CLL is not merely based on patient identity (Fig. 3F and G).

Lastly, as most of our patients with MBL had already relatively elevated WBC counts, we further extended our investigation to include CD5$^+$ sorted cells from three patients with low-count MBL (Supplementary Fig. S7A). Although the signal is expectedly not as strong due to the rare proportion of cells sequenced and the potential contamination with CD5$^+$ normal B cells, we do detect evidence of the same characteristic epigenetic alterations as observed in high-count MBL and CLL (Supplementary Fig. S7B–S7D).

Taken together, our comprehensive analysis of 53 patients and 101 pretreatment RRBS datasets suggests that the transition to the cancer methylome occurs early in the disease. Analysis of patient-matched MBL and CLL shows that no consistent additional DNA methylation changes are seemingly associated with disease progression.

Heterogeneous Expression Patterns Are Present Per Patient but Are Stable across Natural Disease Progression

Complementing our methylation analysis, we profiled the transcriptomes of approximately 60k single cells isolated from healthy donors and the five matched MBL–CLL specimens (Fig. 4A). Unsupervised clustering revealed nine distinct clusters: four clusters representing peripheral blood mononuclear cells of the two healthy donors and the remaining five, with each representing one of five patient B cells (Fig. 4B). From the healthy donors, cell clusters of myeloid and lymphoid origin were readily identifiable based on their marker gene expression. In contrast, the five clusters from patients were identified as CLL/MBL-mixed clusters that showed expression of some B-cell marker genes, although less pronounced (Fig. 4C). When looking more specifically at differentially expressed genes, we found lower and heterogeneous expression for some characteristic B-cell markers and similarly heterogeneous upregulation of genes such as KLF2 and CD27 in the patient samples (Fig. 4D). Of note, the MBL and CLL cells per patient were transcriptionally indistinguishable.

Because the transcription-based clustering could not distinguish the MBL and CLL cells, we instead used barcode information per cell for annotation (Fig. 4E). We observed a remarkable overlap for most clusters, supporting the striking similarity of the MBL and CLL transcriptomes. Only the MBL and CLL cells from patient A were slightly separated in the UMAP visualization. However, upon evaluation of the highest-ranked marker genes for each MBL–CLL set, we found a surprisingly high concordance of expression of even the most differential expressed genes between MBL and CLL cells (Fig. 4F).

Based on the single-cell expression profiles, the differences between MBL and CLL appear to be marginal, which agrees with the lack of separation by clustering on single-cell transcriptomes. Combined with the lack of DNA methylation changes in the MBL-to-CLL transition, it points to an earlier molecular event that sets the normal cells already on the path to tumorigenesis.

Individual MBL and CLL Cells Show Little DNA Methylation Heterogeneity

Our bulk data indicated an early conserved switch in the DNA methylation landscape across patients with CLL, and our single-cell transcriptome data demonstrate the transcriptional similarity between matched MBL and CLL. However, bulk measurements cannot completely distinguish the contributions of diverse cellular subpopulations to the overall picture. Subclonal evolution and genetic heterogeneity are common in CLL (37–39). This understanding motivated us to investigate single-cell methylation maps from two patients with CLL, two patients with MBL, and age-matched B cells collected from two healthy adult volunteers, uncovering a stable level of mean methylation per cell on a global scale (Fig. 5A; Supplementary Table S1; ref. 40).

Analysis of our previously defined DMRs showed the presence of aberrant methylation levels in all MBL/CLL cells with sufficient coverage (Fig. 5B). When comparing naïve to memory with MBL and CLL cells, a gradual gain of methylation in B-cell–related and CLL hyper-DMRs was observed. Conversely, hypomethylated B-cell–related and CLL DMRs appeared slightly stronger in separating normal from diseased (MBL and CLL) cells (Fig. 5B). Phylogenetic clustering separated CLL, memory B, and naïve B cells, with no differences between the B-cell subpopulations with or without the presence of CD5 (Fig. 5C; ref. 12). Moreover, we observed a clear separation between MBL and CLL versus normal, with each forming a tight cluster in line within the observed stability of the methylome per patient. Of note, memory B cells, despite many shared features with the CLL methylome, cluster distinctly next to the naïve B cells and apart from the MBL and CLL cells.

Our genome-wide single-cell methylation analysis thus complements our bulk data by further showing the clear methylation difference between MBL and CLL compared with sorted B-cell subtypes.

DISCUSSION

We show that the aberrant cancer methylome in CLL is already established at the preneoplastic MBL stage and is consistently present at the time of diagnosis across samples collected from 3 low-count and 20 high-count MBL, 5 matched MBL–CLL pairs, and 25 patients with CLL. Although normal B-cell maturation shows some similarities with the CLL methylome, these normal developmental changes are likely insufficient to transform cells into proliferative MBL and CLL. Nonetheless, the shared targets make a better understanding of the underlying mechanism and biological reason highly relevant. We also find a limited set of cancer-specific targets that can be readily applied to distinguish all normal B-cell subtypes from MBL and CLL. These CLL DMRs are

A

Donor	Sample	# Cells	IGHV
HD_1	Blood	8,613	
HD_2	Blood	8,966	
A	MBL/CLL	3,702/3,292	Mutated
B	MBL/CLL	4,340/5,368	Mutated
C	MBL/CLL	5,183/4,654	Unmutated
D	MBL/CLL	4,335/2,508	Unmutated
E	MBL/CLL	4,641/5,028	Mutated
		Σ 60,630	

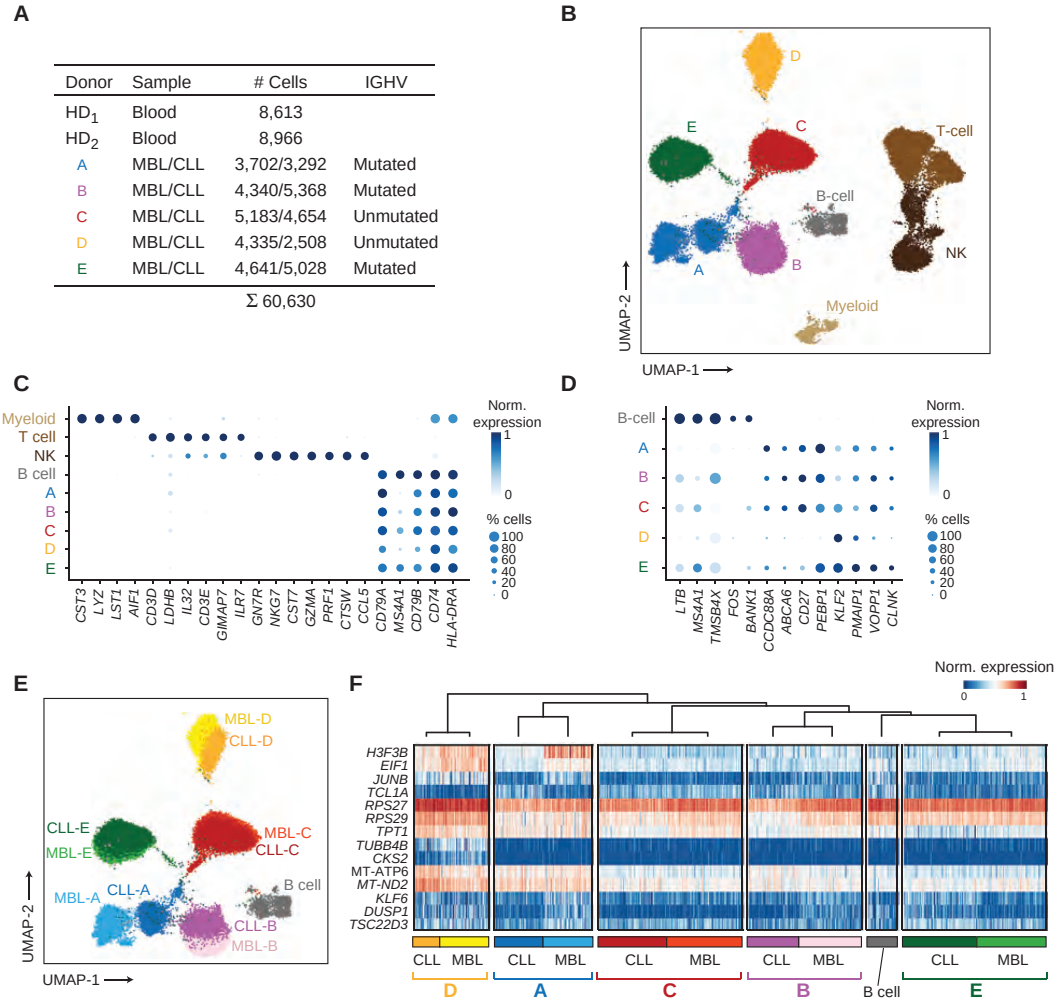


Figure 4. Single-cell transcriptome analysis. **A,** Summary table with details of donors, tissue source, number of cells per sample, and *IGHV* status. Patients were profiled in MBL as well as the CLL state using the 10x Genomics Chromium droplet single-cell RNA sequencing. HD, healthy donor. **B,** UMAP displaying the groups found using the Louvain algorithm. The healthy donor cells split into B cells, T cells, myeloid cells, and natural killer (NK) cells. The MBL and CLL cells of patients build distinct groups but are within a patient not distinguishable. **C,** Normalized gene expression level and the number of positive cells of marker genes used to identify normal cell types in **B**. B-cell–specific genes show an aberrant expression profile in clusters derived from patient cells. **D,** Normalized gene expression level and the number of positive cells of genes identified as marker genes between B cells and all patient cells (Wilcoxon rank-sum test). **E,** UMAP of MBL, CLL, and B cells with artificially introduced identification of MBL and CLL cells. MBL and CLL cells cluster farther apart from the B cells than from other cells of the same patient and are highly overlapping for almost all patients. **F,** Heatmap displaying single-cell gene expression of highest-ranking marker genes between MBL and CLL cells of the same patient (Wilcoxon rank-sum test). Expression levels are very similar among cells of the same patient as compared with other patients or B cells, a parameter that is also supported by hierarchical clustering.

overrepresented among pathways involved in proliferation, cell survival, and growth. Although it remains technically challenging to experimentally explore if, for instance, the addition of just these CLL DMRs alone is sufficient to drive the tumorigenic transition or facilitate extended and rapid proliferation, we anticipate that these targets are certainly worthy of future exploration, including with emerging epigenome editing tools.

Our results provide a comprehensive picture of the DNA methylation alterations in MBL and CLL and demonstrate that the switch to an abnormal landscape has consistently occurred before any of our measured time points. This notably expands findings from prior array-based studies (41, 42) and complements recent work on the genetic evolution across the 21 CLL samples (28). Similar early alterations of the methylome have also been noted in colorectal

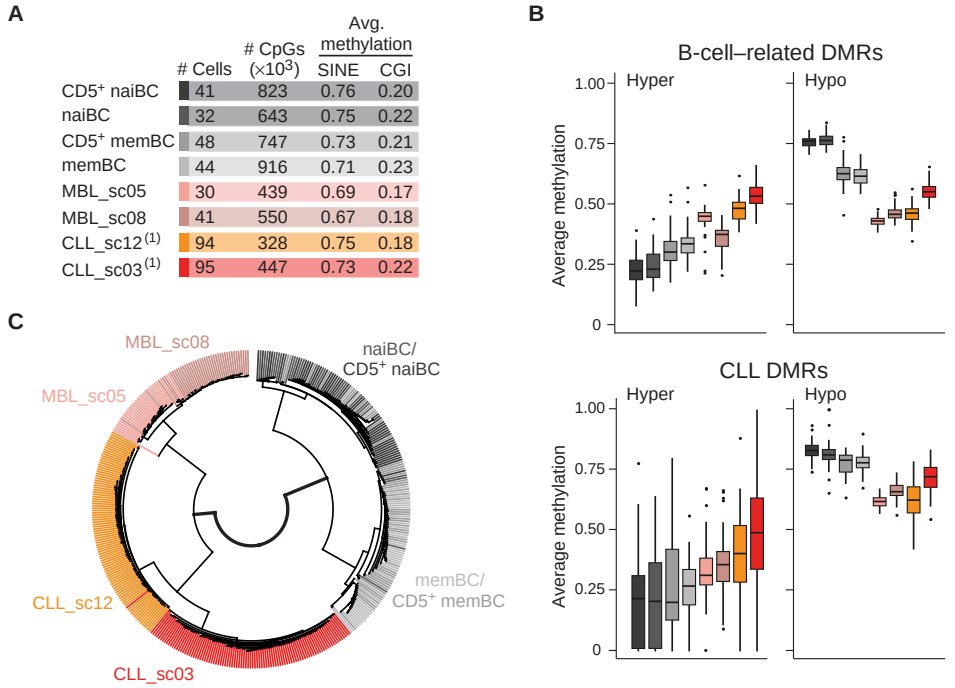


Figure 5. Single-cell DNA methylation analysis of MBL, CLL, and normal B cells. **A,** Summary table of methylation data generated from MBL (rose, brown), CLL (red, orange), and flow cytometrically isolated normal B-cell subpopulations [naïve B cells, $CD19^+CD27^-$ (lightest gray), $CD5^+$ naïve B cells ($CD19^+CD27^-$, light gray), memory B cells ($CD19^+CD27^+$, dark gray), $CD5^+$ memory B cells ($CD19^+CD27^+$, darkest gray)]. $n = 611{,}452$ CpGs covered on average. (1)Data for CLL were taken from Gaiti and colleagues (40). **B,** Average methylation levels for B-cell–related and CLL DMRs per cell confirm aberrant methylation of these regions being consistently observed across all cells. CLL and MBL cells show strong patterns of hypomethylation, whereas the small number of hypermethylated DMRs ($n = 888$ and 70, respectively) is already present in MBL but also seems to be slightly more prevalent in CLL cells. **C,** Unsupervised hierarchical clustering of MBL, CLL, and normal B cells. MBL and CLL cluster together into one clade. Although the clonal MBL and CLL separate by donor, naïve and memory B cells are intermingled with their CD5-positive counterparts.

cancer where the aberrant methylation landscape is already detectable in premalignant colorectal adenomas and amplifies upon the colorectal cancer state (43). Taken together with the near universality of the altered cancer methylome (5), the possibly conserved early emergence points to an important role for the epigenetic change in a tumorigenic transition. Although it is difficult to establish causality, we speculate that the altered landscape may provide a receptive platform for the disease progression. Alternatively, the cancer methylome may simply be a consequence of a developmental program that may regulate numerous cellular attributes, including methylation (5). In the latter case, it may well be other features driving the tumorigenic transition, and the methylome is only one biological consequence of the entire program. Although this seems possible, it should be noted that this altered DNA methylation landscape is maintained across patients sometimes for decades and found in nearly all cancer types, raising the question of why it is not diverging if it has no functional role. Another possibility that we can consider here is that the altered methylome presents an optimized epigenetic state to maintain viability with maximum proliferation and the minimal energy requirement for DNA methylation maintenance.

Aside from these considerations, we note that the MBL and CLL methylome and transcriptome are extremely stable once acquired. In contrast to the dramatic fluctuations in tumor burden (estimated by the changes in the level of WBC counts) across disease course, methylation levels are not consistently affected by clonal expansion or treatment-induced bottlenecks. The latter may reflect that cells surviving treatment represent either the average of all subclones or that limited methylation heterogeneity is present across all subclones. From a practical standpoint, the stability of the methylome in patients with CLL limits its utility to track disease progression. Still, it may be valuable for early detection and helpful to assess the efficiency of treatments. We observed neither any notable consistency of dynamic CpGs along with the MBL to CLL nor CLL progression and treatment, indicating that the few observed dynamics over the disease progression are possibly an accumulative secondary effect. During disease progression, considerable increases in WBC counts are only juxtaposed with subtle methylome changes. These largely constant methylation levels within each patient indicate that increased clonal expansion occurred without substantial additional departure from the preexisting, already aberrant landscape. The stability of the altered state is further supported by our single-cell

transcriptomes from the five patients that transition from MBL to CLL without any major expression dynamics. Finally, our finding that the cancer methylome also remains mostly unaffected by conventional chemoimmunotherapy or the BCL2 inhibitor venetoclax may keep patients at an elevated risk for relapse, in line with the fact that CLL is rarely cured, although this treatment landscape is continuously evolving (17).

Genetic and epigenetic diversity of normal tissues, tumors, and even clonally amplified cell populations have been most broadly assessed in-depth so far using bulk sequencing. To date, the degree of heterogeneity in methylation levels among distinct genomic regions within a single cell has not been investigated to our knowledge. Here, we applied single-cell methylome analysis and could show that aberrant methylation affects single cells to a surprisingly similar extent. Despite these convincing findings, it needs to be stated that due to technical limitations, such as stochastically missing values caused by unequal coverage, parts of the genome have not been investigated. Nevertheless, we could confirm that hypomethylation is more pronounced across individual MBL and CLL cells, suggesting that the methylation machinery targets a consistent and specific set of regions, though single CpGs are generally affected discordantly (20). As a result, single-cell analysis can help classify tumors and their microenvironment, which provides a more holistic disease picture and may guide more precise treatments in the future.

In sum, our comprehensive exploration over the disease course of CLL, including its precursor stage MBL, highlights several important lessons toward a better mechanistic understanding of the cancer methylome. First, the transition to the altered methylome occurs very early, possibly as a nongenetic precursor lesion that is not yet tumorigenic. Second, it should play at least some facilitating, if not central, role as it was present in all 53 patients evaluated and at all stages with remarkable stability. And finally, its persistence after treatment, though currently limited to the 13 patients that we investigated, suggests that the current chemoimmunotherapy and BCL2 inhibition approach eradicates only some but not all diseased cells. Although it remains to be investigated, the general nature of the epigenetic transformation extends to other cancer types, including many solid tumors, suggesting that this landscape may reflect convergence toward a commonly utilized regulatory mechanism.

METHODS

Human Samples

Heparinized blood samples were obtained from normal donors and patients enrolled in clinical research protocols approved by the Human Subjects Protection Committee of the Dana-Farber Cancer Institute (DFCI), at University of California, San Diego and the Mayo Clinic (CLL Research Consortium), and through the International Cancer Genome Consortium (42) after obtaining written-informed consent. Treatment indication for all 21 patients in the discovery cohort was determined based on International Workshop on Chronic Lymphocytic Leukemia criteria (12, 44). Peripheral blood mononuclear cells (PBMC) from normal donors and patients were isolated by Ficoll/Hypaque density gradient centrifugation. Mononuclear cells were used fresh or cryopreserved with 10% DMSO FBS and stored in vapor-phase liquid nitrogen until the time of analysis. $CD19^+$ B cells from normal volunteers and CLL samples with WBC $\leq 50 \times 10^9$/L were isolated by immunomagnetic selection (Miltenyi Biotec) and stained with anti–CD19-phycoerythrin (PE; BioLegend) prior to FACS sorting for live single cells in the presence of DAPI. MBL cells and naïve and memory B cells from age-matched healthy donors were isolated as follows: Cryopreserved PBMCs were thawed and stained with anti–CD19-PE, CD5-FITC, and CD27-Allophycocyanin (APC; BioLegend). Cells were gated for naïve B cells ($CD19^+$, $CD27^-$, and CD5), memory B cells ($CD19^+$, $CD27^+$, and $CD5^-$), or MBL ($CD19^+$ and $CD5^+$; Supplementary Fig. S7A).

Bulk RRBS Library Generation and Data Processing

RRBS libraries were generated from 25 to 100 ng of input DNA using the Ovation Methyl-Seq System (NuGen) following the manufacturer's recommendation. We used NuGen unique molecular identifier (UMI) technology to measure the rate of PCR duplicates on one patient (four samples) and found the duplicate rate to be below 2%, even at an input of only 25 ng of DNA. On average, 15.7M fragments, resulting in 31.4M paired-end 101–base pair (bp) reads, were sequenced per sample on an Illumina HiSeq2500. These reads were aligned to the human hg19 genome using BSmap (45) with flags -v 0.05 -s 16 -w 100 -S 1 -p 8 -u. An average of 21.1M reads per sample was aligned correctly. Custom scripts written in Perl were used to count the number of times a CpG was observed to be methylated. The methylation percentage for each CpG was calculated as the number of times the CpG appeared methylated divided by the total times the CpG was covered in sequencing reads. Finally, we converted the resulting CpG level files to bigWig files, filtering out all CpGs covered with less than five reads. An average of 3.4M CpGs was covered per sample at an average depth of 14×.

Multiplexed Single-Cell RRBS Library Generation and Data Processing

Single-cell RRBS libraries were prepared by combining the first steps (cell lysis and physical separation of DNA and mRNA) of the single-cell methylome and transcriptome sequencing protocol (46) with multiplexed single-cell RRBS (26) using double MspI+HaeIII digestion. Single cells were sorted into 5 μL RLT plus buffer (QIAGEN) containing 1 U SUPERase in RNase inhibitor (Invitrogen) in 96-well PCR plates, flash-frozen on dry ice, and stored at −80°C. Upon thawing, 5 μL of QIAGEN RLT plus buffer and 10 μL M-280 streptavidin beads conjugated to a biotinylated oligo-dT primer were added to each well. After 30 minutes at 25°C, the plates were transferred to a magnet to capture bead-bound mRNA, and the DNA-containing supernatant was transferred to a new 96-well plate. Beads in the original wells were washed twice with 15 μL of washing buffer (50 mmol/L Tris-HCl, pH 8, 75 mmol/L KCl, 3 mmol/L $MgCl_2$, 10 mmol/L DTT, and 0.5% Tween-20), and each wash was added to the DNA plate. To clean up the DNA, 1 volume of a 1:5 dilution of AMPure XT SPRI beads (Beckman Coulter) in 20% PEG/2.5 mol/L NaCl and 0.5 μL Proteinase K (0.8 U/μL, NEB) were added. After 30 minutes at 25°C with mixing, the beads were washed with 80% ethanol and genomic DNA eluted with 8 μL H_2O, with the beads remaining in the well during library prep. After addition of 2 μL 1x CutSmart buffer (NEB) containing 10 U of MspI (NEB), or 5 U of MspI plus 5 U of HaeIII (NEB), DNA was digested for 2 hours at 37°C, followed by heat inactivation for 15 minutes at 65°C. MspI sites were filled in and fragment ends adenylated by adding 2 μL 1× CutSmart containing 2.5 U Klenow fragment (3′–5′exo-, NEB), 0.4 μL of dNTP mixture (10 mmol/L dATP, 1 mmol/L dCTP, and 1 mmol/L dGTP) followed by a two-step incubation for 25 minutes at 30°C and 30 minutes at 37°C and heat inactivation at 70°C for 10 minutes. After addition of 3 μL 1× CutSmart containing 800 U T4 DNA ligase (NEB), 0.1 μL of 100 mmol/L ATP (Roche), 1.5 μL of 0.1 μmol/L custom 5mC-substituted and indexed (inline barcode) adapter, overnight ligation at 16°C, and heat inactivation (20 minutes at 65°C), 24 separately indexed ligation reactions were pooled. After addition of 3 μL sheared and dephosphorylated *Escherichia coli*

carrier DNA (27), DNA was cleaned up with 1.8 volumes of AMPure XP beads (Beckman Coulter), eluted off the beads, and bisulfite converted (EpiTect Fast Bisulfite kit, QIAGEN) following the manufacturer's recommendations with extended conversion time (20 minutes each cycle). Each pool of RRBS libraries from 24 single cells was PCR amplified using KAPA HiFi Uracil+ DNA Polymerase, a universal P5, and a pool-specific indexed P7 primer for a total of 17 cycles. The thermoprofile was 98°C denaturation for 45 seconds, 6 cycles of 98°C for 20 seconds, 58°C annealing for 30 seconds, and 72°C extension for 1 minute, followed by 11 cycles of 98°C for 20 seconds, 65°C annealing for 30 seconds, and 72°C extension for 1 minute, and a final extension at 72°C for 5 minutes. To minimize size bias during sequencing, multiple PCR products, each representing 24 single cells, were pooled together and size selected on a 2% NuSieve agarose gel into two fractions (150–400 bp and 400–800 bp) that were sequenced in separate lanes with a 10% spike-in of a library with a balanced base composition, which is typically 2 lanes (1.5 plus 0.5 lanes for the low and high size cut, respectively) for 96 cells. On average, 4.5M fragments, resulting in 9M paired-end 75-bp reads, were generated per sample on an Illumina HiSeq4000.

Sequencing reads were demultiplexed using the inline barcode, adapters were trimmed, and reads were trimmed for quality. These reads were aligned to the human hg19 genome using BSmap with flags -v 0.1 -s 12 -w 100 -S 1 −q 20 −u −R. An average of 8.4M reads (4.2M pairs) per sample was aligned. To determine the methylation state of all CpGs captured and assess the bisulfite conversion rate, we used the mcall module in the MOABS software suite with standard parameter settings (47). Finally, we converted the resulting CpG level files to bigwig files, filtering out all CpGs covered with more than 250 reads resulting in an average of 1.1M CpGs covered per sample.

10x Single-Cell RNA Library Generation and Data Processing

PBMCs were thawed in Roswell Park Memorial Institute 1640 medium supplemented with 10% FBS and centrifuged at 1,500 rpm for 5 minutes. Each sample was filtered through a 70-μm filter. Cells were resuspended in PBS–0.04% BSA and stained with anti-human CD5 (FITC), CD19 (PE), CD27 (APC), and 7-aminoactinomycin D for 15 minutes on ice (BioLegend). The samples were washed and resuspended in PBS–0.04% BSA at a concentration of 10×10^6 cells/mL. Samples from the same patient were processed and sorted in parallel on the same day using two FacsAria II cytometers (Becton Dickinson). Cells were sorted through a 70-μm nozzle into 1.5-mL Eppendorf tubes with 10-μL PBS–0.04% BSA and immediately stored on ice. Cellular suspensions were loaded on a 10x Genomics Chromium Controller platform (10x Genomics, Inc.) to generate a single-cell Gel bead in Emulsion (GEM). Single-cell RNA sequencing (scRNA-seq) libraries were prepared as previously described (48).

The Cell Ranger pipeline (10X Genomics, Inc.) was used for each scRNA-seq dataset to demultiplex the raw base call files, generate the fastq files, perform the alignment against the mouse reference genome hg19, filter the alignment, and count barcodes and UMIs. Outputs from multiple sequencing runs were also combined using Cell Ranger functions.

Data Analysis

If not stated otherwise, all statistics and plots are generated using R version 3.5.1 "Feather Spray." In all boxplots, the centerline is median; boxes, first and third quartiles; whiskers, 1.5× interquartile range; and data beyond the whiskers' end are omitted.

Bed files were processed using UCSCtools and bedtools (v2.25.0).

Additional Data

Whole genome bisulfite sequencing (WGBS) data of normal B cells were obtained from the European Genome-phenome Archive (EGA) under accession EGAS00001001196 for comparison in the phylogenetic and average methylation analysis. Methylation data were filtered for minimum coverage of 10× read coverage, and coordinates were converted to hg19 using bedtools liftOver (49). Additional WGBS data of normal B cells were obtained from Beekman and colleagues (30) and downloaded from http://resources.idibaps.org/paper/the-reference-epigenome-and-regulatory-chromatin-landscape-of-chronic-lymphocytic-leukemia. Methylation data were filtered for a minimum of 10× read coverage, and coordinates were converted to hg19 using bedtools liftOver. Data were used in phylogenetic comparisons, average methylation analysis, and all comparative analysis, e.g., detecting differential methylated regions and genomic region visualizations. To ensure accurate comparison among samples, all published WGBS data were reduced to positions covered in any of our patient RRBS data, resulting in a set of approximately 5 million comparable CpGs.

Single-cell RRBS and RNA-seq data of the two CLL samples were obtained from Gaiti and colleagues (40).

Feature Annotations

Chromatin states were defined by the standard 15-state model using the ChromHMM algorithm (33) and were downloaded from the UCSC Genome Browser (50). The average methylation rate for each sample and per chromatin state was calculated as the mean of all methylation rates overlapping with a particular chromatin state. CpG islands were downloaded from the UCSC Genome Browser; shores were defined as adjacent 2 kb regions and shelves as the next adjacent 2 kb regions. Gene annotations were downloaded from the UCSC Genome Browser (gencode v19), and promoter regions were defined as 5,000 nt upstream to 2,500 nt downstream of annotated transcription start site. DMRs were assigned to genes if overlapping with the promoter or gene body with at least one shared base. For unique annotation of DMRs and if a DMR overlapped more than one feature, the ranking was: promoter, gene body, last intergenic; or CpG island, shelf, and last shore. For chromatin state annotation of DMRs, the 15 chromatin states were collapsed into the 6 main categories (Active Promoter, Poised Promoter, Enhancer, Transcription, Insulator, and Heterochromatin), and each DMR was assigned to the region with its maximal overlap.

Mutation and subclone information was taken from ref. 28.

Phylogenetic Tree

The phylogenetic analysis of DNA methylation was performed as previously described (51). In brief, the phylogenetic trees were inferred using the fastme.bal function in the R package ape, which is based on the minimal evolution method. Trees were computed by applying the fastme.bal function on the Euclidean distance matrices of the methylation rates of all samples in the tree. To always capture the maximal information, the subset of CpGs considered was adapted to the sample shows, resulting in n = (i) 28,343,743; (ii) 5,227,401; and (iii) 3,490,971 CpGs for (i) normal B cells, (ii) normal B cells + first time point CLL, and (iii) normal B cells + first time point CLL + MBL; normal B cells + all CLL time points, respectively.

Scatter Plots and Correlation

Scatter plots were created using the smoothScatter function of R, and correlations were calculated using the cor function of R. For the first figure, the average methylation per group was used (n = 3 for naïve and memory B cells, n = 20 and 21 for unmutated and mutated CLL, respectively). Missing values were removed from the mean calculation. For the matched MBL–CLL correlation, missing values were

not omitted from the naïve, memory, and the five matched MBL and CLL samples, resulting in 1,801,907 CpGs that were used for correlation analysis and scatter plots. Statistics for the full CLL cohort are given in Supplementary Fig. S5A.

Genome Region Visualization

Visualization of methylation levels per CpG at genomic regions was done using the plotTracks function of the R package Gviz (52). Track data were grouped by cell type, and for the curve, representation was plotted as a smoothed line.

Differential Methylation Analysis

DMRs were called using metilene version 0.2–7 (32). DMRs were defined to have an absolute minimum difference in methylation of 0.25 with a maximum distance of 100 nt between CpGs within a DMR and a minimum of 3 CpGs per DMR (parameter −M 100 −m 3 −d 0.25). DMRs were calculated between (i) the normal B cells from Beekmann and colleagues (30), (ii) CLL samples as well as between the first and last pretreatment and the last pre- and first posttreatment time points, and (iii) for each sample individually between the first and last pretreatment and the last pre- and first posttreatment time points. More specifically, DMRs were calculated for normal B cells to CLL between the naïve B cells ($n = 3$) and the first time point of the CLL samples with unmutated *IGHV* as well as between the memory B cells ($n = 3$) and the first time point of the CLL samples with mutated *IGHV*. Only positions on the autosomes (chr1–22) were taken into account that were covered by all three normal B-cell samples (naïve or memory) and 90% of the CLL samples ($n = 18$ unmutated/$n = 19$ mutated CLL), respectively. For within-patient sample versus sample DMRs, all positions that were covered by both samples were taken into account. After DMR calling, all *P* values were adjusted for multiple testing using the R function p.adj, and regions with an adjusted *P* value < 0.05 were considered DMRs. As previously described, DMRs were separated into DMRs that overlap CpGs that are dynamic during B-cell differentiation (difference between all normal B cells > 0.25) and subsequently called B-cell–related DMRs, and those that do not overlap any dynamic position are called CLL DMRs. No DMRs (FDR < 0.05) were found when comparing the five matched MBL samples with the respective five CLL samples, when comparing the unmutated and mutated CLL, or when comparing the first pretreatment time points to the last pretreatment time points.

For heatmap visualization, methylation levels of DMRs were calculated as the mean methylation of all CpGs with a DMR for all samples and plotted using the heatmap function of the R package ComplexHeatmap (53). Row annotations were based on overlap with features; see Feature Annotations section above. The enrichment analysis of genes affected by DMRs was done using the online Web tool WebGestalt (54). An overrepresentation enrichment analysis (ORA) was calculated for the CLL DMRs with all DMRs as background for Panther pathways and the patient-specific DMRs by comparing recurrently hit genes among more than two patients with all DMR genes.

Single-Cell Analysis. For comparison of the RRBS single-cell experiments, only positions of the double-digest data (naïve and memory B cells) were considered that were also covered by the single-digest data (CLL).

10x scRNA-seq. The single-cell RNA was analyzed using the python toolkit "Scanpy" with default parameters for clustering and UMAP generation (55). Gene expression profiles were generated using parameters for normalized gene expression representation for dotplot and heatmap representations.

Data Accessibility

Raw methylation sequence data from patients are deposited in the database of Genotypes and Phenotypes (dbGAP) record # phs001431.v1.p to allow controlled access and maintain patient privacy.

ScRNA-seq and processed methylation data are available under Gene Expression Omnibus (GEO) accession GSE125499. Go to https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125499.

Authors' Disclosures

A. Biran reports grants from Leukemia & Lymphoma Society and Lymphoma Research Foundation during the conduct of the study. N. Purroy reports other from AstraZeneca outside the submitted work. K. Clement reports a patent for US10801070B2 issued to Harvard College General Hospital Corp., Dana-Farber Cancer Institute Inc, and Broad Institute Inc. E. Braggio reports personal fees from DASA outside the submitted work. T.D. Shanafelt reports grants from Genentech, Pharmacyclics, and AbbVie outside the submitted work, and his institution, Mayo Clinic, has a patent for green tea extract issued and based on the research of T.D. Shanafelt and N.E. Kay. N.E. Kay reports grants and personal fees from AbbVie (data and safety monitoring committee) and Pharmacyclics (advisory board); personal fees from Agios (data and safety monitoring committee), AstraZeneca (data and safety monitoring committee/advisory board), Cytomx Therapy (data and safety monitoring committee/advisory board), Dava Oncology (advisory board), Juno Therapeutics (advisory board), Oncotracker (advisory board), and Rigel (data and safety monitoring committee); and grants from Bristol-Myers Squibb, MEI Pharma, MorphoSys, TG Therapeutics, and Tolero Pharmaceuticals outside the submitted work. J.R. Brown reports personal fees from AbbVie (consulting), Acerta/AstraZeneca (consulting), BeiGene (consulting), Catapult Therapeutics (consulting), Dynamo Therapeutics (consulting), Eli Lilly and Company (consulting), Genentech/Roche (consulting), Juno/Celgene (consulting), Kite (consulting), Loxo (consulting), MEI Pharma (consulting), Nextcea (consulting), Novartis (consulting), Octapharma (consulting), Pfizer (consulting), Pharmacyclics (consulting), Rigel (consulting), Sunesis (consulting), TG Therapeutics (consulting), Janssen (honoraria), and Teva (honoraria); grants and personal fees from Gilead (consulting and research funding) and Verastem (consulting and research funding); personal fees and other from MorphoSys (consulting and data safety monitoring committee service); grants from Loxo/Lilly and Sun Pharmaceuticals; and other from Invectys (data safety monitoring committee service) outside the submitted work. S. Li reports grants from NCI during the conduct of the study. K.J. Livak reports grants from NIH/NCI during the conduct of the study. D.S. Neuberg reports grants from NIH P01 CA206978 during the conduct of the study, as well as other from Pharmacyclics (research support) and Madrigal Pharmaceuticals (stock ownership) outside the submitted work. E. Campo reports grants from Spanish Ministry of Science during the conduct of the study. A. Gnirke reports grants from NIH/NCI during the conduct of the study. C.J. Wu reports other from BioNTech (equity holder) outside the submitted work. No disclosures were reported by the other authors.

Authors' Contributions

H. Kretzmer: Conceptualization, resources, supervision, funding acquisition, investigation, visualization, writing–original draft, project administration, writing–review and editing. **A. Biran:** Conceptualization, resources, investigation, visualization, methodology, writing-original draft, writing–review and editing. **N. Purroy:** Resources, investigation, methodology. **C.K. Lemvigh:** Resources, data curation, investigation. **K. Clement:** Resources, data curation, investigation. **M. Gruber:** Resources, data curation, investigation. **H. Gu:** Resources, data curation, investigation, methodology. **L. Rassenti:** Resources, data curation, methodology. **A.W. Mohammad:** Resources, data curation, methodology. **C. Lesnick:** Resources, data

curation, methodology. **S.L. Slager:** Resources, data curation. **E. Braggio:** Resources, data curation. **T.D. Shanafelt:** Resources, data curation. **N.E. Kay:** Resources, data curation. **S.M. Fernandes:** Resources, data curation, methodology. **J.R. Brown:** Resources, methodology. **L. Wang:** Resources, methodology. **S. Li:** Resources, methodology. **K.J. Livak:** Resources, methodology. **D.S. Neuberg:** Resources, methodology. **S. Klages:** Resources, methodology, project administration. **B. Timmermann:** Supervision, funding acquisition, project administration. **T.J. Kipps:** Supervision, funding acquisition. **E. Campo:** Supervision, funding acquisition, methodology, project administration. **A. Gnirke:** Conceptualization, resources, supervision, funding acquisition, investigation, methodology, writing–original draft, project administration, writing–review and editing. **C.J. Wu:** Conceptualization, resources, supervision, funding acquisition, investigation, methodology, writing–original draft, project administration, writing–review and editing. **A. Meissner:** Conceptualization, resources, supervision, funding acquisition, visualization, methodology, writing–original draft, project administration, writing–review and editing.

Acknowledgments

The authors thank members of the Meissner lab, in particular Jocelyn Charlton, and both Erin M. Parry and Inaki Subero-Martin for critical reading of the article. The authors thank Jerome Ritz and the DFCI Pasquarello Tissue Bank in Hematologic Malignancies for the prospective collection and processing of blood samples from healthy donors. This work was supported in part by the NCI (5P01CA081534-14). A. Biran was supported by the Lymphoma Research Foundation. C.K. Lemvigh acknowledges support from the Fishman Family Fund. M. Gruber received a Marie-Curie International Outgoing Fellowship from the European Union (PIOF-2013-624924). T.D. Shanafelt and N.E. Kay were supported by the NIH (R01CA197120). E. Campo is supported by Instituto de Salud Carlos III, Madrid Spain (PMP15/00007), "La Caixa" Foundation (grant CLLEvolution-HR17-00221), Health Research 2017, and European Research Council (ERC) BCLLatlas - 810287 and is a Researcher of the "Institució Catalana de Recerca i Estudis Avançats" (ICREA) of the Generalitat de Catalunya. S.L. Slager and E. Braggio were supported by R01 CA235026. J.R. Brown is supported by NCI RO1 CA 213442 and the Rosenbach Fund for Lymphoma Research. S. Li is supported by the NCI Research Specialist Award (R50CA251956-01). The single-cell analysis was supported by a SPARC grant of the Broad Institute (to A. Gnirke). C.J. Wu acknowledges support from the CLL Global Research Foundation, NHLBI (1RO1HL103532-01), and NCI (1R01CA155010-01A1 and UG1 CA233338), and is a Scholar of the Leukemia & Lymphoma Society. A. Meissner acknowledges support from the Starr Foundation, the New York Stem Cell Foundation, and the Max Planck Society.

Received November 11, 2019; revised October 28, 2020; accepted November 18, 2020; published first December 3, 2020.

REFERENCES

1. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev 2002;16:6–21.
2. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet 2013;14:204–20.
3. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nat Rev Genet 2006;7:21–33.
4. Easwaran H, Tsai HC, Baylin SB. Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. Mol Cell 2014;54:716–27.
5. Smith ZD, Shi J, Gu H, Donaghey J, Clement K, Cacchiarelli D, et al. Epigenetic restriction of extraembryonic lineages mirrors the somatic transition to cancer. Nature 2017;549:543–7.
6. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet 2018;50:591–602.
7. Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature 2013;500:477–81.
8. Rajewsky K. Clonal selection and learning in the antibody system. Nature 1996;381:751–8.
9. Natkunam Y. The biology of the germinal center. Hematology Am Soc Hematol Educ Program 2007:210–5.
10. Kulis M, Merkel A, Heath S, Queiros AC, Schuyler RP, Castellano G, et al. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. Nat Genet 2015;47:746–56.
11. Oakes CC, Seifert M, Assenov Y, Gu L, Przekopowitz M, Ruppert AS, et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. Nat Genet 2016;48:253–64.
12. Hallek M, Cheson BD, Catovsky D, Caligaris-Cappio F, Dighiero G, Dohner H, et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. Blood 2008;111:5446–56.
13. Mowery YM, Lanasa MC. Clinical aspects of monoclonal B-cell lymphocytosis. Cancer Control 2012;19:8–17.
14. Landgren O, Albitar M, Ma W, Abbasi F, Hayes RB, Ghia P, et al. B-cell clones as early markers for chronic lymphocytic leukemia. N Engl J Med 2009;360:659–67.
15. Rawstron AC, Bennett FL, O'Connor SJ, Kwok M, Fenton JA, Plummer M, et al. Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. N Engl J Med 2008;359:575–83.
16. Zenz T, Mertens D, Kuppers R, Dohner H, Stilgenbauer S. From pathogenesis to treatment of chronic lymphocytic leukaemia. Nat Rev Cancer 2010;10:37–50.
17. Kipps TJ, Stevenson FK, Wu CJ, Croce CM, Packham G, Wierda WG, et al. Chronic lymphocytic leukaemia. Nat Rev Dis Primers 2017;3:16096.
18. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. Blood 1999;94:1848–54.
19. Kulis M, Heath S, Bibikova M, Queiros AC, Navarro A, Clot G, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet 2012;44:1236–42.
20. Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer Cell 2014;26:813–25.
21. Kretzmer H, Bernhart SH, Wang W, Haake A, Weniger MA, Bergmann AK, et al. DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. Nat Genet 2015;47:1316–25.
22. Gama-Sosa MA, Slagel VA, Trewyn RW, Oxenhandler R, Kuo KC, Gehrke CW, et al. The 5-methylcytosine content of DNA from human tumors. Nucleic Acids Res 1983;11:6883–94.
23. Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 1983;301:89–92.
24. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, et al. Increased methylation variation in epigenetic domains across cancer types. Nat Genet 2011;43:768–75.
25. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 2008;454:766–70.
26. Charlton J, Downing TL, Smith ZD, Gu H, Clement K, Pop R, et al. Global delay in nascent strand DNA methylation. Nat Struct Mol Biol 2018;25:327–32.
27. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc 2011;6:468–81.
28. Gruber M, Bozic I, Leshchiner I, Livitz D, Stevenson K, Rassenti L, et al. Growth dynamics in naturally progressing chronic lymphocytic leukaemia. Nature 2019;570:474–9.

29. Guièze R, Liu VM, Rosebrock D. Mitochondrial reprogramming underlies resistance to BCL-2 inhibition in lymphoid malignancies. Cancer Cell 2019;36:369–84.
30. Beekman R, Chapaprieta V, Russinol N, Vilarrasa-Blasi R, Verdaguer-Dot N, Martens JHA, et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. Nat Med 2018;24:868–80.
31. Bock C, Beerman I, Lien WH, Smith ZD, Gu H, Boyle P, et al. DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. Mol Cell 2012;47:633–47.
32. Juhling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S. Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Res 2016;26:256–62.
33. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011;473:43–9.
34. Landan G, Cohen NM, Mukamel Z, Bar A, Molchadsky A, Brosh R, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat Genet 2012;44:1207–14.
35. Souers AJ, Leverson JD. ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. Nat Med 2013;19:202–8.
36. Wolf C, Garding A, Filarsky K, Bahlo J, Robrecht S, Becker N, et al. NFATC1 activation by DNA hypomethylation in chronic lymphocytic leukemia correlates with clinical staging and can be inhibited by ibrutinib. Int J Cancer 2018;142:322–33.
37. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell 2013;152:714–26.
38. Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, et al. Mutations driving CLL and their evolution in progression and relapse. Nature 2015;526:525–30.
39. Nadeu F, Clot G, Delgado J, Martin-Garcia D, Baumann T, Salaverria I, et al. Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. Leukemia 2018;32:645–53.
40. Gaiti F, Chaligne R, Gu H, Brand RM, Kothen-Hill S, Schulman RC, et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. Nature 2019;569:576–80.
41. Cahill N, Bergh AC, Kanduri M, Goransson-Kultima H, Mansouri L, Isaksson A, et al. 450K-array analysis of chronic lymphocytic leukemia cells reveals global DNA methylation to be relatively stable over time and similar in resting and proliferative compartments. Leukemia 2013;27:150–8.
42. Puente XS, Bea S, Valdes-Mas R, Villamor N, Gutierrez-Abril J, Martin-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature 2015;526:519–24.
43. Bormann F, Rodriguez-Paredes M, Lasitschka F, Edelmann D, Musch T, Benner A, et al. Cell-of-origin DNA methylation signatures are maintained during colorectal carcinogenesis. Cell Rep 2018;23:3407–18.
44. Hallek M, Cheson BD, Catovsky D, Caligaris-Cappio F, Dighiero G, Dohner H, et al. iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. Blood 2018;131:2745–60.
45. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 2009;10:232.
46. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods 2016;13:229–32.
47. Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, et al. MOABS: model based analysis of bisulfite sequencing data. Genome Biol 2014;15:R38.
48. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049.
49. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2.
50. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, et al. The UCSC Genome Browser Database. Nucleic Acids Res 2003;31:51–4.
51. Brocks D, Assenov Y, Minner S, Bogatyrova O, Simon R, Koop C, et al. Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. Cell Rep 2014;8:798–806.
52. Hahne F, Ivanek R. Visualizing genomic data using Gviz and Bioconductor. Methods Mol Biol 2016;1418:335–51.
53. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016;32:2847–9.
54. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res 2017;45:W130–W7.
55. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15.

PAPER IV

Evolutionary history of transformation from chronic lymphocytic leukemia to Richter syndrome

Parry, E. M.*, Leshchiner, I.*, Guièze, R.*, Johnson, C., Tausch, E., Parikh, S. A., **Lemvigh, C. K.**, Broséus, J., Hergalant, S., Messer, C., Utro, F., Levovitz, C., Rhrissorrakrai, K., Li, L., Rosebrock, D., Yin, S., Deng, S., Slowik, K., Jacobs, R., Huang, T., Li, S., Fell, G., Redd, R., Lin, Z., Knisbacher, B. A., Livitz, D., Schneider, C., Ruthen, N., Elagina, L., Taylor-Weiner, A., Persaud, B., Martinez, A., Fernandes, S. M., Purroy, N. Z., Anandappa, A., Ma, J., Hess, J., Rassenti, L. Z., Kipps, T. J., Jain, N., Wierda, W., Cymbalista, F., Feugier, P., Kay, N. E., Livak, K. J., Danysh, B. P., Stewart, C., Neuberg, D., Davids, M. S., Brown, J. R., Parida, L., Stilgenbauer, S.**, Getz, G.**, Wu, C. J.**.

Accepted, Nature Medicine, 2022

The results from the work carried out presented in this thesis is presented in manuscript in Figure 5d-e and Extended Data Figure 9b-d, 10b.

*,** = equal contribution

1. Extended Data

Figure #	Figure title One sentence only	Filename This should be the name the file is saved as when it is uploaded to our system. Please include the file extension. i.e.: *Smith_ED_Fig1.jpg*	Figure Legend If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Fig. 1	**Clonal deconvolution process.**	Parry_ED_Fig1.tif	a, distinguishing RS from CLL clones after inferring subclonal composition of paired CLL and RS samples. b, inferring phylogenetic tree from cancer cell fraction using *PhylogicNDT*. c, sample composition d, mapping copy number variations to clones using *CopyNumber2Tree*.
Extended Data Fig. 2	**Phylogenetic reconstruction and somatic genomic alterations.**	Parry_ED_Fig2.tif	For each of the patient trios with WES data, the left panel shows the phylogenetic tree tracing the transformation history from CLL to RS. The magenta frame denotes the Richter clones. The middle top panel represents the subclonal composition inferred after clustering alterations with similar cancer cell fractions as previously reported4. The middle bottom panel indicates the timeline with RS and CLL sampling time and CLL therapeutic lines. (F, fludarabine; C, cyclophosphamide; R, rituximab; P, pentostatin; O/Ofa, ofatumumab; HDMP, high-dose methylprednisolone; A, alemtuzumab; Auto, autologous stem cell transplantation; CLB, chlorambucil; B, bendamustine; CHOP, cyclophosphamide, doxorubicin, vincristine, prednisone; ESHAP, etoposide, methylprednisolone, high-dose cytarabine, cisplatin; CHP, cyclophosphamide, doxorubicin, prednisone; Len, lenalidomide; Ob, obinutuzumab; idela; idelalisisb; D, dexamethasone; Adria, adriamycin). The right panel is composed of allelic fraction plots and allelic copy ratio plots showing clonal assignment of somatic copy number events to CLL and RS clones. Cases with whole genome doubling in Extended Data Fig. 2 and clonal unrelated cases in Extended Data Fig. 3.
Extended Data Fig. 3	**Phylogenetic reconstruction and somatic genomic alterations**	Parry_ED_Fig3.tif	For each of the patient trios with WES data, the left panel shows the phylogenetic tree tracing the transformation history from CLL to RS. The magenta frame denotes the Richter clones. The middle top panel represents the subclonal composition inferred after clustering alterations

			with similar cancer cell fractions as previously reported4. The middle bottom panel indicates the timeline with RS and CLL sampling time and CLL therapeutic lines. (F, fludarabine; C, cyclophosphamide; R, rituximab; P, pentostatin; O/Ofa, ofatumumab; HDMP, high-dose methylprednisolone; A, alemtuzumab; Auto, autologous stem cell transplantation; CLB, chlorambucil; B, bendamustine; CHOP, cyclophosphamide, doxorubicin, vincristine, prednisone; ESHAP, etoposide, methylprednisolone, high-dose cytarabine, cisplatin; CHP, cyclophosphamide, doxorubicin, prednisone; Len, lenalidomide; Ob, obinutuzumab; idela; idelalisisb; D, dexamethasone; Adria, adriamycin). The right panel is composed of allelic fraction plots and allelic copy ratio plots showing clonal assignment of somatic copy number events to CLL and RS clones. Cases with whole genome doubling in Extended Data Fig. 2 and clonal unrelated cases in Extended Data Fig. 3.
Extended Data Fig. 4	**Putative RS driver genes**	Parry_ED_Fig4.tif	**a-x,** individual protein mutation maps for selected putative Richter drivers, showing gene mutation subtype (for example, missense), position and evidence of mutational hotspots. Panels were generated by using the cBioPortal for Cancer Genomics tool.
Extended Data Fig. 5	**RS sCNAs and genomic clustering.**	Parry_ED_Fig5.tif	GISTIC2-defined recurrent copy number gains (red, left) and losses (blue, right) are visualized for focal events for RS samples (a) and RS clones (b) (RS samples with CLL events subtracted, bottom). Chromosomes are shown on the vertical axis. Green line denotes a near significant q value of 0.25 and significant events ($q<0.1$) are annotated in text along with putative driver genes contained within the peak (Supplementary Table 5) **c,** NMF clustering of RS with DLBCL (304 de novo DLBCL samples[14] shows clonal related RS clusters separately from DLBCL and closes to DLBCL from C2[14]. Clonal unrelated RS clusters across DLBCL subtypes and separate from RS. Samples were annotated for clonal relationship (related RS, gray, unrelated RS, black), cohort (DLBCL, light purple; RS, dark purple) and DLBCL clusters (C1, purple; C2, yellow, C3, pink, C4, blue, C5, green)[14]. **d,** NMF clustering of RS shows 5 distinct genomic subtypes of transformation

Extended Data Fig. 6	**Transcriptome supports distinct RS molecular subtypes.**	Parry_ED_Fig6.tif	**a,** Supervised clustering of transcriptome data from 36 RS patients by molecular subtype highlights differentially regulated genes in subtype 1 and 3 (Supplementary Table 8). Samples are annotated for cohort (Discovery, pink; Validation, yellow), clonal relationship (unrelated, black, related, white), and sample purity by WES (green gradient). **b,** Unsupervised consensus clustering of RS transcriptome data (n=36) shows 5 clusters. (Discovery, pink; Validation, yellow), RS molecular subtype (1, purple; 2, blue; 3, orange; 4, green; and 5, pink), and sample purity by WES (green gradient).**c,** 5 x 5 table showing association between molecular subtype of RS and unsupervised transcriptome clusters (2 sided Fisher's exact test, P=0.038) **d,** Kaplan-Meier curve showing OS of clonal unrelated RS compared to clonal related RS. P value is log rank (2 sided Mantel Cox).
Extended Data Fig. 7	**Phylogenetic trees showing CLL and RS clones from WGS of paired samples.**	Parry_ED_Fig7.tif	**a,** Phylogenetic tree and CCF plot for 9 patients based on WGS data showing clonal related RS (magenta box). **b,** Phylogenetic tree and CCF plot for 2 patients based on WGS demonstrating clonally unrelated RS **c,** Representative phylogenetic trees and CCF plot for 3 patients from UK cohort[8] based on WGS.
Extended Data Fig. 8	**WGS Circos plots with or without chromothripsis.**	Parry_ED_Fig8.tif	**a,** chromothripsis and kataegis in RS sample (Pt 42) with whole genome doubling. Circos plots showing structural variants (interchromosomal, blue; deletion, red; inversion, yellow; tandem duplication, green; long range, teal), allelic copy number (middle), rainfall plot with kataegis regions (red) and chromosomes (outside). Adjacent rainfall plots show kategis regions (C to G, red; C to T, yellow; C to A, teal) with corresponding allelic copy number fragmentation. **b,** Circos plots from RS WGS samples showing structural variants (interchromosomal, blue; deletion, red; inversion, yellow; tandem duplication, green and long range, teal), allelic copy number (middle), rainfall plot with kataegis regions (red) and chromosomes (outside). SVs impacting known genes and translocation partners are labeled (Supplementary Table 7k).
Extended Data Fig. 9	**Single cell processing and transcriptome analysis of RS**	Parry_ED_Fig9.tif	**a,** flow sorting strategy for RS single-cell samples. Flow sorting to separate RS and CLL cells by size for Patient 19 and Patient 41 (lymph node, LN; peripheral blood, PB; bone marrow,

	samples at single cell resolution.		BM). Flow sorting viable cells for Pt 43, Pt 4 and Pt10. Representative flow plots below demonstrate CLL and RS cells were included in sorted population. **b,** B-cell receptor (BCR) clonotypes plotted for RS and CLL clusters on UMAP visualization. **c,** Representative example from patient 10 showing CNVsingle identifies malignant B cell clusters (5 and 6) separate from immune cell clusters (0,1,2,3,4,7,9). **d,** UMI/cell and Gene/cell plots for CLL and RS single-cell clusters. RS demonstrates higher UMI/cell (P<2.2 x 10-16 see Methods, Supplementary Table 8). **e,** RNA inference of directional trajectories is shown on UMAP visualization for Pts 43 and 10. **f,** copy number variation heatmap inferred in each cluster from scRNA-seq data using our CNVSingle algorithm for Pts 43 and 10 (Methods)
Extended Data Fig. 10	**Single-cell transcriptome and copy number analysis of RS patients.**	Parry_ED_Fig10.tif	UMAP visualization of single-cells from patient 4 (left) with associated allelic copy number ratio plot inferred by CNVsingle (top right) and RS WES (bottom right). **b,** UMAP visualization of CLL and RS cells from Patient 18 (left top panel) with flow-sorting annotations (right top panel). Inferred CNAs from CNVSingle (bottom panel) are shown as heatmap with CLL (green) and RS (pink) events highlighted. **c,** UMAP visualization of CLL and RS cells from Patient 41 (left top panel) with flow-sorting annotations (right top panel). Inferred CNAs from CNVSingle (bottom panel) with CLL (green) and RS (pink) events highlighted. **d,** Plasma of patient 44 shows RS specific sCNVs on chromosome 9 and 13 leading up to RS diagnosis, which are not reflected in circulating CLL **e,** Plasma of patient 99 at the start of CLL-directed therapy (top) and just ahead of diagnosis of RS (bottom) during CLL response. **f,** Chromothripsis in post-transplant RS plasma cfDNA at time of relapse (Pt 112). **g,** Plot showing allele frequency of RS (purple) and CLL (green) mutations in RS WES (bottom) and plasma cfDNA WES (top) for patient 5 (top) and patient 44 (bottom)

2. Supplementary Information:

A. Flat Files

Item	Present?	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Yes	Supplementary Figures 1-2.pdf	Supplementary Figures 1-2
Reporting Summary	Yes	Nr-reporting-summary 100522.pdf	
Peer Review Information	No	*OFFICE USE ONLY*	

B. Additional Supplementary Files

Type	Number If there are multiple files of the same type this should be the numerical indicator. i.e. "1" for Video 1, "2" for Video 2, etc.	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. i.e.: *Smith_ Supplementary_Video_1.mov*	Legend or Descriptive Caption Describe the contents of the file
Supplementary Table	1-10	SupplementaryTables1-10.xlsx	Supplementary tables 1-10 combined into single Excel document

Evolutionary history of transformation from chronic lymphocytic leukemia to Richter syndrome

Erin M Parry MD PhD[1,2,3]*, Ignaty Leshchiner PhD[2,4]*, Romain Guièze MD PhD[1,2,5,6]*, Connor Johnson[2], Eugen Tausch MD[7], Sameer A. Parikh MD[8], Camilla Lemvigh[1,9], Julien Broséus MD PhD[10,11], Sébastien Hergalant[10], Conor Messer[2], Filippo Utro PhD[12], Chaya Levovitz MD PhD[12], Kahn Rhrissorrakrai PhD[12], Liang Li[2], Daniel Rosebrock[2], Shanye Yin PhD[1,3], Stephanie Deng[1], Kara Slowik[2], Raquel Jacobs[2], Teddy Huang[1,13], Shuqiang Li PhD[1,2,13], Geoff Fell[14], Robert Redd[14], Ziao Lin[2], Binyamin A. Knisbacher PhD[2], Dimitri Livitz[2], Christof Schneider[7], Neil Ruthen[1,13], Liudmila Elagina[2], Amaro Taylor-Weiner PhD[2], Bria Persaud[2], Aina Martinez[2], Stacey M Fernandes[1], Noelia Purroy MD PhD[1,2,3], Annabelle J Anandappa MD[1,3], Jialin Ma[2], Julian Hess[2], Laura Z Rassenti PhD[15], Thomas J Kipps MD PhD[15], Nitin Jain MD[16], William Wierda MD PhD[16], Florence Cymbalista MD PhD[17], Pierre Feugier MD PhD [10,18], Neil E. Kay MD[8], Kenneth J. Livak PhD[1,13], Brian P. Danysh PhD[2], Chip Stewart PhD[2], Donna Neuberg ScD[14], Matthew S. Davids MD[1,3], Jennifer R. Brown MD PhD[1,3], Laxmi Parida PhD[12], Stephan Stilgenbauer MD[7]**, Gad Getz[2,3,19] PhD **, Catherine J. Wu MD[1,2,3,20]**

*= These authors contributed equally
**= These authors jointly supervised this work

Affiliations:
[1] Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA.
[2] Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
[3] Harvard Medical School, Boston, MA 02215, USA.
[4] Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA.
[5] CHU de Clermont-Ferrand, F-63000, Clermont-Ferrand, France.

[6] Université Clermont Auvergne, EA7453 CHELTER, F-63000, Clermont-Ferrand, France.
[7] Division of CLL, Dept. of Internal Medicine III, Ulm University, Ulm, Germany
[8] Division of Hematology, Mayo Clinic, Rochester, MN, USA.
[9] Department of Health Technology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
[10] Inserm UMRS1256 Nutrition-Génétique et Exposition aux Risques Environnementaux (N-GERE), Université de Lorraine, Nancy, France.
[11] Université de Lorraine, CHRU-Nancy, service d'hématologie biologique, pôle laboratoires, Nancy, France.
[12] IBM Research, Yorktown Heights, New York, NY, USA.
[13] Translational Immunogenomics Lab, Dana-Farber Cancer Institute, Boston, MA, USA.
[14] Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA.
[15] Moores Cancer Center, Medicine, University of California, San Diego, La Jolla, CA, USA.
[16] Department of Leukemia, The University of Texas MD Anderson Cancer Center, TX, USA.
[16] Laboratoire d'hématologie, Hôpital Avicenne –AP-HP, INSERM U978- Université Sorbonne Paris Nord, Bobigny, France.
18 Université de Lorraine, CHRU Nancy, service d'hématologie clinique, Nancy, France
[19] Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA.
[20] Department of Medicine, Brigham and Women's Hospital, Boston, MA 02215, USA.

Correspondence should be addressed to:
Catherine J. Wu, MD
Department of Medical Oncology
Dana-Farber Cancer Institute
450 Brookline Avenue
Dana Building, Room DA-520
Boston MA 02115
E-mail: cwu@partners.org

Gad Getz, PhD
Broad Institute of MIT and Harvard,
415 Main St.
Cambridge MA 02141
E-mail: gadgetz@broadinstitute.org
Keywords: Richter syndrome, transformation, chronic lymphocytic leukemia, clonal evolution, cell free DNA, whole-exome sequencing, single cell RNA sequencing

ABSTRACT

Richter syndrome (RS) arising from chronic lymphocytic leukemia (CLL) exemplifies an aggressive malignancy that develops from an indolent neoplasm. To decipher the genetics underlying this transformation, we computationally deconvoluted admixtures of CLL and RS cells from 52 patients with RS, evaluating paired CLL-RS whole-exome sequencing data. We discovered RS-specific somatic driver mutations (including *IRF2BP2, SRSF1, B2M, DNMT3A,* and *CCND3*), recurrent copy number alterations beyond del(9p21)[*CDKN2A/B*],

recurrent whole genome duplication and chromothripsis, which were confirmed in 45 independent RS cases and in an external set of RS whole-genomes. Through unsupervised clustering, clonally-related RS was largely distinct from diffuse large B cell lymphoma (DLBCL). We distinguished pathways that were dysregulated in RS versus CLL, and detected clonal evolution of transformation at single-cell resolution, identifying intermediate cell states. Our study defines distinct molecular subtypes of RS and highlights cell-free DNA analysis as a potential tool for early diagnosis and monitoring.

Transformation to a high-grade malignancy accounts for therapeutic resistance and rapid disease progression across cancers[1-4]. Richter syndrome (RS), an aggressive lymphoma developing in patients with chronic lymphocytic leukemia (CLL), is a striking example of transformation[3]. RS is associated with median overall survival of less than one year, even in the modern era[3]. Despite advanced genomic characterization of CLL[5,6], understanding of the genetic factors driving evolution of CLL to RS remains limited. This has been partly from the difficulties in acquiring RS tissue and paired antecedent CLL cells. These challenges have precluded comparative evolutionary analysis, and limited the ability to define the molecular events underlying transformation beyond alterations in *TP53*, *NOTCH1*, *CDKN2A/B* and *MYC*[4,7-9]. While a subset of RS is believed to be clonally unrelated based on IGHV sequencing[3,9], a genome-wide analysis to exclude shared ancestry has not been yet performed. Finally, RS biopsies contain admixtures of RS and CLL cells, mandating development of tools for *in silico* deconvolution of RS and CLL genetic changes. To definitively delineate factors contributing to high-grade transformation, we analyzed exomes from matched RS and CLL DNA from 52 patients and confirmed our findings in 45 independent RS patients and 14 external RS cases[4].

RESULTS

Developing an analytic framework to discover RS drivers

We assembled a discovery cohort of 53 patients with paired CLL and RS samples of diffuse large B-cell histology (DLBCL), the most common form of transformation[3] **(Fig. 1a, Supplementary Table 1-2**). Forty-five (83%) patients received prior CLL-directed therapies, with 11 (21%) having received targeted agents. Thirty-nine (72%) patients had unmutated immunoglobulin heavy-chain variable region gene (*IGHV*) CLL (U-CLL). Whole-exome sequencing (WES) was completed for 186 DNA samples (from 53 patients) and whole-genome sequencing (WGS) for 30 samples (11 patients) (**Supplementary Table 3**). WES data from 42 patients originated from matched CLL-RS-germline samples ("trios") and 10 from paired CLL-RS ("duos"). As validation, we performed WES on 45 independent RS cases, 17 of which were duos (**Fig. 1b, Supplementary Table 1-3**).

To delineate the driver events giving rise to RS, we employed established WES analysis tools and 3 additional steps: (i) *deTiN*[10], to recover somatic mutations filtered due to tumor-in-normal contamination;[5] (ii) an optimized tool to detect somatic copy number alterations (sCNAs); (iii) *PhylogicNDT*[11] to establish the clonal composition per patient sample and infer the phylogenetic tree (**Fig. 1c***;* **Extended Data Fig. 1a-c**). RS clones were defined as new clones arising in the RS sample, not present in the antecedent CLL sample, and distinct based on somatic single nucleotide variants (sSNVs) and sCNAs. Within the CLL compartment, phylogenetic trees identified the ancestral (CLL^{ANC}), intermediate (CLL^{INT}, which expanded to give rise to RS that arose from CLL^{ANC}), and divergent (CLL^{DIV}) clones (**Fig. 1d**). RS was identified as related to CLL if at least one common CLL^{ANC} clone was shared.

These tools were applied to infer the CLL and RS clonal structure and relatedness (**Extended Data Fig. 1d, 2-3; Supplementary Figure 1-2).** We identified instances of clonal unrelatedness to the antecedent CLL (**Fig. 1e, Extended data Fig. 2**), previously classified based on *IGHV* sequencing in ~20% of RS [3,9]. Most RS were clonally related to the antecedent CLL (n=45, 87%) (**Fig. 1f**). Evolutionary relationships were secondarily determined by comparing the immunoglobulin gene sequence (**Supplementary Table 4**), largely in line with the WES-based phylogenies.

Defining the genomic landscape of RS

To determine the pure RS genomic landscape, we identified: (i) events strictly present in RS cells through computational isolation of the RS lineage separate from CLL^{DIV} (**Fig. 2a-grey outline**); and (ii) events newly acquired in RS clones (**Fig. 2a-magenta outline**). To uncover drivers of transformation, we applied *MutSig2CV*[12] and GISTIC2.0[13] (**Fig. 2b-c, Supplementary Table 5**)[14-16]

From our discovery cohort, we observed mutations in known CLL drivers (*NOTCH1, TP53, SF3B1;* **Fig. 2b**) and identified new candidate RS drivers (**Fig. 2c, Extended Data Fig. 4a-x**). These included mutations in *IRF2BP2* (n=7), which encodes an IRF2-dependent transcriptional corepressor, that is mutated in the N1 subtype of DLBCL[15] and primary mediastinal B cell lymphoma[17] (**Extended Data Fig. 4d**). Inactivating mutations in the DNA methyltransferase *DNMT3A* (8%) were previously reported as a single case in RS[8]; genetically engineered mice modelling this alteration have confirmed its CLL-driving function and impact on NOTCH signaling[18,19](**Extended Data Fig. 4e**). *B2M* loss through inactivating mutations, a mechanism of immune escape across cancers,-[14,20,21 21,22] was observed in 3 patients (**Extended Data Fig. 4f**). The detected mutations in the MYC-interacting[23] splicing factor *SRSF1* (n=4) did not co-occur with mutated-*SF3B1*, consistent with mutual exclusivity of splicing factor mutations across cancers[24] (**Extended Data Fig. 4g**). *EZH2* hotspot alterations were found in 2 clonally unrelated RS cases, as in DLBCL[14,15], while *EZH2* frameshift was seen in one case (**Extended Data Fig. 4h**).

Strikingly, we detected numerous sCNAs (**Fig. 2b-c, Extended Data Fig. 5a-b; Methods**), including *del(*17p) [*TP53*, 63%] and *del*(9p21.3) [*CDKN2A/B*, 19%], with arm-level loss of 9p in 5 additional patients. Recurrent focal events beyond common CLL drivers included *del*(15q13.1l) [*MGA* and *B2M*, 21%], amplification (amp) of chromosome 8q24 [*MYC*, 15%], *del*(7q36) [*EZH2, POT1*, *KMT2C* 11.5%], and *amp*(13q31.2) [*ERCC5*, miR-17-92 12%], which have been described in high-risk CLL[25]. Changes not previously reported in CLL or RS included *amp*(9p24) [*PDL1/L2*, 8%], *del*(16q12) (11.5%), *del*(18q22) (8%), and *amp*(7q21.2) [*CDK6,* 11.5%], *del*(1p), *amp*(11q) [*POU2AF1, SDHD*] and *amp*(1q23). Whole-genome doubling (WGD) was noted in 15% of cases

(**Extended Data Fig. 2-3**). The recurrent RS-specific gene mutations, sCNAs, and WGD were confirmed in our validation cohort (n=45) (**Fig. 2b-right bar, Supplementary Table 5, Extended Data Fig. 4a-x**) and 14 external RS genomes[4] (**Extended Data Fig. 4a-x; Supplementary Table 5)**.

Combined analysis of our discovery and validation cohorts provided power to further detect novel RS drivers, with *CCND3*, *TET2* and *BRAF* mutations and additional focal sCNAs emerging as significant (**Extended Data Fig. 4v-x, Fig. 2d**). Comparison of our 45 clonally related cases with prior large-scale CLL analyses[26] revealed predisposing lesions for RS, given their relative enrichment in $CLL^{ANC+INT}$, including mutated *TP53* and *NOTCH1*, *del*(17p) and *del*(14q32) but not *tri*(12), mut-*SF3B1,* or *del*(11q) (all $Q<0.05$, **Fig. 2e**, **Supplementary Table 5**). Compared to DLBCL[14,15], the driver distribution in these 45 cases was enriched for *TP53*, *del*(17p), *NOTCH1*, *del*(13q14.2), *del*(1p), *amp*(19p13.2), *SF3B1, EGR2,* and *GNB1* (**Fig. 2f,** all $Q<0.05$). Mut-*IRF2BP2,* -*MGA* and -*DNMT3A* frequency was higher in RS compared to 304 *de novo* DLBCLs[14].

Evaluation of the relative timing of each putative driver event in 58 related RS cases revealed *ATM* mutations, *tri(*12) or *SF3B1* mutations as already present in CLL^{ANC}*;* alterations in *TP53* (mutations and/or *del*(17p)) or *NOTCH1* and *del*(15q15.1) [*MGA*] were predominantly CLL events ($P<0.05$, **Supplementary Table 6**). In contrast, *del*(9p21), *del*(9p), *del*(9q), *del*(2q37), *amp*(1q23), and *del*(6q) were most frequently observed as new RS events ($P<0.05$, **Supplementary Table 6**); WGD was restricted to the RS clones (**Fig. 2g**, **Extended Data Fig. 2-3, Supplementary Figure 1-2**). By systematically identifying preferred genomic trajectories driving transformation, we calculated the probability for acquiring any of the RS drivers per CLL driver, via network analysis (**Fig. 2h**, **Supplementary Table 6**). Significant trajectories from CLL to RS included *NOTCH1* to *del*(1p), *NOTCH1* to *del*(14q32) and *del*(14q32) to *amp*(16q23) ($P<0.05$; $Q<0.4$).

Overall, our findings indicate mutations of *NOTCH1*, DNA damage response and the MAPK pathway as preexisting in CLL, and alterations in epigenetics, interferon/inflammatory signaling, cell cycle deregulation and immune evasion - whether by sSNVs or by sCNAs - as newly occurring at transformation (**Fig. 3a**).

Profiling RS emerging following targeted therapies

Therapies targeting BTK, BCL2 or PI3K-delta pathways have revolutionized CLL therapy, and yet have failed to prevent transformation. Since RS is a recognized mechanism of therapeutic resistance[27-29], we evaluated the 15 patients within our cohort presenting transformation to RS while receiving targeted agents. No typical resistance mutations to targeted agents were detected (*BTK*, *BCL2*), while one ibrutinib-exposed patient had *del*(8p) and one venetoclax-treated patient had *amp*(1q), both previously described sCNA drivers of resistance[30,31] (**Fig. 3b, Extended Data Fig. 2-3, Supplementary Fig. 1-2).**

To track disease tempo over time, we analyzed serial samples procured in the years prior to RS from 2 patients receiving targeted agents. Pt 26 illustrates the potential impact of *EZH2* inactivation to transformation while on venetoclax since the RS specimen carried both an inactivating *EZH2* frameshift mutation and deletion of the *EZH2* locus through *del*(7q36) (**Fig. 3c**). Pt 3 developed nodal RS that evolved from *TP53*-mutated CLL while on ibrutinib. The RS clone emerged from an aggressive CLL subclone (clone 3) marked by focal loss of the *CDKN2A/B* locus (**Fig. 3d**). Newly acquired genetic changes in the RS clone included inactivating mutations in chromatin modifiers (*CHD2, SRSF1)*, *NFKBIE,* and *del*(15q15) [*MGA* loss]. Transformation on targeted agents appears as an heterogenous process, distinct from acquired resistance in CLL and characterized by complex evolution marked by accumulation of multiple events.

RS is characterized by distinct molecular subtypes

To assess the degree of similarity between RS and DLBCL, we performed unsupervised non-negative matrix factorization (NMF) clustering on our 97 RS cases along with 304 DLBCL samples[14] based on our identified RS genetic alterations together with known DLBCL drivers. Most RS (75 of 97 cases) clustered together, largely separately from DLBCL (**Extended Data Fig. 5c**). The DLBCL cases closest to RS comprised DLBCL C2, previously reported with biallelic *TP53* inactivation, frequent *CDKN2A/B* loss and *del*(13q14)[Rb1][14]. Seven of 8 clonally unrelated RS clustered with DLBCL (Fisher's exact test, P=6.75 x 10^{-6}), with membership across the DLBCL clusters[14], highlighting unrelated RS as a diverse entity genetically similar to *de novo* DLBCL.

Analysis of the combined RS validation and discovery WES data by unbiased NMF consensus clustering defined 5 RS molecular subtypes (**Extended Data Fig. 5d**). Three (RS1, RS3, RS5) were enriched in *TP53* and/or *del*(17p) and displayed higher rates of sCNAs and genome alterations **(Fig. 4a)**. RS1 (13.4%) was marked by WGD and fractured genomes (P <0.001 **Supplementary Table 7f; Methods**), along with arm level loss of 1p and 9p and *MYC* amplification. It comprised 6 of 15 M-CLL patients, highlighting WGD as a mechanism of transformation in M-CLL (Fisher's exact test P=4.6x10^{-3}). RS3 (20.6%) was enriched for *del*(17p) and mutations in *TP53*, *NOTCH1*, and *IRF2BP2*, and frequently contained sCNAs, including *del*(14q32.11), *del*(9q), *del*(15q15.2) [*MGA*], *amp*(16q23.2) [*IRF8*] and *del*(2q37.1). RS5 (22.7%) similarly displayed high rates of *del*(17p) and *TP53* alterations and frequent sCNAs including *del*(16q12.1), *del*(1p35.2) and *amp*(7p) but lacked *NOTCH1* mutation. By contrast, subtypes RS2 and RS4 showed lower fraction of genome altered (P <0.001 [RS2 vs 1, 3, 5]; P<0.005 [RS4 vs 1, 3, 5]; **Supplementary Table 7**). RS2 (26.8%) was predominantly marked by *tri*(12) co-occurring with *SPEN*/*NOTCH1* and *KRAS* mutations. RS4 (16.5%) was marked by *SF3B1* and *EGR2* mutations on a background of *del*(13q).

Evaluation of differential expression between subtypes from matched transcriptomes from 36 RS cases identified distinct signatures defining RS1 (n=25 genes) and RS3 (n=188) (**Extended Data Fig. 6a**, **Supplementary Table 8**). RS3 displayed signatures of cell cycle and inflammatory/interferon signaling processes in line with its enrichment for *IRF2BP2* mutations (**Supplementary Table 8**). Unsupervised consensus clustering identified 5 transcriptional clusters that associated with RS molecular subtypes (Fisher's exact test, P=0.038, **Extended Data Fig. 6b-c**). RS2 and RS4 associated with improved overall survival (log-rank P = 0.0082) (**Fig. 4b)**. Clonally related cases had shorter median OS than unrelated ones (log-rank P = 0.0094) (**Extended Data Fig. 6d**).

Mutational processes underlying transformation

Evaluation of mutational profiles from the combined CLL and RS WES data revealed signatures of aging and activation-induced cytidine deaminase (AID), like previous studies[6,16,32]. We detected a dominant signature of polymerase epsilon (*POLE*) mutation in an unrelated RS case (Pt 30), with deleterious *POLE* mutation and >2,000

sSNVs (**Figure 4c**). We further analyzed WGS generated from 11 RS trios since WGS-determined phylogenetic trees improved resolution of clones; these remained concordant with the WES phylogenies, as previously reported[33] **(Extended Data Fig. 7a**). CLL and RS clones of the two unrelated RS cases did not share a distant non-coding evolutionary history, definitively establishing them as unrelated malignancies (**Extended Data Fig. 7b**). Of 14 external WGS RS cases[4], 2 were clonally unrelated cases (**Extended Data Fig. 7c, Supplementary Table 7**). Mutational analysis of the CLL clones from 10 of 11 evaluable patients revealed signatures similar to the WES analysis (**Fig. 4c**). However, the RS clones revealed expanded mutational signatures, including prior chemotherapy (SBS17b), reactive oxygen species (SBS18) and defective DNA mismatch repair (SBS44, as previously reported[4]). Kataegis was recently reported in RS[4], and we indeed identified this across 4 of 11 RS genomes with clustered AID-related mutations (**Fig. 4d-e, Supplementary Table 7**).

From the WGS samples, we observed chromothripsis as a common defining feature of TP53-altered RS genomes (**Fig. 4f, Extended Data Fig. 8a**). Chromothripsis was detected in regions likely contributing to RS pathogenesis, including 7q21 (*CDK6*) (Pt 41), 11q13 (*CCND1*) (Pt 29) and 9p24.1 (*PD-L1/L2*) (Pt 41); most regions were patient-specific (**Extended Data Fig. 8b**).

Dynamics of transformation at single-cell resolution

We identified 292 upregulated and 111 downregulated transcripts associated with transformation from analysis of bulk RNA-seq data generated from paired high-purity RS and CLL RNA (n=5; $\log_2$ fold change| > 1, adjusted $P < 0.05$) (**Fig. 5a-b**, **Supplementary Table 9**). The larger RS cells contained more expressed transcripts and at higher abundances. Their most upregulated transcripts included regulators of mitosis, spindle assembly and cytokinesis (*AURKA, AURKB, CDK1*, *CDK2)*, activation-induced cytidine deaminase (*AIDCA*), and DNA repair regulators (*BRCA1*, *XRCC2*) (**Fig. 5b, Supplementary Table 9**). Overexpression of several of these genes have been implicated in aneuploidy in cancer[34]. By contrast, CLL showed higher relative expression of BCR-signaling pathway genes.

To examine CLL transforming to RS at high-resolution, we performed scRNA-seq of flow cytometry-sorted RS diagnosis biopsy specimens from 5 additional patients that contained clonally related RS and CLL cells within the same microenvironment (**Fig 5c**; **Extended Data Fig. 9a-b**). Given the numerous RS-defining sCNAs in our WES data, we devised a tool, *CNVSingle,* to identify the expression clusters representing RS versus CLL clones based on detection of sCNA events in scRNA-seq data. *CNVSingle* is not dominated by reference, but rather utilizes segmentation of SNP-heterozygous sites to infer the sCNAs across a cluster of cells. This approach greatly improved the signal-to-noise ratio over other methods[35] and robustly detected tumor-specific sCNAs in malignant cells, with additional events in clusters of RS cells compared to those of CLL, and the absence of sCNAs in normal immune cells (**Extended Data Fig. 9c**).

Compared to CLL, the RS-identified clones across the evaluated patient samples displayed higher UMI/cell (*i.e.*, mean 9000 vs. 3193 for Pt 43, $P<10^{-14}$, Wilcoxon) and genes/cell (mean 2909 vs. 1074, $P< 10^{-14}$) (**Extended Data Fig. 9d**). Differential expression analyses of the RS versus CLL clusters showed enrichment in pathways mapping to MYC targets, cell cycle, inflammatory response and STAT signaling pathways (**Supplementary Table 9**). Directional trajectories inferred using RNA-velocity[36] supported a transition in cell states from CLL to RS (**Extended Data Fig. 9e**). The expression patterns in CLL and RS cells were sufficiently distinct that a Random Forest classifier could predict CLL vs. RS identity of individual cells (mean $F_1 \pm \sigma = 0.92 \pm 0.01$; **Methods**).

Strikingly, sCNA assignments mapped to transcriptionally identified cell populations. For example, the LN cells of Pt 43 (RS5 subtype), formed two groups of clusters consistent with CLL and RS (**Fig. 5d-top middle**). CLL cluster 2 exhibited gene expression intermediate between cluster 1 and the RS clusters, including an increase in cell cycle genes **(Fig. 5d-top right).** Accordingly, the copy number profile of cluster 1 resembled the quieter CLL^{ANC} clone in WES (green), while cluster 2 showed acquisition of sCNAs of the CLL^{INT} clone in WES (light blue) that subsequently gave rise to RS. RS clusters (3, 4) displayed additional sCNAs, consistent with the chromothripsis seen in WES analysis (*i.e.*, sCNAs on chromosome 2 and 7, 8, and 9 with regional fragmentation)

(**Extended Data Fig. 9f-top)**. Thus, intranodal cells reside on a genetic and transcriptional continuum from indolent to aggressive CLL towards RS.

Pt 10 (RS1) highlighted the rapid evolution of transformation with genomic instability in M-CLL. WES analysis established the lack of sCNAs in circulating CLL before transformation, in contrast to the abundant sCNAs and WGD in RS cells. WES of peripheral blood CLL at the time of RS diagnosis revealed new WGD, but with fewer sCNAs than the LN RS WES. By flow cytometry of the LN at RS diagnosis, both CLL and RS cells were detected (**Extended Data Fig. 9a, right**); single-cell transcriptomes yielded 3 distinct populations. Cluster 1 displayed gene counts per cell consistent with CLL while cluster 3 expressed much higher numbers of genes, in line with RS (**Extended Data Fig. 9d**); Cluster 2 showed an intermediate phenotype (**Fig. 5e top**). Cluster 1 demonstrated both *del*(17p) and WGD, matching the WES profile of the circulating CLL at the time of RS. Cluster 2 showed progressive genomic disorder followed by cluster 3, which highly resembled the CN profile of RS as per WES, with further sCNAs and fragmentation on chromosome 9. Therefore, in this case, *del*(17p) and WGD in an aggressive CLL clone preceded the RS transition, marked by subsequent global copy number shifts and chromothripsis; these observations delineate the stepwise sequence of events leading to RS.

For Pt 4, few RS cells were captured but WGD and frequent sCNAs were nonetheless observed, again demonstrating the genome disorder of the RS1 subtype (**Extended Data Fig. 10a**). For Pt 18 (RS2), clustering identified distinct early RS (clusters 3 and 4) from an RS subclone (cluster 0) containing *del*(6) that had been seen by WES (**Extended Data Fig. 10b**). Pt 41 (RS1) highlighted an intermediate cell state clearly residing within the co-existing forward scatter (FSC)-low CLL population (**Extended Data Fig. 9b,d; Supplementary Table 8**). Indeed, per *CNVsingle*, the intermediate state cells showed acquisition of early RS-specific events (i.e. *del*(3p), *del*(4) and *del*(14q)), while expression data showed enriched cell cycle genes (**Extended data Fig. 10c)**.

Early RS clones are detectable in cell-free DNA

Given the numerous RS-associated genomic features, we assessed the feasibility of non-invasive detection of RS events through cell-free DNA (cfDNA) (**Fig. 6a**). We evaluated 46 plasma samples by ultra-low pass (ULP)-WGS[37] collected from 24 patients within three years of RS diagnosis and through relapse (**Supplementary Table 10**). Samples from 17 patients were collected at the time of RS disease, including 8 at initial diagnosis. Ten were from the discovery cohort and their RS characterization served as positive confirmation for detection of RS-specific alterations. Eight of these also had simultaneous (same blood draw), or contemporaneous circulating CLL cells analyzed by WES, thus offering a controlled way to evaluate the differing contributions of nodal vs circulating disease, since the cfDNA includes DNA shed from both LN and circulating CLL cells.

RS-associated genomic features were indeed detectable in plasma. WGD was observed in the cfDNA of Patient 38 at time of RS diagnosis, matching the RS WES profile, while circulating CLL remained diploid (**Fig. 6b**). The cfDNA of Patient 44 revealed RS-associated sCNAs (*del*(9p), *amp*(13)) that were not in the CLL cells (**Extended Data Fig. 10d**). cfDNA analysis also highlighted RS emergence during therapy in a high-risk CLL patient with *del*(17p) (Patient 99). While the cfDNA profile at the start of CLL-directed therapy showed minimal sCNAs, that at time of RS diagnosis showed abundant new sCNAs including *amp*(8q24) [*MYC*] (**Extended Data Fig. 10e)**. In other patients, chromothripsis was evident in plasma cfDNA (**Fig. 6c; Extended Data Fig. 10f**). Furthermore, for 4 of 4 cases from our discovery cohort in which WES was additionally performed on cfDNA, RS-specific mutations were detected (**Fig. 6d, Extended Data Fig. 10g).**

We queried whether RS changes could be detected in cfDNA in advance of RS diagnosis. For 2 of 7 patients whose plasma was collected 1-10 months prior to RS diagnosis (**Supplementary Table 10**), we could detect RS-associated alterations in the cfDNA, during which time they were undergoing therapies for presumed refractory CLL. In Pt 5, WGD and chromothripsis (chr 6 and 16) were detected in plasma 162 days prior to diagnosis and were absent from CLL (**Fig. 6e-left**). WES of cfDNA (**Extended Data Fig. 10g)** further showed presence of RS-specific mutations. In Pt 20, cfDNA 181 days prior to RS diagnosis showed WGD and sCNAs not present in the corresponding CLL blood sample (day -179) or LN biopsy (CLL) from the prior week (**Fig. 6e-right**).

Finally, we probed the potential for cfDNA analysis to detect early RS relapse. We considered 2 patients who had achieved a state of minimal CLL involvement following allogeneic hematopoietic stem cell transplantation (HSCT), but subsequently relapsed with nodal RS. For Patient 112, cfDNA obtained immediately following HSCT lacked evidence of RS events but by days +83 and +162, new sCNAs, and thus increased fraction genome altered (FGA), were found - consistent with nodal disease emergence (**Fig. 6f**). Ultimately, biopsy-confirmed RS relapse was diagnosed on day +187. With subsequent RS response, the RS-associated cfDNA changes resolved. Pt 111 intermittently had elevated FGA in plasma following HSCT, prior to RS diagnosis, which resolved following RS therapy (**Supplementary Table 10**). Across samples, the highest levels of FGA in cfDNA were observed in RS diagnostic samples (n=8), with decreasing ratio in the preceding 1-10 months (n=7), and even lower ratio in more distant pre-diagnosis samples (>10 months; n=4). In 7 cases, FGA exceeded all values from high-risk CLL cases (n=14 samples from 5 patients) **(Fig. 6g**). Of the 8 patients with cfDNA available at the time of biopsy-proven RS diagnosis, we confidently discerned RS-specific lesions in 6 (75%) using strict criteria.

DISCUSSION

For decades, the RS diagnosis has relied on morphologic characterization of aggressive lymphoma within the context of concurrent or known history of CLL[3]. Herein, through the implementation of advanced analytic approaches that can distinguish between the RS and CLL clones, and through integration of exome, genome and transcriptome data to the largest series of paired CLL and RS specimens to date, we have defined the distinct molecular events that precede and define the RS transition.

Of the new insights gained from this study, one was the identification of novel putative driving events in RS, distinct from CLL, affecting splicing, immune evasion, epigenetics, cell cycle regulation, interferon signaling, and MYC signaling. Epigenetic remodeling has been detected in RS, impacting pathways of BCR signaling, oxidative phosphorylation, cell proliferation and MYC signalling[38]. We further identified instances of driver alterations with potential therapeutic impact, such as those affecting CDK6 or immune checkpoints. Our study

highlights major differences between RS and *de novo* DLBCL despite several shared driver events. We delineated 5 RS subtypes and confirmed these genomic patterns associated with distinct transcriptomes and outcome.

Second, RS is marked by numerous sCNAs and features of genomic instability (i.e. chromothripsis, kataegis, WGD). Near tetraploidy has been identified as an RS risk factor[39], and our detailed genomic and single-cell analysis demonstrates how this unstable state can lead to RS evolution. These features could result from mitosis defects, as suggested by RNA expression data, and WGD may confer potential therapeutic vulnerabilities[40]. We demonstrate how such instability may be used to provide an earlier and non-invasive detection of RS in cfDNA, which should be further evaluated in clinical studies as a cost-effective approach for this difficult-to-diagnose aggressive cancer[41] .

Finally, we confirmed the majority of RS is unrelated to the co-occurring CLL- a facet previously only defined based on differing IGHV clonotypes[3,9] and ultra-deep IGHV sequencing[42]. Unrelated RS has been previously associated with improved clinical outcomes, which suggests distinct disease biology[3,9]. We now demonstrate that by exome- or genome-level analysis, clonal unrelated RS is a *de novo* DLBCL, occurring as an independent lymphoma, lacking any shared distant genetic history with the co-existing CLL. These cases tended to lack *TP53* and *NOTCH1* alterations, were enriched in M-CLL, and clustered with *de novo* DLBCL separately from clonal related RS. These molecular insights may help identify RS patients with a more favorable prognosis.

Altogether, our comprehensive evolutionary tracing enables a molecular definition of transformation that can guide identification, diagnosis and prognosis of RS. Our advanced molecular framework can serve as a model for studying transformed cancers.

—

ACKNOWLEDGEMENTS

We thank Cynthia Hahn, Elisa Ten Hacken, Wandi Zhang, Satyen Gohil, and Lilian Werner for helpful discussions. We thank Candace Patterson, Sam Pollock, Oriol Olive, Conner J. Shaughnessy and Hayley Lyon for assistance in data collection and organization and Saveliy Belkin and Chet Birger for assistance in data storage. We thank Tim Lehmberg, Mikaela McDonough, Christina Galler and Mary Collins for assistance in sample collection and biobanking. We thank the patients, their families and the investigators of the clinical trials for providing samples and clinical data. This study was supported by NIH/NCI P01 CA206978 (to C.J.W. and G.G.) and NCI (1U10CA180861-01) (to C.J.W). The work is partially supported by the Broad/IBM Cancer Resistance Research Project (I.L., G.G, L.P) and a grant from Force Hemato (R.G.). Individual support was provided by DDCF Physician-Scientist Fellowship (E.M.P), Dana-Farber Flames FLAIR fellowship (E.M.P), ASCO Conquer Cancer Young Investigator Award (E.M.P), Fishman Family Fund (R.G., C.L.), EMBO fellowship ALTF 14-2018 (B.A.K), NCI Research Specialist Award R50CA251956 (S.L.), NIH/NCI R21CA267527-01 (S.Y.). Additional research support was provided by NIH R01 CA 213442 (J.R.B), Melton Family Foundation (J.R.B), NIH/NCI R01-CA236361 (T.J.K.), and the Deutsche Forschungsgemeinschaft (DFG) SFB1074 subprojects B10 (E.T.) and subprojects B1 and B2 (E.T., C.S. and St. St.)

AUTHOR CONTRIBUTIONS STATEMENT

E.M.P., I.L., R.G., C.J.,G.G., S.S., and C.J.W. designed and performed the experiments, analyzed data, and wrote the manuscript.
E.M.P, N.P-Z., A.J.A., T.H., and S.L. generated single-cell RNA seq data.
C.Lem., E.M.P., and I.L. analyzed single-cell RNA seq data along with C.Lev., F.U., and K.R.
R.G., E.T., C.Sc., M.S.D., N.J., W.W., L.R.,T.J.K, J.B., S.H., P.F., F.C., N.K., S.P., J.R.B., and S.S. provided patient samples.
K.J.L. and S.L. designed targeted NGS assay for detecting NOTCH1 3'UTR mutation. N.R. performed mapping and analysis of this NGS data.
D.R., F.U., C.Lev., and S.Y. analyzed RNAseq data. S.D. analyzed mutational data under the supervision of E.M.P. and C.J.W.
C.M., J.M., J.H., L.L. and Chi.S. analyzed WGS data.
C.J., I.L., B.P., L.E., D.R., A.T-W., A.M., D.L., E.M.P., R.G., and C.J.W. performed sequencing data analyses, assessment of the clonal architecture and inference of phylogenies under the supervision of I.L. and G.G.
E.M.P., R.G., L.R., J.B., and S.F. prepared patient samples.
I.L. developed the analytic tool for determining somatic copy number variations from FFPE samples and CNVsingle for detecting copy-number events in single-cell RNA-seq data.
D.N. performed and supervised statistical analyses. R.R. performed statistical analyses. G.F. provides graphical representation of clinical data.
B.A.K. performed immunogenetic data analyses.
B.P.D., K.R., C.Lev, and L.P. helped to design and guide the research.
B.P.D., R.A.J., and K.S. was involved in managing the project.
Z.L., I.L., and C.J. performed cell-free DNA analyses.
All authors discussed, interpreted results and approved the manuscript.

COMPETING INTERESTS STATEMENT

I.L. serves as a consultant for PACT Pharma Inc. and has stock, is on the board and serves as a consultant for ennov1 LLC., is on the board and holds equity in Nord Bio, Inc.; C.J.W., G.G, B.A.K, Z.L. are inventors on a

patent: “Compositions, panels, and methods for characterizing chronic lymphocytic leukemia” (PCT/US21/45144); C.J.W., G.G., E.M.P., I.L. and R.G. are named as inventors on U.S. provisional patent application serial number 63/244,625, filed on September 15, 2021, and U.S. provisional patent application serial number 63/291,213, filed on December 17, 2021, both of which are entitled, "Diagnosis and Prognosis of Richter's Syndrome."; G.G. is a founder, consultant and holds privately held equity in Scorpion Therapeutics, receives funding support from: IBM and Pharmacyclics, is an inventor on patent applications related to: MSMuTect, MSMutSig, MSIDetect, POLYSOLVER, and SignatureAnalyzer-GPU; R.G. receives funding support from: Abbvie, Janssen, Gilead, AstraZeneca, and Roche; M.S.D. served as a consultant for Abbvie, Adaptive Biotechnologies, Ascentage Pharma, Astra-Zeneca, BeiGene, Bristol-Myers Squibb, Eli Lilly, Genentech/Roche, Janssen, Merck, Ono Pharmaceuticals, Pharmacyclics, Research to Practice, Takeda, TG Therapeutics, Verastem, and Zentalis, receives funding support from: Ascentage Pharma, Astra-Zeneca, Genentech/Roche, MEI Pharma, Novartis, Pharmacyclics, Surface Oncology, TG Therapeutics, and Verastem, recieves funding for travel from: Abbvie, BeiGene, BioAscend, Clinical Care Options, Curio Science, Imedex, ION Solutions, Janssen, MDOutlook, PeerView, PRIME Oncology, Research to Practice, and TG Therapeutics; JRB has served as a consultant for Abbvie, Acerta/Astra-Zeneca, BeiGene, Bristol-Myers Squibb/Juno/Celgene, Catapult, Eli Lilly, Genentech/Roche, Hutchmed, Janssen, MEI Pharma, Morphosys AG, Novartis, Pfizer, Pharmacyclics, Rigel; received research funding from Gilead, Loxo/Lilly, SecuraBio, Sun, TG Therapeutics.; C.J.W. receives funding support from: Pharmacyclics; holds equity in: BioNTech, Inc; N.E.K. serves as an advisor for: Abbvie, Astra Zeneca, Beigene, Behring, Cytomx Therapy, Dava Oncology, Janssen, Juno Therapeutics, Oncotracker, Pharmacyclics and Targeted Oncology, receives funding support from: Abbvie, Acerta Pharma, Bristol Meyer Squib, Celgene, Genentech, MEI Pharma, Pharmacyclics, Sunesis, TG Therapeutics, Tolero Pharmaceuticals, participates on the DSMC (Data Safety Monitoring Committee) for: Agios Pharm, AstraZeneca, BMS-Celgene, Cytomx Therapeutics, Dren Bio, Janssen, Morpho-sys, and Rigel; T.J.K. is on the advisory board and receives funding support from: Abbvie and Roche, serves on the Speakers Bureau for Janssen, Abbvie, and Roche; E.T. serves as an advisor and on the Speakers Bureau for: Janssen, Abbvie, and Roche, receives funding support from: Abbvie and Roche; S.S. is on the advisory board and receives funding and travel support, and speaker fees from: AbbVie, AstraZeneca, BeiGene, BMS, Celgene, Gilead, GSK, Hoffmann-La Roche, Janssen, Novartis, and Sunesis; N.J. receives research funding from: Pharmacyclics, AbbVie, Genentech, AstraZeneca, BMS, Pfizer, Servier, ADC Therapeutics, Cellectis, Precision BioSciences, Adaptive Biotechnologies, Incyte, Aprea Therapeutics, Fate Therapeutics, Mingsight, Takeda, Medisix, Loxo Oncology, Novalgen and serves on Advisory Board /Honoraria: Pharmacyclics, Janssen, AbbVie, Genentech, AstraZeneca, BMS, Adaptive Biotechnologies, Precision BioSciences, Servier, Beigene, Cellectis, TG Therapeutics, ADC Therapeutics, MEI Pharma; W.G.W. reports funding from GSK/Novartis, Abbvie, Genentech, Pharmacyclics LLC, AstraZeneca/Acerta Pharma, Gilead Sciences, Juno Therapeutics, KITE Pharma, Sunesis, Miragen, Oncternal Therapeutics, Inc., Cyclacel, Loxo Oncology, Inc., Janssen, Xencor. S.A.P. has received research funding to the institution from Pharmacyclics, Janssen, AstraZeneca, TG Therapeutics, Merck, AbbVie, and Ascentage Pharma for clinical studies in which S.A.P. is a principal investigator. S.A.P has received honoraria for participation in consulting activities/advisory board meetings for Pharmacyclics, Merck, AstraZeneca, Genentech, GlaxoSmithKline, Adaptive Biotechnologies, Amgen, and AbbVie (no personal compensation). K.J.L. holds equity in Standard BioTools Inc. D.N. has stock ownership in Madrigal Pharmaceuticals. C.S. serves on speaker’s bureau for Astra Zeneca and AbbVie. D.L. holds stock in and consults for ennov1. N.P is currently an employee at Bristol Meyers Squibb.

J.B. none; K.S. none; R.J. none; C.J. none; C.M. none; D.R. none; S.L. none; D.L. none; J.L. none; J.H. none; C.St. none; L.Z.R. none; C.Sc. none; S.Y. none; G.F. none; N.R. none; C.Lem. none; F.C. none; F.U. none; K.R. none; C.Lev. none; L.P. none; A.T.-W. none, T.H. none, R.R. none, L.E. none, B.P. none, J.M. none, S.F. none, A.J.A. none, S.H. none, S.D. none, L.L. none, P.F. none, and B.P.D. none.

Figure 1

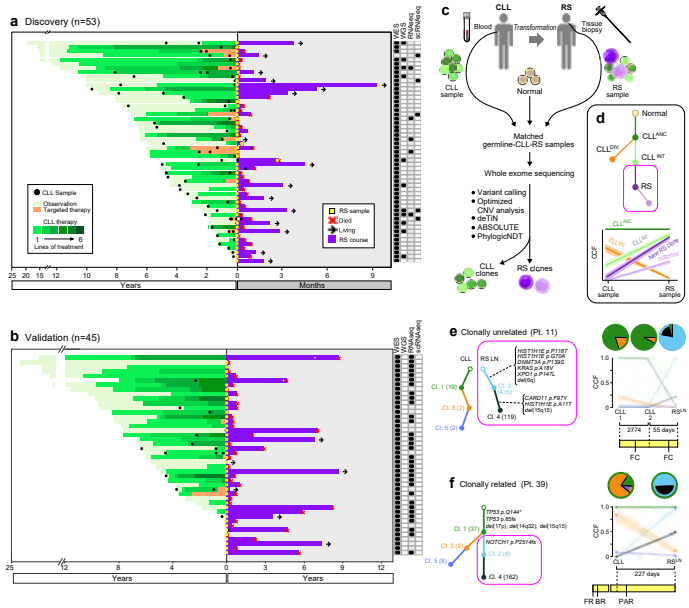

a Discovery (n=53)
b Validation (n=45)
c
CLL
RS
Blood
Transformation
Tissue biopsy
CLL sample
Normal
RS sample
Matched germline-CLL-RS samples
Whole exome sequencing
Variant calling
Optimized CNV analysis
deTiN
ABSOLUTE
PhylogicNDT
CLL clones
RS clones
d
Normal
CLL^ANC
CLL^DIV
CLL^INT
RS
CCF
CLL sample
RS sample
e Clonally unrelated (Pt. 11)
f Clonally related (Pt. 39)
Years
Months
CLL Sample
Observation
Targeted therapy
CLL therapy
Lines of treatment
RS sample
Died
Living
RS course
FC
FR
BR
PAR
55 days
227 days

Figure 2

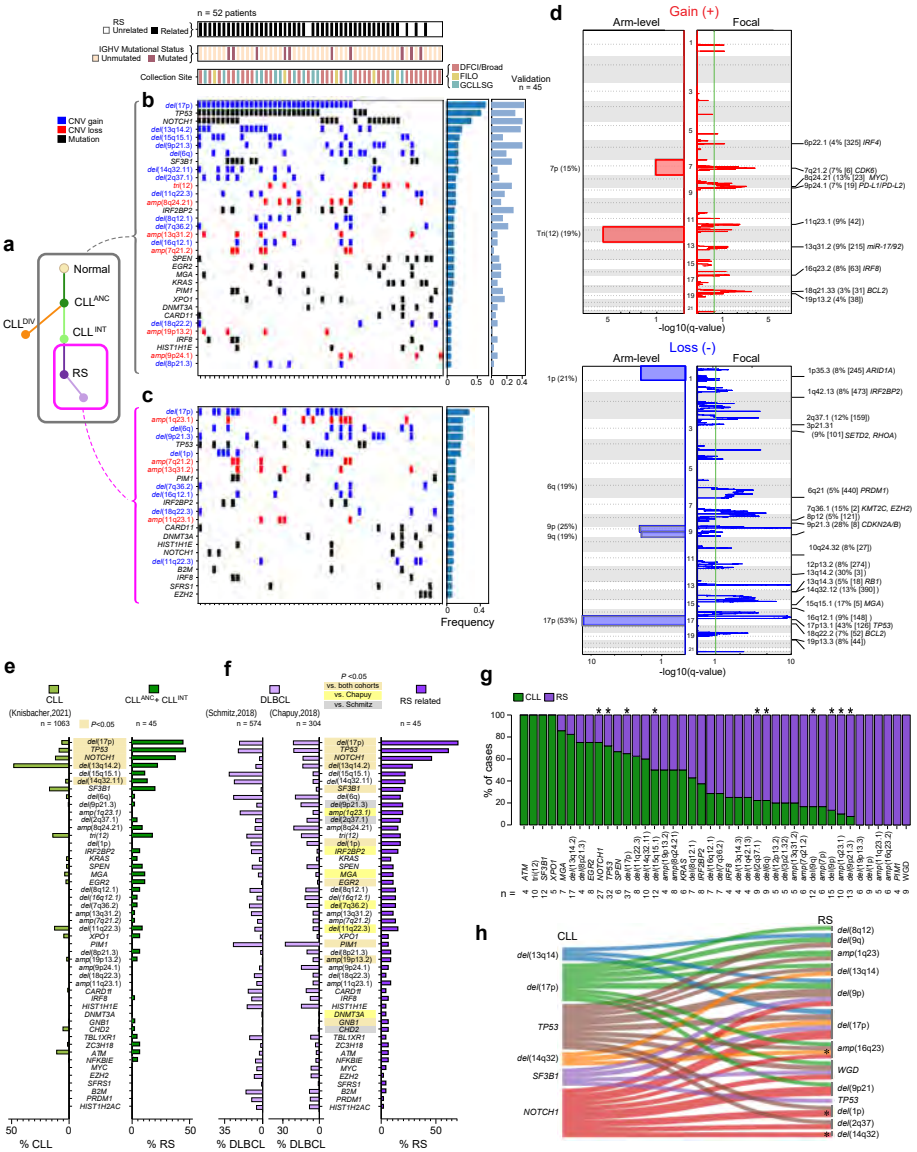

a
Normal
CLL ANC
CLL INT
CLL DIV
RS
b
c
CNV gain
CNV loss
Mutation
n = 52 patients
RS
Unrelated
Related
IGHV Mutational Status
Unmutated
Mutated
Collection Site
DFCI/Broad
FILO
GCLLSG
Validation
n = 45
Frequency
d
Gain (+)
Arm-level
Focal
-log10(q-value)
Loss (-)
e
CLL
(Knisbacher,2021)
n = 1063
CLL ANC + CLL INT
n = 45
P<0.05
% CLL
% RS
f
DLBCL
(Schmitz,2018)
n = 574
(Chapuy,2018)
n = 304
RS related
n = 45
P <0.05
vs. both cohorts
vs. Chapuy
vs. Schmitz
% DLBCL
g
CLL
RS
% of cases
h
CLL
RS

Figure 3

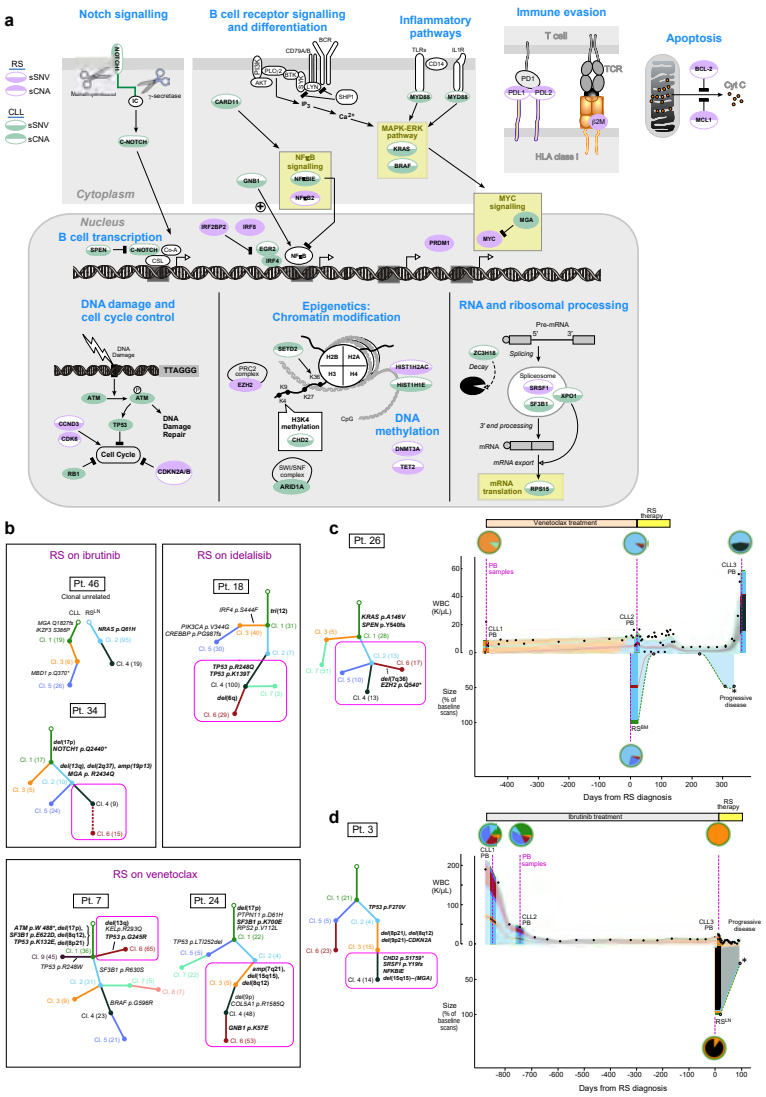

a
Notch signalling
B cell receptor signalling and differentiation
Inflammatory pathways
Immune evasion
Apoptosis
RS
sSNV
sCNA
CLL
sSNV
sCNA
Cytoplasm
Nucleus
B cell transcription
DNA damage and cell cycle control
Epigenetics: Chromatin modification
RNA and ribosomal processing
DNA methylation
b
RS on ibrutinib
Pt. 46
Clonal unrelated
Pt. 34
RS on idelalisib
Pt. 18
RS on venetoclax
Pt. 7
Pt. 24
c
Pt. 26
Venetoclax treatment
RS therapy
PB samples
WBC (K/μL)
Size (% of baseline scans)
Days from RS diagnosis
Progressive disease
d
Pt. 3
Brutinib treatment
RS therapy
PB samples
WBC (K/μL)
Size (% of baseline scans)
Days from RS diagnosis
Progressive disease

Figure 4

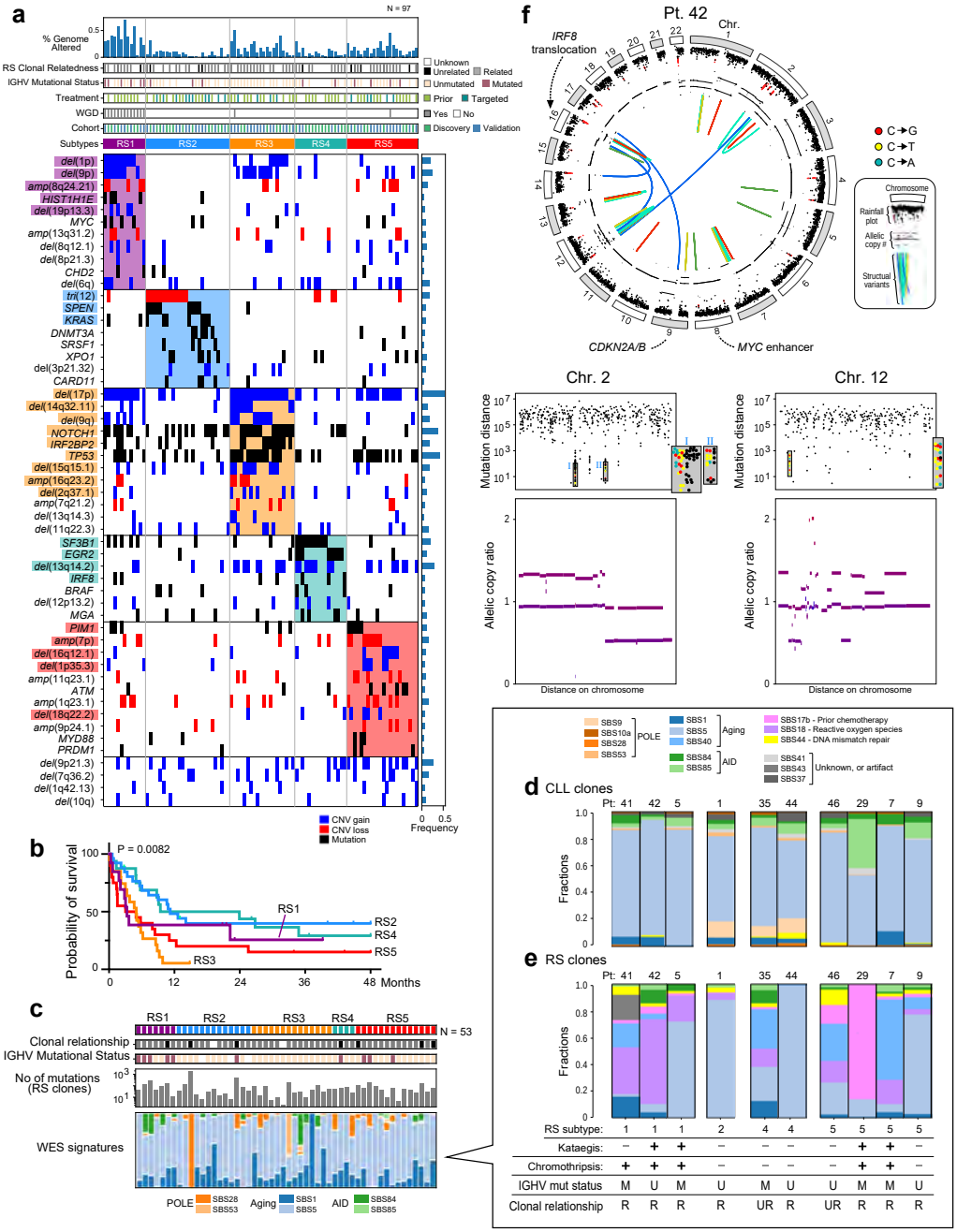

a
N = 97
% Genome Altered
RS Clonal Relatedness
IGHV Mutational Status
Treatment
WGD
Cohort
Subtypes
RS1
RS2
RS3
RS4
RS5
Unknown
Unrelated
Related
Unmutated
Mutated
Prior
Targeted
Yes
No
Discovery
Validation
del(1p)
del(9p)
amp(8q24.21)
HIST1H1E
del(19p13.3)
MYC
amp(13q31.2)
del(8q12.1)
del(8p21.3)
CHD2
del(6q)
tri(12)
SPEN
KRAS
DNMT3A
SRSF1
XPO1
del(3p21.32)
CARD11
del(17p)
del(14q32.11)
del(9q)
NOTCH1
IRF2BP2
TP53
del(15q15.1)
amp(16q23.2)
del(2q37.1)
amp(7q21.2)
del(13q14.3)
del(11q22.3)
SF3B1
EGR2
del(13q14.2)
IRF8
BRAF
del(12p13.2)
MGA
PIM1
amp(7p)
del(16q12.1)
del(1p35.3)
amp(11q23.1)
ATM
amp(1q23.1)
del(18q22.2)
amp(9p24.1)
MYD88
PRDM1
del(9p21.3)
del(7q36.2)
del(1q42.13)
del(10q)
CNV gain
CNV loss
Mutation
Frequency
b
P = 0.0082
Probability of survival
RS1
RS2
RS4
RS5
RS3
Months
c
Clonal relationship
IGHV Mutational Status
No of mutations (RS clones)
WES signatures
N = 53
POLE
Aging
AID
f
Pt. 42
Chr.
IRF8 translocation
C→G
C→T
C→A
Chromosome
Rainfall plot
Allelic copy #
Structural variants
CDKN2A/B
MYC enhancer
Chr. 2
Chr. 12
Mutation distance
Allelic copy ratio
Distance on chromosome
SBS9
SBS10a
SBS28
SBS53
SBS1
SBS5
SBS40
SBS84
SBS85
SBS17b - Prior chemotherapy
SBS18 - Reactive oxygen species
SBS44 - DNA mismatch repair
SBS41
SBS43
SBS37
Unknown, or artifact
d CLL clones
e RS clones
Pt: 41 42 5 1 35 44 46 29 7 9
Fractions
RS subtype: 1 1 1 2 4 4 5 5 5 5
Kataegis: – + + – – – – + + –
Chromothripsis: + + + – – – – + + –
IGHV mut status M U M U M U U M M U
Clonal relationship R R R R UR R UR R R R

Figure 5

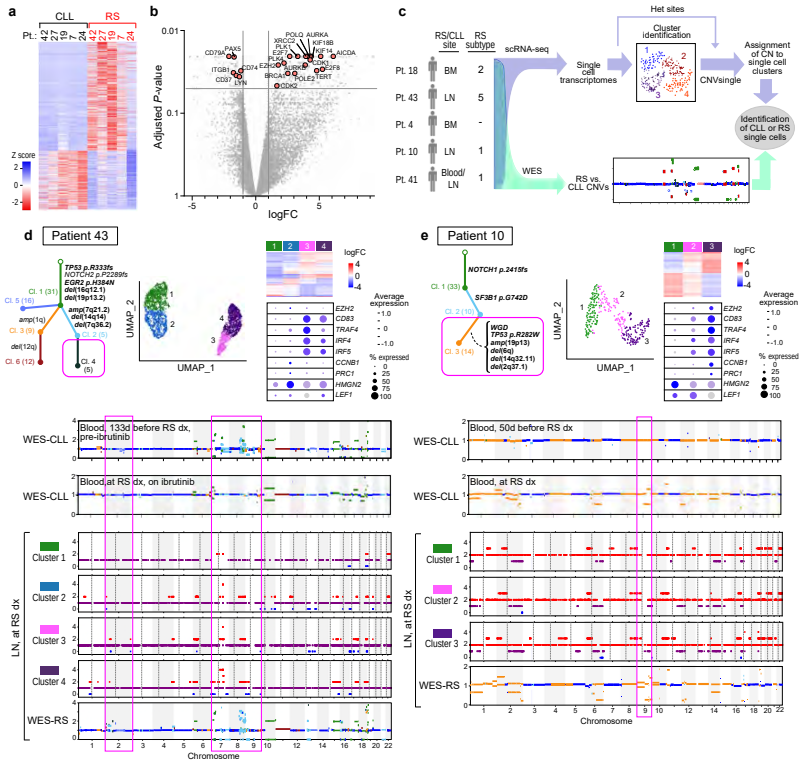

a
CLL
RS
Pt.:
Z score
b
Adjusted P-value
logFC
c
RS/CLL site
RS subtype
scRNA-seq
Pt. 18
Pt. 43
Pt. 4
Pt. 10
Pt. 41
BM
LN
Blood/ LN
Single cell transcriptomes
Het sites
Cluster identification
CNVsingle
Assignment of CN to single cell clusters
Identification of CLL or RS single cells
WES
RS vs. CLL CNVs
d
Patient 43
UMAP_1
UMAP_2
logFC
Average expression
% expressed
EZH2
CD83
TRAF4
IRF4
IRF5
CCNB1
PRC1
HMGN2
LEF1
e
Patient 10
WES-CLL
Blood, 133d before RS dx, pre-ibrutinib
Blood at RS dx, on ibrutinib
Blood, 50d before RS dx
Blood, at RS dx
Cluster 1
Cluster 2
Cluster 3
Cluster 4
WES-RS
LN, at RS dx
Chromosome

Figure 6

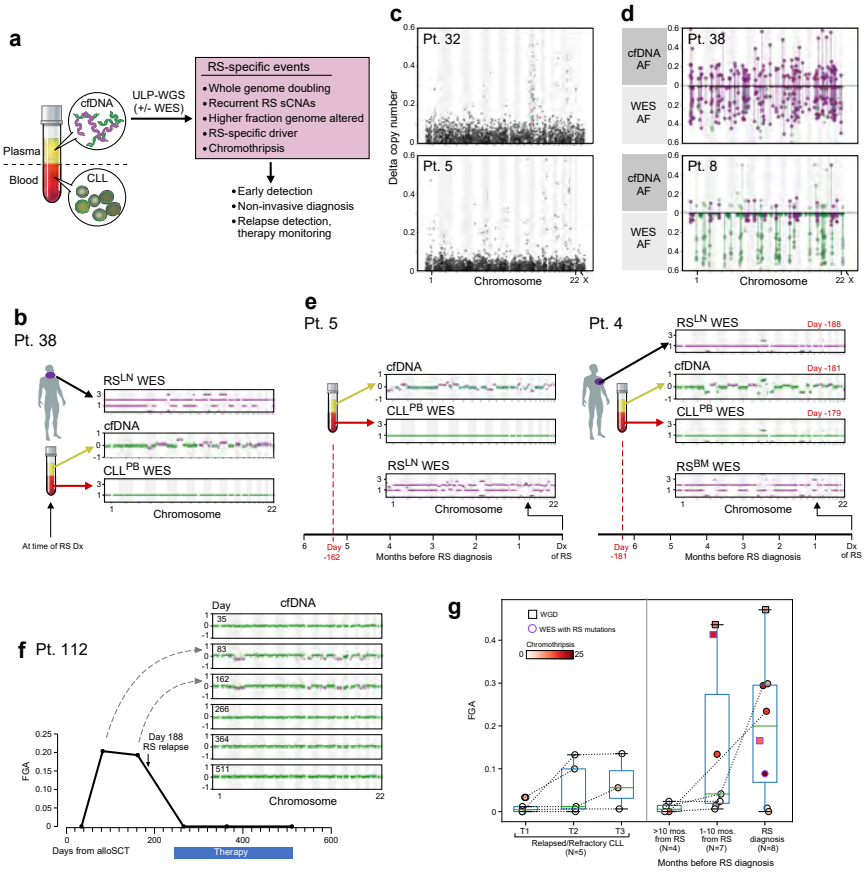

a
cfDNA
ULP-WGS
(+/- WES)
Plasma
Blood
CLL
RS-specific events
• Whole genome doubling
• Recurrent RS sCNAs
• Higher fraction genome altered
• RS-specific driver
• Chromothripsis
• Early detection
• Non-invasive diagnosis
• Relapse detection, therapy monitoring
c
Pt. 32
Pt. 5
Delta copy number
Chromosome
d
Pt. 38
Pt. 8
cfDNA AF
WES AF
Chromosome
b
Pt. 38
RS^LN WES
cfDNA
CLL^PB WES
Chromosome
At time of RS Dx
e
Pt. 5
Pt. 4
RS^LN WES
cfDNA
CLL^PB WES
RS^BM WES
Day -188
Day -181
Day -179
Day -162
Months before RS diagnosis
Dx of RS
f
Pt. 112
Day
cfDNA
FGA
Day 188
RS relapse
Days from alloSCT
Therapy
g
WGD
WES with RS mutations
Chromothripsis
FGA
T1
T2
T3
Relapsed/Refractory CLL (N=5)
>10 mos. from RS (N=4)
1-10 mos. from RS (N=7)
RS diagnosis (N=8)
Months before RS diagnosis

FIGURE LEGENDS

Fig. 1. Developing an analytic framework for detecting Richter Syndrome (RS)-specific clones. a, Disease course of 53 RS patients from CLL diagnosis in relationship to lines of therapy and sample collection. **b**, Disease course of 44 of 45 RS validation cohort patients from CLL diagnosis in relationship to lines of therapy and sample collection (1 patient with missing data). **c**, Computational schema for deciphering CLL and RS clones within RS biopsy samples. **d,** Inset shows labeled sample phylogenetic tree with associated sample cancer cell fraction (CCF) plot. Phylogenetic trees with CCF clustering, clonal abundance and associated patient disease course in representative clonally unrelated (**e**) and related (**f**) cases.

Fig. 2. The landscape of putative driver mutations in RS. a, Phylogenetic tree schema demonstrating clones comprising RS history (gray box) and RS-specific clones (magenta box) (ANC, ancestor clone; INT, CLL intermediate clone; DIV, CLL divergent clone; RS, RS clone). **b-c**,-Somatic mutation information across the putative driver genes and recurrent somatic copy number alterations (rows) for 52 RS patients (columns) that underwent WES, ranked by frequency (right) for both (**b**) RS history, alterations detected in RS cells and (**c**) RS clones, alterations acquired at transformation. Samples were annotated for sequencing site (DFCI/Broad, red; German CLL Study Group (GCLLSG), blue; French Innovative Leukemia Organization (FILO), yellow), IGHV status (maroon, mutated; peach unmutated), and clonal relationship (black, related; white, unrelated). Light blue frequency bars adjacent to RS history represent frequency in validation cohort of each alteration (n=45) **d,** GISTIC2.0 plots showing arm level (right panel) and focal (left panel) amplifications (red, top) and deletions (blue, bottom) for RS samples in the combined discovery and validation cohorts (n=97). Discovery cohort GISTIC2.0 plots are located in Extended Data Figure 4. **e,** Frequencies of somatic alterations in CLL clones from related RS cases (n=45, dark green bars) compared to CLL driver frequencies[16] using 2 sided exact binomial test with Benjamini-Hochberg multiple hypothesis testing correction **f**, RS somatic alteration frequencies (dark purple) compared to DLBCL event frequencies (light purple) from DLBCL cohorts[15,17] using 2 sided exact binomial test with Benjamini-Hochberg multiple test correction. **g,** Proportion in which a recurrent driver is found as present in $CLL^{ANC+INT}$ (green) or acquired in RS (purple) across 58 related cases (only drivers affecting at least 4 patients are shown) (**Supplementary Table 6**). * denotes $P<0.05$ (McNemar test, one-sided). **h,** Sankey plot showing trajectories from CLL driver to acquired RS driver. Only driver pairs with at least 4 co-occurrences across the cohort are displayed and tested for statistical significance (**Supplementary Table 6**). * denotes $P<0.05$ (Fisher's exact test, two-sided) and $Q < 0.4$.

Fig. 3. Tracing evolution of RS on targeted agent therapy
a, Pathways altered in CLL transformation to RS include CLL phase alterations (light green) and new drivers identified in RS (light purple). sSNV (top shading) and sCNA (bottom shading). **b,** Trees depicting clonal evolution of CLL to RS in seven select patients who developed RS on novel agents. Recurrent RS drivers indicated in bold. **c-d,** Evolution of RS from CLL showing clonal composition and absolute tumor burden over time based on serial sampling for two patients. Left panel - a phylogenetic tree with associated driver events. (Magenta square, RS clones). Right panel - relative abundance of CLL in peripheral blood by white blood cell count (1000 cells/microliter) (top) and relative abundance of RS in bottom plot (by PET/CT scan tumor metrics) with clonal evolution dynamics. Pie charts reflect composition of each sampling timepoint. (pink dotted line, sampling time; Top bar, treatment history; PB, peripheral blood; BM, bone marrow)

Fig. 4. Molecular mechanisms underlying transformation to RS. a, genomic classification of RS. For 97 patients (columns), 5 patterns of RS identified by consensus NMF clustering are depicted with respective somatic mutations and copy number alterations (rows). Samples are annotated for prior treatments (chemoimmunotherapy, light green; targeted agent, dark green; no prior therapy, white); *IGHV* status (mutated, brown; unmutated, beige; white, not determined); clonal relatedness (related, gray; unrelated, black; unknown by WES, white); and the presence of whole genome doubling (gray). Fraction genome altered per sample is shown (top). Event frequencies are indicated as blue bars on the right side for each alteration. Genes that met significance for association with a cluster by Fisher's exact Test (**Supplementary Table 7**) are highlighted by cluster association (subtype 1, purple; subtype 2, blue; subtype 3, orange; subtype 4, green; subtype 5, red). **b,** Overall survival according to the RS genomic pattern. Kaplan-Meier curves for each subtype according to color legends. *P* value is from log-rank (Mantel Cox) testing. **c,** WES signatures for RS samples from discovery cohort (n=52). **d-e,** WGS signatures for CLL (**c**) and RS (**d**) clones in 10 evaluable patients. *IGHV* status (mutated, M; unmutated, UM) and clonal relationship (R, related; UR, unrelated) is indicated at bottom. **f,** Chromothripsis and kataegis in RS sample (Pt 42) with whole genome doubling**.** Circos plots showing structural variants (interchromosomal, blue; deletion, red; inversion, yellow; tandem duplication, green; long range, teal), allelic copy number (middle), rainfall plot with kataegis regions (red) and chromosomes (outside). Adjacent rainfall plots show kategis regions (C to G, red; C to T, yellow; C to A, teal) with corresponding allelic copy number ratio plot showing corresponding fragmentation.

Fig. 5. Transformation to RS at single-cell resolution. a, Heatmap of differentially expressed transcripts with FDR<0.1 and absolute $\log_2$ fold change > 1 in analysis between paired RS and CLL samples from Pts 27, 7, 42, 20 and 24. **b,** Volcano plot of transcript expression changes in RS compared to CLL. Differentially expressed genes were assessed using limma-voom (**Methods**) in paired mode using sample read counts. logFC denotes log2FC and *P*-values are adjusted for multiple comparisons. Pink dots denote select relevant transcripts. **c,** Schema for assignment of copy number changes to single-cells to enable identification of CLL vs RS cells. **d-e,** Single-cell data shows transcriptional differences between RS and CLL from Pt 43 in **d**, and Pt 10 in **e**, and highlights intermediate states. Phylogenetic tree showing clonal structure of RS from WES data (top left) and UMAP visualization of RS and CLL single-cells (top middle). Heatmap representation of differential regulated genes between clusters (top right) and dot plot showing cluster expression of representative genes in dysregulated pathways (**Supplementary Table 9**) (purple shading, relative expression; dot size, percent of single-cell cells expressing transcript). Inferred allelic copy number from *CNVsingle* for each single-cell cluster (bottom) depicted adjacent to WES allelic copy number plots color-coded to show copy number events assigned to CLL and RS clones (**Methods**).

Figure 6. cfDNA isolated from plasma of RS patients shows evidence of transformation. a, Schema showing how RS specific DNA events can be identified separately from cell-free DNA and different from circulating CLL cells. **b**, cfDNA in RS Pt 38 shows WGD of clonally unrelated RS, which is not seen in circulating CLL disease at time of diagnosis. **c**, Chromothripsis is observed in cfDNA of RS patients, as demonstrated by plotting the difference between copy number state changes across the genome (Pt 32 top, Pt 5 bottom) **d**, Allele frequencies for RS (purple) and CLL (green) mutations found in RS WES sample (bottom) and RS plasma sample cfDNA WES (top) for patient 38 (top panel) and patient 8 (bottom panel). **e**, Plasma from patients shows early detection of RS. Pt 5 (top) shows RS-related WGD and chromothripsis fragmentation 162 days prior to RS diagnosis, which is not seen in corresponding co-sampled CLL cells. Plasma from Pt 20 (bottom panel) examined 181 days prior to RS shows RS-related WGD and sSCNVs which are not seen in co-sampled CLL or in lymph node biopsy taken from prior week. **f**, sCNAs become detectable prior to post-transplant relapse in Pt 112, as seen by plot of fraction genome altered and corresponding cfDNA samples showing emergence of new sCNVs despite continued remission of circulating and marrow CLL. **g**, Metrics of RS in cfDNA are plotted for RS samples leading up to diagnosis. Y axis is fragment genome altered, color scale shows presence of chromothripsis, square represents whole genome doubled (WGD) sample and purple outline indicates samples for which RS mutations were detected on WES of cfDNA. CLL samples at left of figure depict 13 samples from 5 relapsed/refractory CLL patients. RS samples (right) show 19 samples divided by time leading up to RS in 14 RS patients. Number of samples per each category is indicated on the figure by N. Dashed lines denote serial samples from same patients. Box plots show median values as horizontal line and whiskers showing maximum and minimum values with boundaries of box showing the interquartile range.

REFERENCES

1. Offin, M., *et al.* Concurrent RB1 and TP53 Alterations Define a Subset of EGFR-Mutant Lung Cancers at risk for Histologic Transformation and Inferior Clinical Outcomes. *J Thorac Oncol* **14**, 1784-1793 (2019).
2. Volta, A.D., *et al.* Transformation of Prostate Adenocarcinoma Into Small-Cell Neuroendocrine Cancer Under Androgen Deprivation Therapy: Much Is Achieved But More Information Is Needed. *J Clin Oncol* **37**, 350-351 (2019).
3. Parikh, S.A., Kay, N.E. & Shanafelt, T.D. How we treat Richter syndrome. *Blood* **123**, 1647-1657 (2014).
4. Landau, D.A., *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525-530 (2015).
5. Puente, X.S., *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519-524 (2015).
6. Chigrinova, E., *et al.* Two main genetic pathways lead to the transformation of chronic lymphocytic leukemia to Richter syndrome. *Blood* **122**, 2673-2682 (2013).
7. Fabbri, G., *et al.* Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. *J Exp Med* **210**, 2273-2288 (2013).
8. Klintman, J., *et al.* Genomic and transcriptomic correlates of Richter transformation in chronic lymphocytic leukemia. *Blood* **137**, 2800-2816 (2021).
9. Rossi, D., *et al.* The genetics of Richter syndrome reveals disease heterogeneity and predicts survival after transformation. *Blood* **117**, 3391-3401 (2011).

10. Taylor-Weiner, A., *et al.* DeTiN: overcoming tumor-in-normal contamination. *Nat Methods* **15**, 531-534 (2018).
11. Leshchiner, I., *et al.* Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. *bioRxiv*, 508127 (2018).
12. Lawrence, M.S., *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
13. Mermel, C.H., *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
14. Chapuy, B., *et al.* Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat Med* **24**, 679-690 (2018).
15. Schmitz, R., *et al.* Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N Engl J Med* **378**, 1396-1407 (2018).
16. Knisbacher, B.A., *et al.* Molecular map of chronic lymphocytic leukemia and its impact on outcome. *Nat Genet* (2022).
17. Chapuy, B., *et al.* Genomic analyses of PMBL reveal new drivers and mechanisms of sensitivity to PD-1 blockade. *Blood* **134**, 2369-2382 (2019).
18. Biran, A., *et al.* Activation of Notch and Myc Signaling via B-cell-Restricted Depletion of Dnmt3a Generates a Consistent Murine Model of Chronic Lymphocytic Leukemia. *Cancer Res* **81**, 6117-6130 (2021).
19. Mahajan, V.S., *et al.* B1a and B2 cells are characterized by distinct CpG modification states at DNMT3A-maintained enhancers. *Nat Commun* **12**, 2208 (2021).
20. Challa-Malladi, M., *et al.* Combined genetic inactivation of beta2-Microglobulin and CD58 reveals frequent escape from immune recognition in diffuse large B cell lymphoma. *Cancer Cell* **20**, 728-740 (2011).
21. Sade-Feldman, M., *et al.* Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat Commun* **8**, 1136 (2017).
22. Gettinger, S., *et al.* Impaired HLA Class I Antigen Processing and Presentation as a Mechanism of Acquired Resistance to Immune Checkpoint Inhibitors in Lung Cancer. *Cancer Discov* **7**, 1420-1435 (2017).
23. Singh, K., *et al.* c-MYC regulates mRNA translation efficiency and start-site selection in lymphoma. *J Exp Med* **216**, 1509-1524 (2019).
24. Lee, S.C., *et al.* Synthetic Lethal and Convergent Biological Effects of Cancer-Associated Spliceosomal Gene Mutations. *Cancer Cell* **34**, 225-241 e228 (2018).
25. Edelmann, J., *et al.* Genomic alterations in high-risk chronic lymphocytic leukemia frequently affect cell cycle key regulators and NOTCH1-regulated transcription. *Haematologica* **105**, 1379-1390 (2020).
26. Knisbacher, B.A., *et al.* The CLL-1100 Project: Towards Complete Genomic Characterization and Improved Prognostics for CLL,. *Blood* **136**, 3-4 (2020).
27. Anderson, M.A., *et al.* Clinicopathological features and outcomes of progression of CLL on the BCL2 inhibitor venetoclax. *Blood* **129**, 3362-3370 (2017).
28. Jain, P., *et al.* Long-term outcomes for patients with chronic lymphocytic leukemia who discontinue ibrutinib. *Cancer* **123**, 2268-2273 (2017).
29. Maddocks, K.J., *et al.* Etiology of Ibrutinib Therapy Discontinuation and Outcomes in Patients With Chronic Lymphocytic Leukemia. *JAMA Oncol* **1**, 80-87 (2015).
30. Burger, J.A., *et al.* Clonal evolution in patients with chronic lymphocytic leukaemia developing resistance to BTK inhibition. *Nat Commun* **7**, 11589 (2016).
31. Guieze, R., *et al.* Mitochondrial Reprogramming Underlies Resistance to BCL-2 Inhibition in Lymphoid Malignancies. *Cancer Cell* **36**, 369-384 e313 (2019).
32. Kasar, S., *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun* **6**, 8866 (2015).
33. Gruber, M., *et al.* Growth dynamics in naturally progressing chronic lymphocytic leukaemia. *Nature* **570**, 474-479 (2019).

34. Zhang, N., *et al.* Overexpression of Separase induces aneuploidy and mammary tumorigenesis. *Proc Natl Acad Sci U S A* **105**, 13033-13038 (2008).
35. Patel, A.P., *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).
36. Bergen, V., Lange, M., Peidli, S., Wolf, F.A. & Theis, F.J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408-1414 (2020).
37. Adalsteinsson, V.A., *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, 1324 (2017).
38. Nadeu, F., *et al.* Detection of early seeding of Richter transformation in chronic lymphocytic leukemia. *Nat Med* **28**, 1662-1671 (2022).
39. Miller, C.R., *et al.* Near-tetraploidy is associated with Richter transformation in chronic lymphocytic leukemia patients receiving ibrutinib. *Blood Adv* **1**, 1584-1588 (2017).
40. Quinton, R.J., *et al.* Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature* **590**, 492-497 (2021).
41. Soilleux, E.J., *et al.* Diagnostic dilemmas of high-grade transformation (Richter's syndrome) of chronic lymphocytic leukaemia: results of the phase II National Cancer Research Institute CHOP-OR clinical trial specialist haemato-pathology central review. *Histopathology* **69**, 1066-1076 (2016).
42. Favini, C., *et al.* Clonally unrelated Richter syndrome are truly de novo diffuse large B-cell lymphomas with a mutational profile reminiscent of clonally related Richter syndrome. *Br J Haematol* (2022).

ONLINE METHODS

Patient sample collection and processing

CLL, RS, and normal germline (i.e. non-tumor) samples were collected from patients following written informed consent through sample collection protocols or from clinical trial NCT03619512 from the French Innovative Leukemia Organization (FILO) with the approval of the following institutional review boards (IRBs): Dana-Farber Cancer Institute IRB, University of California San Diego IRB, Mayo Clinic IRB, MD Anderson Cancer Center IRB, Ethics Committee of Ulm University, Ulm Germany or University Hospital of Nancy with the approval of the Comité de Protection des personnes (CPP) Ouest IV (Nantes, France). All biospecimen collection protocols were conducted in accordance with the principles of the Declaration of Helsinki and with the approval of the Institutional Review Boards of the respective institutions. Patient and sample characteristics are provided (**Supplementary Tables 1-2**). Sex was self-report.

RS samples. RS samples were collected from BM, LN, lymphoid tissue or PBMCs and included both fresh frozen and FFPE samples. Freshly collected tissue samples were disaggregated by GentleMACs digestion (Miltenyi Biotec) before cryopreservation with FBS/10% DMSO and storage in liquid nitrogen or directly stored as whole tissue blocks in liquid nitrogen. Blood and BM specimens were isolated by Ficoll/Hypaque density gradient centrifugation prior to cryopreservation with FBS/10% DMSO and storage in liquid nitrogen. For viably frozen samples of low purity (<30% tumor), RS cells were isolated by fluorescence activated cell sorting (FACS) (Aria II instrument, Becton Dickinson) based on CD5+ and CD19+ co-expression on cells with increased forward

scatter (FSC) (Biolegend, CD5-FITC cat#364022, CD19-PE-Cy7 cat#302216). For FFPE specimens, samples from each submitting center were reviewed for >50% purity prior to sequencing.

CLL samples. CLL samples were obtained from PBMCs. Samples with higher CLL purity (WBC >25 x 10^3/microliter or ALC >20 x 10^3/microliter) were processed without CD19 selection, and PBMCs were isolated by Ficoll/Hypaque density gradient centrifugation and then cryopreserved with FBS/10% DMSO and stored in liquid nitrogen until the time of analysis. Samples with WBC <25,000/uL or ALC <20,000/uL underwent CD19 selection (RosetteSep Human B-cell enrichment, Stem Cell Technologies) or as previously described[5] or FACS sorting to enrich for CD5+CD19+ populations.

Germline samples. Sources of non-tumor germline DNA included saliva (Oragene Discover [ORG500 or ORG600] kit, DNA Genotek), remission bone marrow [5] or *in vitro* expanded T cells. For the latter, CD19- CD4+ or CD19-CD3+ cells were collected by FACS (Aria II, BD; Biolegend, cat#300330, cat#300506, #363006). The cells were plated and expanded *in vitro* in RPMI (Gibco) containing phytohemagglutinin (PHA) (1.5:100), IL-7 (20 ng/mL), IL-2 (100 U/mL), 10% human serum and beta-2-mercaptoethanol (1/1000).

Genomic DNA sequencing

Whole-exome sequencing (WES)

A total of 143 samples were processed and sequenced at the Broad Institute (Cambridge, MA). For these fresh blood and bone marrow samples and cryopreserved suspension cells, genomic DNA and RNA was extracted per manufacturer's recommendations (Qiagen). DNA was quantified in triplicate using a standardized PicoGreen® dsDNA Quantitation Reagent (Invitrogen) assay. The quality control identification check was performed using fingerprint genotyping of 95 common SNPs by Fluidigm Genotyping (Fluidigm, San Francisco, CA). Library construction from double-stranded DNA was performed using the KAPA Library Prep kit, with palindromic forked adapters from Integrated DNA Technologies. Libraries were pooled prior to hybridization. Hybridization and capture were performed using the relevant components of Illumina's Rapid Capture Enrichment Kit, with a 37Mb target. All library construction, hybridization and capture steps were automated on the Agilent Bravo liquid handling system. After post-capture enrichment, library pools were denatured using 0.1N NaOH on the Hamilton Starlet. Cluster amplification of DNA libraries was performed according to the manufacturer's protocol (Illumina) using HiSeq 4000 exclusion amplification chemistry and HiSeq 4000 flowcells. Flowcells were sequenced utilizing Sequencing-by-Synthesis chemistry for HiSeq 4000 flowcells. The flowcells were then analyzed using RTA v.2.7.3 or later. Each pool of whole-exome libraries was sequenced on paired 76 cycle runs with two 8 cycle index reads across the number of lanes needed to meet coverage for all libraries in the pool. Output from Illumina software was processed by the Picard data-processing pipeline to yield BAM files containing demultiplexed,

aggregated aligned reads. Standard quality control metrics, including error rates, percentage-passing filter reads, and total Gb produced, were used to characterize process performance before downstream analysis.

Twenty-seven samples were processed and sequenced at University of Ulm, Germany. Exome Enrichment was performed through biotinylated RNA oligomer libraries, which are part of the SureSelectXT Human All Exon V5 capture library. The preparation workflow with the SureSelectXT reagent kit included DNA Fragmentation via Covaris supersonic shearing, end repair, ligation, library hybridization, indexing, QPCR based quantification and multiplexing (per protocol version 1.7). Multiple quality controls via Agilent Bioanalyzer were implemented into the process. Libraries were amplified to produce clonal clusters and sequenced using massively parallel sequencing on the Illumina HiSeq2000 Sequencing System. Eleven samples were processed (SureSelect QXT Agilent kit) and sequenced on a HiSeq 1000 instrument at the University of Nancy, France.

A subset of our WES data had reduced coverage in the GC-rich region of *NOTCH1*. For these, targeted deep sequencing of the *NOTCH1* 3' UTR was performed, as previously described[26].

Whole-genome sequencing (WGS)

Preparation of libraries for cluster amplification and sequencing (PCR-Free). 350ng of genomic DNA in 50µL of solution was processed by fragmentation through acoustic shearing (Covaris focused ultrasonicator), targeting 385bp fragments, and additional size selection was performed using a SPRI 80 cleanup. Library preparation (Hyper Prep without amplification module, KAPA Biosystems, #KK8505) was performed as above for WES. Libraries were then quantified using quantitative PCR (KAPA Biosystems) with probes specific to the ends of the adapters, normalized to 1.7nM, and then pooled into 24-plexes.

Preparation of libraries for cluster amplification and sequencing (PCR-Plus). An aliquot of genomic DNA (100ng in 50µL) was used as the input into DNA fragmentation. Shearing was performed as described above in the PCR-free procedure. Library preparation was performed using a commercially available kit provided by KAPA Biosystems (KAPA Hyper Prep with Library Amplification Primer Mix, product KK8504), and with palindromic forked adapters using unique 8-base index sequences embedded within the adapter (Roche). The libraries were then amplified by 10 cycles of PCR. Following sample preparation, libraries were quantified using quantitative PCR (KAPA Biosystems) with probes specific to the ends of the adapters. This assay was automated using Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 2.2nM and pooled into 24-plexes.

Cluster amplification and sequencing (NovaSeq 6000). Sample pools were combined with NovaSeq Cluster Amp Reagents DPX1, DPX2 and DPX3 and loaded into single lanes of a NovaSeq 6000 S4 flowcell cell using

the Hamilton Starlet Liquid Handling system. Cluster amplification and sequencing occurred on NovaSeq 6000 Instruments utilizing sequencing-by-synthesis kits to produce 151bp paired-end reads. Output from Illumina software was processed by the Picard data-processing pipeline to yield CRAM or BAM files containing demultiplexed, aggregated aligned reads. All sample information tracking was performed by automated LIMS messaging.

Circulating DNA sequencing. Whole blood was collected by routine phlebotomy. Plasma was separated within 1-4 days of collection through density centrifugation and stored at -80°C until DNA extraction (QIAsymphony DSP Circulating DNA Kit, QIAGEN), which was performed according to the manufacturer's instructions. Library preparation was performed (KAPA HyperPrep Kit with Library Amplification, KAPA Biosystems) using duplex UMI adapters (IDT), starting with 2-3 cc of plasma. Samples were normalized and pooled using equivolume pooling, with up to 95 samples per pool. Cluster amplification was performed according to the manufacturer's protocol (Illumina) using Exclusion Amplification cluster chemistry and HiSeqX flowcells. Flowcells were sequenced on v2 Sequencing-by-Synthesis chemistry for HiSeqX flowcells. The flowcells were then analyzed using RTA v.2.7.3 or later. Each pool of ultra-low pass whole genome libraries was run on one lane using paired 151bp runs.

Analysis of UK WGS. BAM files were obtained from prior analysis and realigned them to the Broad Institute's build of hg19 (known as b37: https://gatk.broadinstitute.org/hc/en-us/articles/360035890711-GRCh37-hg19-b37-humanG1Kv37-Human-Reference-Discrepancies)[43]. Out of the 17 sample trios obtained from the UK group, 14 samples completed WGS (3 failed due to data quality and realignment issues). The standard pipeline as previously described[43] was applied to these FFPE samples except for detection of sCNAs. Formalin damage results in extremely noisy read coverage profiles, confounding traditional copy number segmentation pipelines. To mitigate this, we applied a modified sCNA calling method that relies on segmentation of allelic imbalance at germline het sites (as opposed to segmentation of total coverage) as its primary signal. Although total coverage is extremely noisy, the fraction of reads supporting alternate versus reference alleles at heterozygous sites is undistorted, allowing for clean allelic imbalance segmentation. Within each segment of allelic imbalance, we binned total coverage on a megabase scale, which is coarse enough to average over formalin-induced coverage fluctuations, which typically manifest as sharp coverage spikes at the 10-100 kilobase scale. SV and phylogenetic analysis were completed for 12/14 samples.

Sequence data processing and analyses

WES/WGS alignment and quality control. Sequencing was conducted using standard methods (Supplementary Note)[16,33,44]. All DNA sequence data were processed through Broad Institute pipelines, such that data from multiple libraries and flow cell runs were combined into a single BAM file. This file contained reads aligned to

the human genome hg19 genome assembly (version b37, using BWA-MEM [version 0.7.15-r1140]) provided by the Picard and Genome Analysis Toolkit (GATK) developed at the Broad Institute[45], a process that involves marking duplicate reads, recalibrating base qualities and realigning around indels.

WES analysis. Sequences were analyzed by the Broad Institute's Cancer Genome Analysis WES Characterization Pipeline, in which aligned BAM files were inputted into a standard WES somatic variant-calling pipeline[44] and included *MuTect* for calling somatic single nucleotide variants (sSNVs), *Strelka2*[46] for calling small insertions and deletions (indels), *deTiN* for estimating tumor-in-normal (TiN) contamination, *ContEst* for estimating cross-patient contamination, *AllelicCapSeg* for calling allelic copy number variants, and *ABSOLUTE* for estimating tumor purity, ploidy, cancer cell fractions, and absolute allelic copy number. Artifactual variants were filtered out using a token panel-of-normals (PoN) filter, a blat filter, and an oxoG filter. For tumor samples without a matching normal control, a "no-normal" pipeline was used, as previously described[14]. Several FFPE samples exhibited lower DNA quality, resulting in noisier profiles with standard methods. For these samples, we applied an additional filtering technique of identifying the most correlated targets across a set of FFPE samples and performing tangent normalization[47] on samples that showed consistent behavior, thus excluding artifactual copy number targets.

WGS Analysis. WGS analysis was performed as previously described[43]. Due to the large amount of computational resources required to efficiently process cancer whole genomes, we ran these analysis pipelines on an elastic high performance computing (HPC) cluster on Google Cloud VMs, comprising thousands of CPU cores.

For structural variation (SV) identification, our pipeline integrates evidence from three SV detection algorithms (*Manta*[48], *SvABA*[49] and *dRanger*[50]) to generate a list of SV events with high confidence from WGS data. Subsequently, we applied *BreakPointer*[51] to pinpoint the exact breakpoint at base-level resolution. Breakpoint information was aggregated per sample to identify: (i) balanced translocations, defined as those with breakpoints on reverse strands within 1 kb of each other; (ii) inversions supported on both ends; (iii) complex events, based on the number of clustered events within 50 kb of each other. Breakpoints were annotated by intersection with our lists of CLL driver genes and significant sCNA regions, and with genes in the COSMIC Cancer Gene Census (v90)[52].

Identification of regions of kataegis and chromothripsis. In the WGS, kataegis regions were defined by genomic regions with at least 6 mutations within 2 standard deviations of the median chromosomal intermutational distance, as previously described[53]. For FFPE samples, to account for increased background sequencing artifacts, we considered only mutations with VAF > 0.15. Regions of chromothripsis were identified based on integrated evaluation of rainfall plots, allelic CN plots and SV calls.

Determining evolutionary relationships between RS and CLL and identifying RS specific genetic alterations. The *PhylogicNDT*[11,33] suite of tools was used to generate posterior distributions on cluster positions and mutation membership to calculate the ensemble of possible trees that support the phylogenetic relationship of detected cell populations. Through applying this tool across a set of CLL and RS samples per patient, the most likely tree was identified using probabilistic modeling and thus parent-child relationships among clones. Furthermore, all mutations were assigned to clones based on the match of mutational and clone CCF distribution. The RS clone was defined as a novel emerging clone first detected in the RS sample and absent in a preceding CLL sample. In the rare cases without close antecedent CLL samples, RS clones were conservatively identified from distal tree branches and through integrating available information on RS purity from pathology assessment. If a shared CLL historical clone was identified between samples, the RS was determined to be clonally related. If a shared clone was not identified across samples, the RS was determined to be clonally unrelated. In WGS *PhylogicNDT* results only, clusters with fewer than 20 mutations were removed along with clusters with low cancer cell fraction (CCF < 15%).

Mapping CN alterations to RS and CLL clones. Once clonal structure was established, subclonal sCNAs were mapped to clones using *PhylogicNDT CopyNumber2Tree*. Posterior probability was calculated based on CN profiles and allele-fraction distributions of heterozygous SNP sites across samples to assign likelihood of each event to belong to a clone with a particular CCF. The RS and CLL specific clonal events (both sSNVs and sCNAs) were thus identified.

Discovery of significantly mutated genes in RS and CLL clones. *MutSig2CV*[12] was run to identify driver genes from the filtered WES Mutation Annotation Format (MAF) file of both the RS history and RS clones. Divergent CLL clones were thus excluded, allowing for the identification of recurrent drivers contained within RS cells and clones. To further improve power to detect known variants, we ran *MutSig2CV* on a restricted set of hypotheses through utilizing list of CLL[16] and *de novo* DLBCL drivers[14,15]. For the validation cohort, *MutSig2CV* results were reported for new drivers that met significance and were present in at least one patient from the discovery cohort.

Identification of recurrent RS focal and arm-level copy number events. Somatic copy number alterations (sCNAs) were detected using the GATK4 CNV pipeline (http://github.com/gatk-workflows/gatk4-somatic-cnvs), comprised of the CalculateTargetCoverage, NormalizeSomaticReadCounts, and Circular Binary Segmentation (*CBS*) algorithms[54] for genome segmentation, with additional normalization for FFPE samples as described for WES analysis. To identify significantly amplified or deleted genomic regions in RS samples, *GISTIC2.0*[13] was applied, both before and after subtracting the CLL sample segment changes, to produce a list of candidate RS

sCNA driver regions. In parallel, the antecedent CLL sCNA drivers were examined through *GISTIC*. Significant events were reported with a Q value threshold of 0.1. A force-calling process was applied to identify the presence/absence of each sCNA driver event across tumor samples (https://github.com/getzlab/GISTIC2_postprocessing). This force calling process was then applied to all DLBCL recurrent sCNAs in RS and to identify RS recurrent sCNAs in DLBCL, both for frequency comparisons and to build a consensus matrix for clustering.

Signature analysis. Mutational signatures were determined using *SignatureAnalyzer* (https://github.com/getzlab/getzlab-SignatureAnalyzer). We furthermore compared the identified signatures with those in COSMIC (v3.2)[52] based on cosine similarity.

Immunogenetic analysis

To determine the clonal relationships between CLL and RS, we inferred the DNA sequences of immunoglobulin genes from WES/WGS data as previously described[16]. (**Supplementary Table 4**).

Consensus clustering of genetic alterations

Generation of gene sample matrix

All significantly mutated genes (MutSig2CV, $Q \leq 0.1$ and frequency ≥ 4 cases), significant regions of sCNAs (GISTIC2.0, $Q \leq 0.1$ and frequency ≥ 4 cases) were assembled into a gene-by-sample matrix (**Supplementary Table 7c**). The entries in the gene-by-sample matrix represent mutations and CN events as follows: non-synonymous mutations, 2; synonymous mutations, 1; no-mutation, 0; high-grade CN gain [CN $\geq$ 3.4 copies], 2; low-grade CN gain [3.4 copies $\geq$ CN $\geq$ 2.1 copies], 1; CN neutral, 0; low-grade CN loss [1.1 $\leq$ CN $\leq$1.9 copies], 1; high-grade CN loss [CN $\leq$ 1.1 copies], 2; WGD, 5.

NMF clustering

The 7 samples without genetic drivers in the gene-by-sample matrix were assigned to cluster C0. In addition, we identified marker genes differentially expressed across clusters by applying a Fisher's exact test (2×5 table with variant present or absent as one dimension and cluster as the second dimension) and corrected the p-values for multiple hypothesis testing using the BH-FDR procedure[55] (**Supplementary Table 7f**). Features with a q-value $\leq$0.1 were selected as cluster features and visualized as a color-coded heatmap. Features were annotated with their maximally positive associated cluster, determined by computing the 2x2 Fisher Exact test for all 5 clusters (2×2 table with variant present or absent as one dimension and within-cluster or outside-cluster the second dimension) (**Supplementary Table 7f**). To ensure robustness given the sample size of 97, we performed 100 subsampling iterations by randomly removing 8 patients in each iteration and calculated a sample-by-sample similarity matrix that reflects the frequency that each of two samples were clustered together in the 100 runs. Finally, we performed

UPGMA hierarchical clustering using 1-similarity as a distance metric. To define the final cluster membership, we cut the resulting dendrogram based on the modal number of clusters across the 100 subsampled consensus NMF clustering runs.

Mutual exclusivity/co-occurrence estimations.

For each gene of interest, the significance of the co-occurrence or mutual exclusivity for each pair of different events (mutations, amplification, deletion) that affects that gene was calculated using Fisher's exact test, and then false discovery rate was calculated using the Benjamini-Hochberg method.

Non-negative matrix factorization consensus clustering

To robustly identify clusters of tumors with shared genetic features, we applied a non-negative matrix consensus clustering algorithm[56] with slight modifications. Briefly, we passed the gene-by-sample matrix to the NMF consensus clustering algorithm (testing number of clusters k=2 to 10) and skipped the matrix normalization step so that the distance is calculated directly based on the values in the gene-by-sample matrix. The consensus NMF method was run as 20 iterations of NMF starting with different random seeds. The NMF consensus clustering algorithm provided the cluster membership of each sample, the cophenetic coefficient for k=2 to k=10 clusters and silhouette values for the optimal number of clusters, which was k=5 (**Supplementary Table 7d, Supplementary Note 2**).

Bulk RNA sequencing and data analyses

High-quality RNA from CLL/RS pairs was extracted, as previously described[5]. Total RNA was quantified using the Quant-iT™ RiboGreen® RNA Assay Kit and normalized to 5 ng/µl. Following plating, 2 µL of ERCC controls (using a 1:1000 dilution) were spiked into each sample. An aliquot of 200ng for each sample was transferred into library preparation which uses an automated variant of the Illumina TruSeq™ Stranded mRNA Sample Preparation Kit. This method preserves strand orientation of the RNA transcript. It uses oligo dT beads to select mRNA from the total RNA sample, followed by heat fragmentation and cDNA synthesis from the RNA template. The resultant 400bp cDNA then goes through dual-indexed library preparation: 'A' base addition, adapter ligation using P7 adapters, and PCR enrichment using P5 adapters. After enrichment, the libraries were quantified using Quant-iT PicoGreen (1:200 dilution). After normalizing samples to 5 ng/µL, the set was pooled and quantified using the KAPA Library Quantification Kit for Illumina Sequencing Platforms. The entire process was in a 96-well format and all pipetting is done by either Agilent Bravo or Hamilton Starlet. Pooled libraries were normalized to 2 nM and denatured using 0.1 N NaOH prior to sequencing. Flowcell cluster amplification and sequencing were performed according to the manufacturer's protocols using either the HiSeq 2000 or HiSeq 2500 instrument. Each run generated a 101bp paired-end with an eight-base index barcode read. Data was analyzed using the Broad Picard Pipeline, which includes de-multiplexing and data aggregation.

Bulk RNA-sequencing of validation cohort

RNA was extracted with Macherey Nagel RNA extraction kit (Macherey-Nagel, Düren, Germany). Total RNA-Seq libraries were generated from 500 ng of total RNA using TruSeq Stranded Total RNA LT Sample Prep Kit with Ribo-Zero Gold (Illumina, San Diego, CA), according to manufacturer's instructions. The final cDNA libraries were checked for quality and quantified using capillary electrophoresis prior to sequencing with HiSeq 4000 sequencing using

RNA-seq data analyses. RNA-seq reads were aligned to the human reference genome hg19 using STAR (v2.4.0.1)[57]. Lowly expressed genes with CPM < 1 in all samples were filtered out. Differentially expressed (DE) genes were assessed using *limma-voom*[58] in paired mode using sample read counts, with $|\log_2 FC|>1$ and adjusted p-value<0.25 as a cutoff. To ensure robustness of the analysis, for the 5 pairs of RS and CLL samples analyzed, DE genes were recalculated iteratively, each time leaving out one sample pair. Genes were rank ordered by their *t* statistic multiplied by the frequency they were found significant ($|\log_2 FC|>1$ and adjusted P <0.1) in the leave-one-out analysis. This was used as input for pre-ranked *GSEA* on HALLMARK pathways (1,000 permutations, weighted enrichment statistics, MsigDB v7.4)[59].

RNA clustering of RS samples and integration with genetic subtypes

Gene counts were pre-processed with *ComBat-seq* (v3.42.0)[60] to eliminate possible batch effects and one sample was removed as an outlier. TPMs were computed and genes were filtered out if TPM = 0 in at least one sample, median TPM over samples <= 0.5, or median TPM over samples > 1000. TPMs were then $\log_2$ transformed and top genes by variance (z-score of variance > 1) were z-score transformed for downstream analysis. Consensus clustering[61] using the hierarchical clustering (complete linkage) with spearman distance was used to identify the optimal number of clusters (observed as 5 RNA subtypes), and the resulting consensus matrix was transformed into a distance matrix for hierarchical clustering (complete linkage). The agreement between RNA subtypes and genomically-identified clusters was determined by a Fisher's exact test. Supervised analysis for differentially expressed genes for each genomically-identified cluster was performed using *limma-voom(v3.50.3)*[58] as a one-vs-other comparison. Pathway analysis of each genomically-identified cluster was performed using Preranked-GSEA[59] with the MsigDB Hallmark (v7.4) genesets using the LIMMA t-statistic to rank order genes.

Single-cell RNA-sequencing and analysis

Sample preparation. For suspension samples with admixture of both CLL and RS cells, cells were thawed by drop-wise addition of warmed media (RPMI 10% FCS) and stained with antibodies (Biolegend CD5 FITC cat#364022, CD19 PE-Cy7 cat#302216, CD3 PB cat#300330 using 2-4 uL of each antibody per 100 uL test) and

a viability marker (Biolegend 7-AAD cat#420404 at 1:500 or Zombie Violet cat#423114 at 1:1000) before resuspension in PBS-0.04% BSA (Ultrapure NEB/Invitrogen). For Patients 19 and 41, viable CD5+ CD19+ cells were sorted into RS and CLL fractions by size based on the increased forward scatter (FSC) of RS cells (BD FACS Aria II). For Patients 43, 4 and 10, viable cells within the lymphocyte gate were sorted for analysis.

Sequencing. Five to ten thousand single-cells per specimen underwent transcriptome sequencing (Chromium Controller, 10X Genomics) according to the manufacturer's instructions, using either the 3' v2 kit (Patients 19 and 41) or the 5' v2 kit with BCR and TCR sequencing (Patients 43, 4 and 10). Each flow sorted fraction was run as a separate lane on the same chip. Libraries were pooled and sequenced on HiSeqX or NovoSeqS4 (Illumina).

Data processing of scRNA-seq libraries. Reads were processed and aligned to the Hg19 reference genome. All data were filtered using *Cell Ranger* (v2.1.1 for Patient 41; v2.0.0 for Patient 19, and v3.0.2 for Patients 4, 10 and 43). Background, or ambient, RNA was removed using *CellBender* with the exception of Patient 41. Data from each patient was analyzed using *Seurat* (v3.1.4)[62]. QC filtering was applied to remove cells with fewer than 500 UMIs, >50,000 UMIs or more than 10% mitochondrial reads. Potential doublets were detected using *DoubletFinder* (v2.0.2)[63] using default pN and optimal pK for each sample and removed ahead of further analysis. For Patient 41, cell cycle regression was performed followed by data integration using standard methods[62].

Clustering was then performed to identify B cell clusters; these were further sub-clustered for additional analysis of malignant B cells using the presence of standard B cell markers (*CD19, CD20, IGLL5, CD79A, CD79B*) and the absence of T/NK/Myeloid markers (*CD3, CD4, CD8, CD56, CD14, CD16, CD33*). Clustree (v0.4.2) was used to identify stable clusters prior to downstream analysis. UMI/cell and genes/cell for each cluster were calculated with Seurat and the mean values across CLL and RS clusters were compared using a Wilcoxon test.

Inferred copy number across single cells (CNVsingle). We applied a novel tool *CNVsingle* (https://github.com/broadinstitute/CNVsingle) to the above processed Seurat objects. In brief, *CNVsingle* utilized normalization from matched PBMC derived B-cell profiles followed by Savitzky–Golay noise reduction. These profiles alongside the per cell allele counts across common heterozygous SNP sites identified in the samples were utilized by a Hidden Markov Model running in allele specific mode on subsets of cells. Thus *CNVsingle* provides allele-specific copy number profiles for all malignant cell clusters. As validation, different types of normal cells provided copy-neutral profiles. Single cell derived allelic CN across clusters was compared to WES CN profiles and found to be highly concordant. These profiles were then used to identify clusters as CLL, RS or transitional and cluster identities were used for subsequent differential expression testing.

Differential expression testing. Expression analysis was performed on CLL and RS clusters identified as those *CNVsingle* profiles that matched the CLL WES or RS WES samples. Clusters that showed intermediate sCNA profiles were considered potential transitional clusters. Furthermore, genes with non-zero read counts in less than 20 cells were removed. Gene counts for a given cluster were obtained by summing the counts across all the cells in each respective cluster. DE genes were assessed using *limma-voom(v3.50.3)* in paired mode. The ranked gene list sort the *t* statistic from the DE analysis for RS clusters in each scRNAseq sample was submitted to pre-ranked *GSEA* to analyze the HALLMARK pathways (1,000 permutations, weighted enrichment statistics, MsigDB v7.4)[59].

Velocity Analysis. RNA inference of directional trajectories was performed with *scVelo* (v 0.2.4 – with fit_connected_states=False) using the dynamical model on the normalized data. Spliced and unspliced reads were computed via *velocity* (v 0.17.17)[36]. The result of the model was then used to estimate gene latency, which represents the cell's internal clock and is based only on its transcriptional dynamics. The root key parameter has been computed via the *CellRank* (v 1.5.0) library.

Random Forest (RF) Analysis. We used an RF approach to differentiate between CLL and RS cells in the single cell data. Data was preprocessed using the same cell/gene filtering as in the DE analysis. To reduce the impact of cell size differences between CLL and RS, we performed a z-score normalization per cell. We trained an RF (*n_estimators* = 1000, sklearn v1.0.1) on samples LNs from Pts 10 and 43 and predicted on cells from Pts 41 (LN or peripheral blood [PB] samples) and 18 (bone marrow) whose cell labels were determined by FACS sorting described earlier. We ran the RF 20 times and obtained a mean ± σ of 0.92 ± 0.01 when looking at only the LN sample in the test set to avoid any potential microenvironment differences. When we included the PB sample in the test set, the F_1 only slightly decreased to 0.86 ± 0.11; while also adding in Pt 18 yielded an F_1 of 0.66 ± 0.01. The decrease in F_1 score is possibly due to differences in tissues of origin and sequencing platforms. The top discriminative features are defined as genes whose gini impurity scores were at least 3σ above the mean.

Clinical endpoint analysis and statistical analysis

Data analyses were carried out using GraphPad Prism version 9 and R software version 4. To compare RS drivers identified with previously reported CLL (n=1063)[16] and DLBCL (n=304)[14] and (n=574)[15] datasets, a 2-sided exact binomial test was performed with Benjamini-Hochberg multiple hypothesis testing correction. To obtain the frequency of RS events in DLBCL cohorts prior to this comparison, we called RS sCNAs in 304 DLBCLs[14] and the 443 primary DLBCLs[15] for which purity was >20%. Event frequencies were compared when an event was detected in both sample sets. RS and CLL drivers co-occurrences were represented by using a Sankey diagram. Significance was evaluated by calculating the probability for acquiring each of the RS drivers considering the acquisition of a given driver in CLL using Fisher's exact test. To evaluate how often a given driver initially occurs

during the RS stage in the subset of related RS, we performed the McNemar test. Differences were considered significant when a *P* value adjusted for multiplicity of testing was < 0.05. Overall survival (OS) was defined as the interval between date of transformation and death or censored at last follow-up. Survival data were calculated using the method of Kaplan-Meier and curves were compared by log-rank testing.

Cell free DNA (cfDNA) analysis

After sequencing, plasma cfDNA samples were processed and analyzed as reported[37]. To detect RS-specific changes, we undertook the following steps. First, we analyzed delta copy number changes between segments, assigning a positive chromothripsis score when 3 consecutive 1 Mb segments had CN delta ≥ 0.1, suggested locally fractured genome. Second to assess Richter-specific aneuploidy, we evaluated the fraction of genome in non-copy-neutral state by fraction genome altered (FGA), defining a region as altered if the segment had an event as detected by iCHOR analysis and a CN change ≥ 0.1 (to filter out low confidence CN changes) and comparing to a matched CLL sample when available. Third, we assigned WGD to samples where copy-number events had allelic ratios (corrected for iCHOR estimated purities) corresponding to two levels of allele deletions (i.e. 2/0, 1/1 and 2/1 copy-number states) as measured from the main balanced copy-number level (2/2). Lastly, we performed WES on cfDNA, which we then examined for RS clonal alterations detected in bulk through phylogenetic reconstruction.

Data Deposition

WES, RNA-seq, WGS, and scRNAseq data will be deposited in dbgap (accession number phs002458.v1.p1) at the time of publication.

DATA AVAILABILITY STATEMENT

WES, RNA-seq, WGS, scRNAseq and cfDNA data are available at dbgap (https://www.ncbi.nlm.nih.gov/gap/) using accession number phs002458.v2. RNA-seq data from validation cohort can be accessed by registering for an EGA account (https://ega-archive.org/) and contacting the Data Access Committee under study EGAS00001005495 and accession number EGAD00001007922 (https://ega-archive.org/datasets/EGAD00001007922).

The following figures have associated raw data : Fig. 2 b-h, Fig 4 a,f, Fig. 5 a,b,d,e, Fig 6h, Extended Data 4 a-x, Extended Data 5 a-c, Extended Data Fig. 6a, Extended Data Fig. 8, Extended Data Fig. 9, Extended Data Fig. 10.

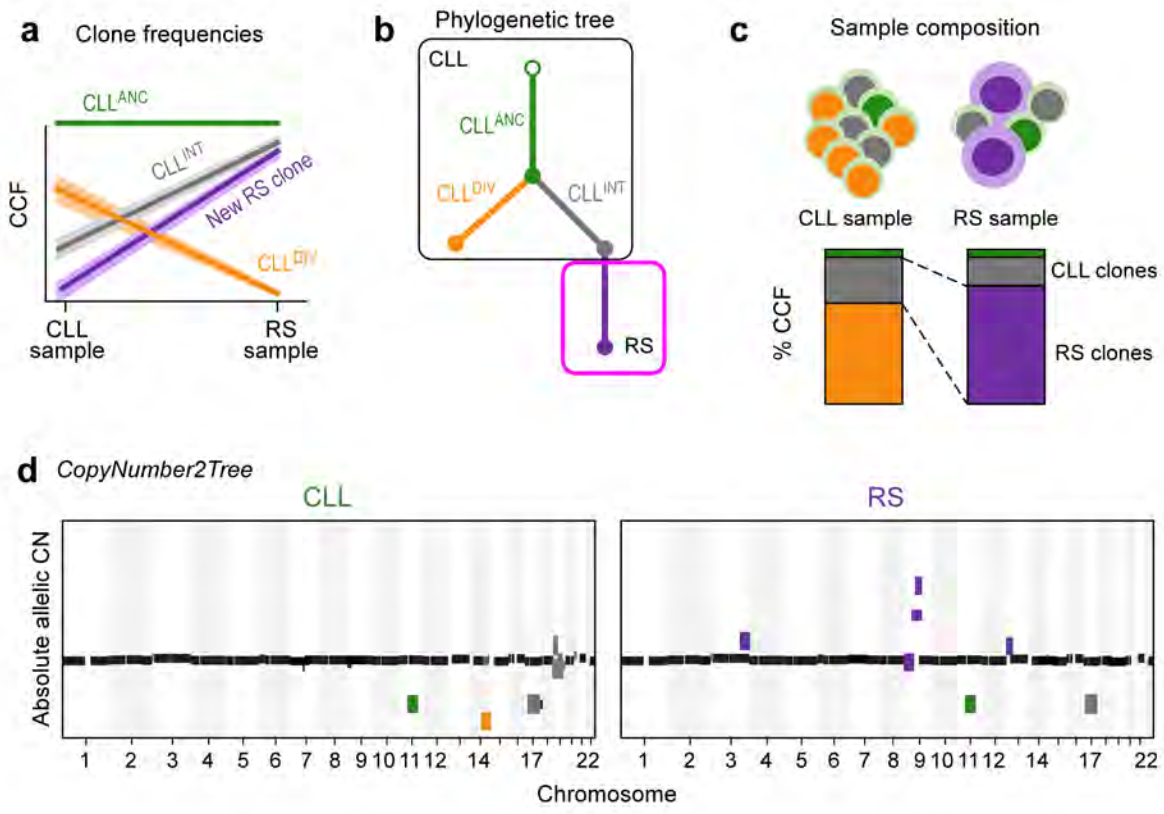
CODE AVAILABILITY STATEMENT

Code is available for CNVsingle (https://github.com/broadinstitute/CNVsingle) and *PhylogicNDT CopyNumber2Tree* (https://github.com/broadinstitute/PhylogicNDT).

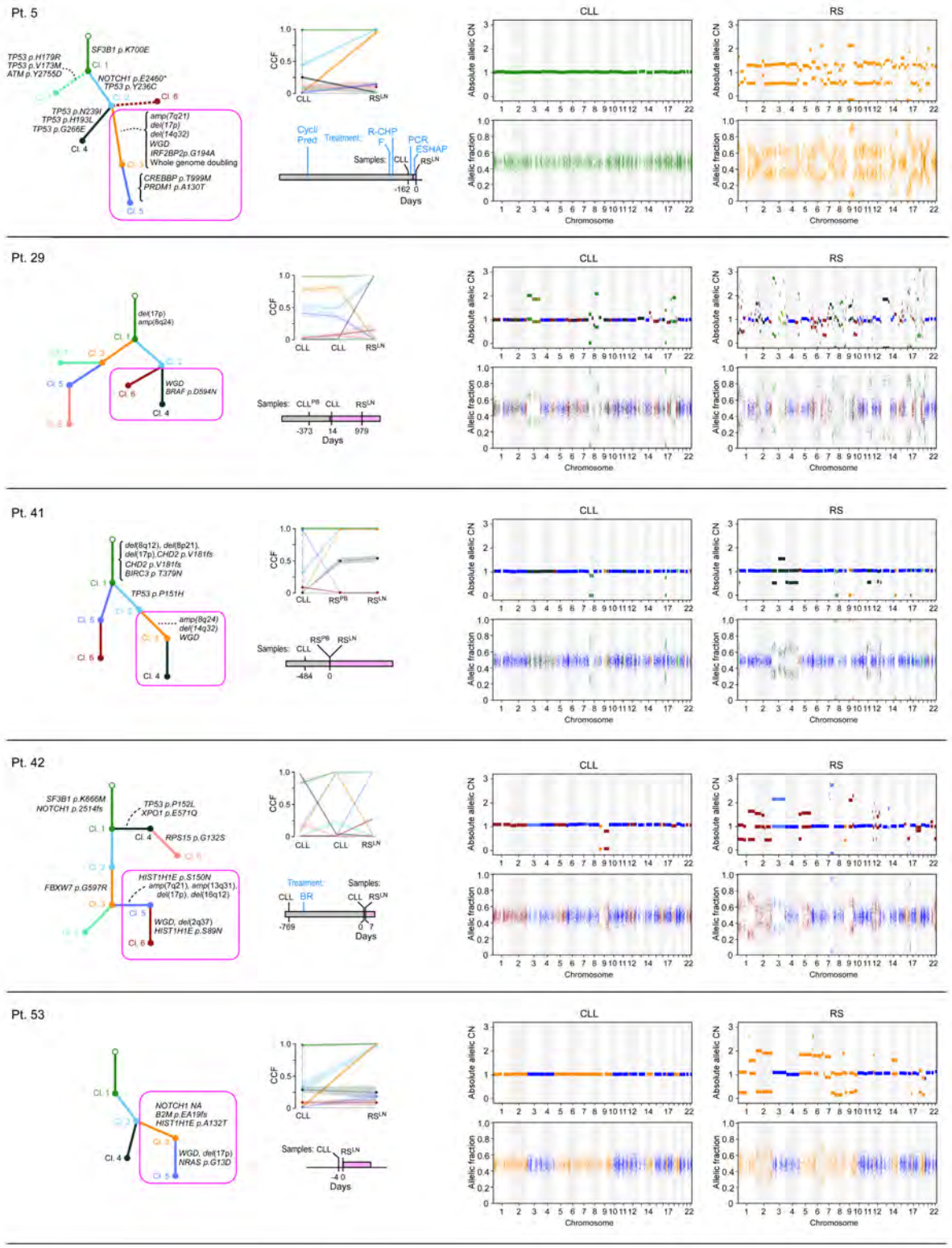
METHODS ONLY REFERENCES

43. Parikh, A.R., *et al.* Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nat Med* **25**, 1415-1421 (2019).
44. McKenna, A., *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
45. Kim, S., *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594 (2018).
46. Tabak, B., *et al.* The Tangent copy-number inference pipeline for cancer genome analyses. *bioRxiv*, 566505 (2019).
47. Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).
48. Chen, X., *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016).
49. Wala, J.A., *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* **28**, 581-591 (2018).
50. Bass, A.J., *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* **43**, 964-968 (2011).
51. Drier, Y., *et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res* **23**, 228-235 (2013).
52. Sondka, Z., *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705 (2018).
53. Alexandrov, L.B., *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
54. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572 (2004).
55. Brunet, J.P., Tamayo, P., Golub, T.R. & Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* **101**, 4164-4169 (2004).
56. Dobin, A., *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
57. Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).
58. Subramanian, A., *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
59. Zhang, Y., Parmigiani, G. & Johnson, W.E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* **2**, lqaa078 (2020).
60. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* **52**, 91-118 (2003).
61. Stuart, T., *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).
62. McGinnis, C.S., Murrow, L.M. & Gartner, Z.J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e324 (2019).

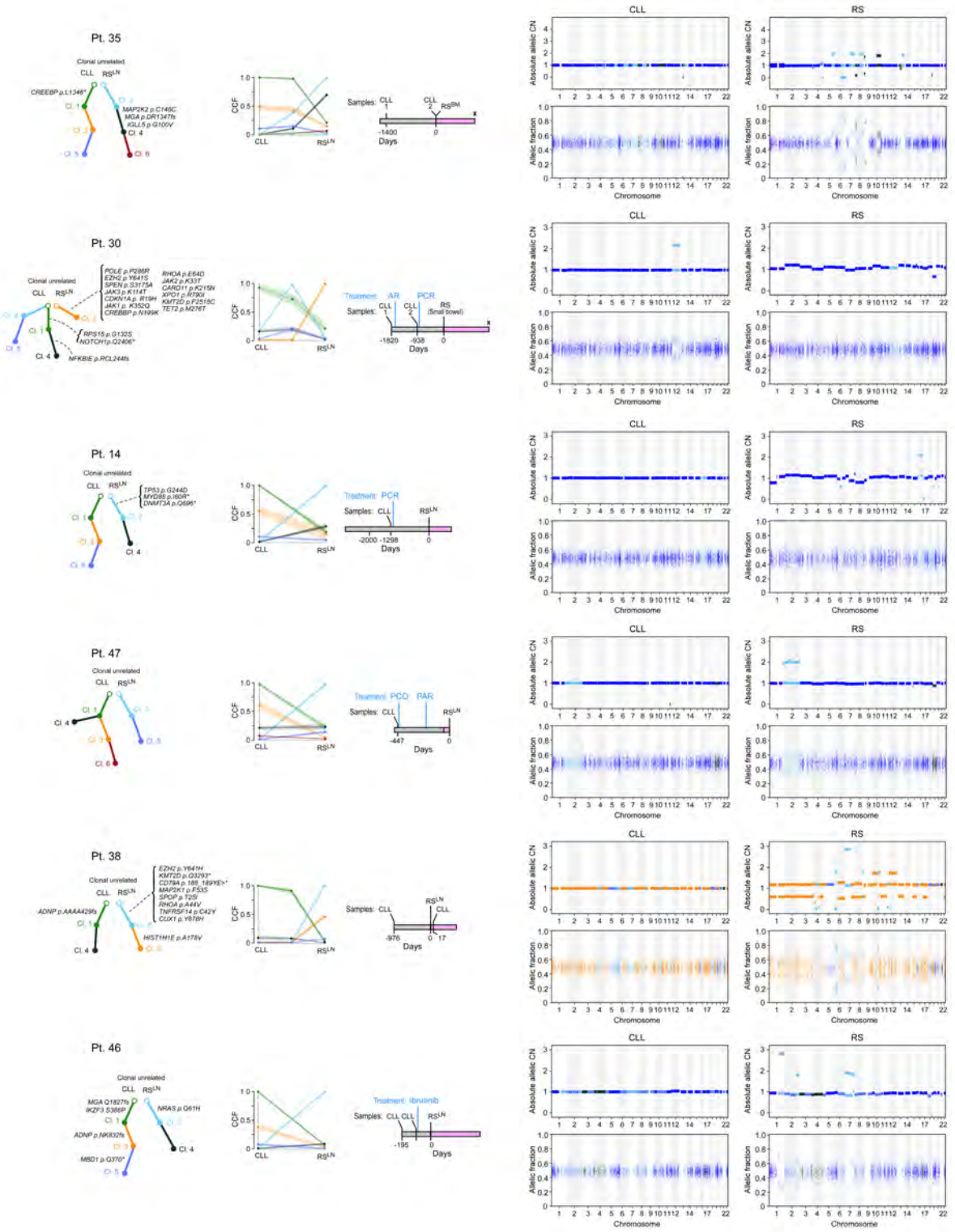
Extended Data 1



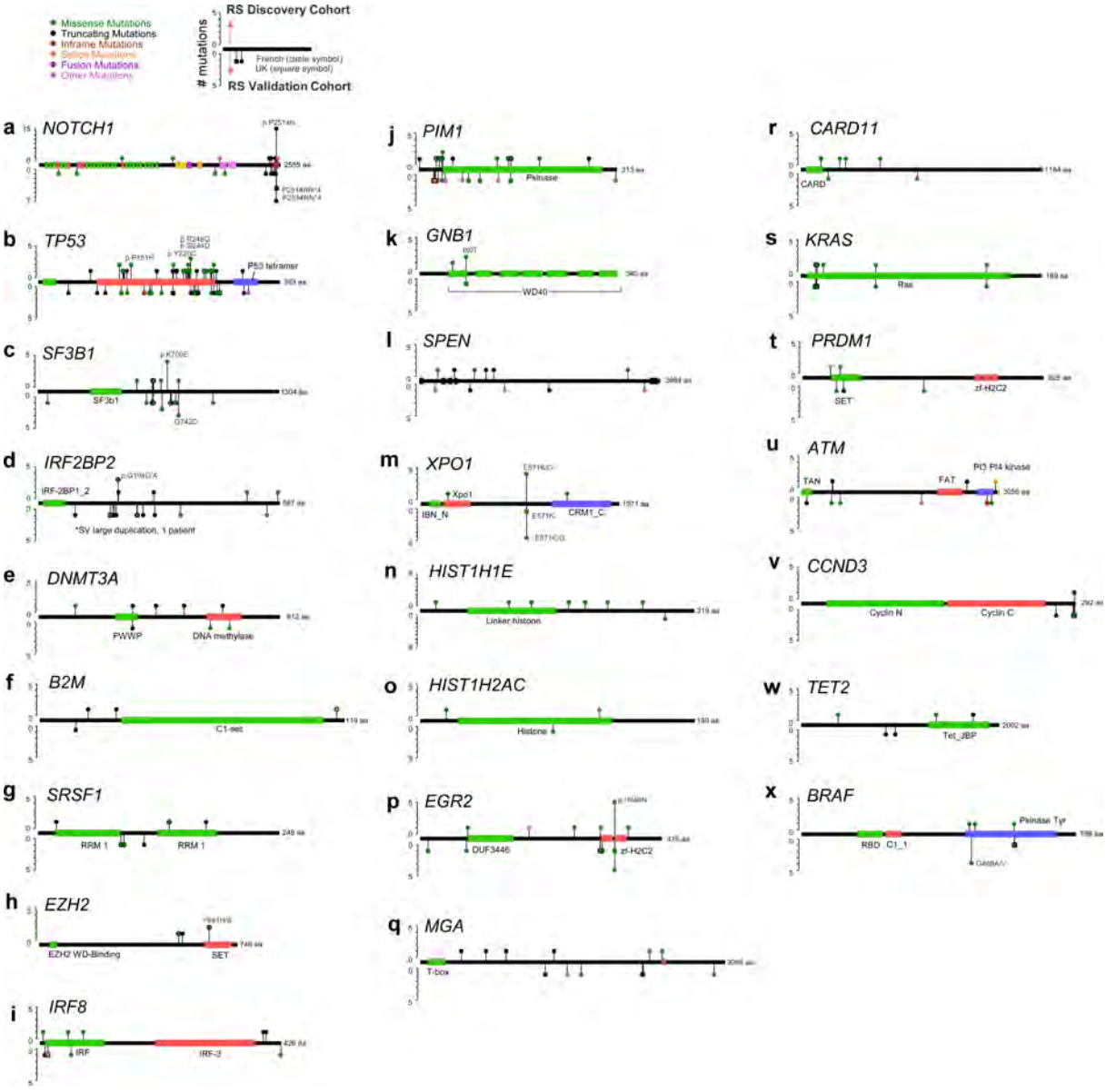
Extended Data 2



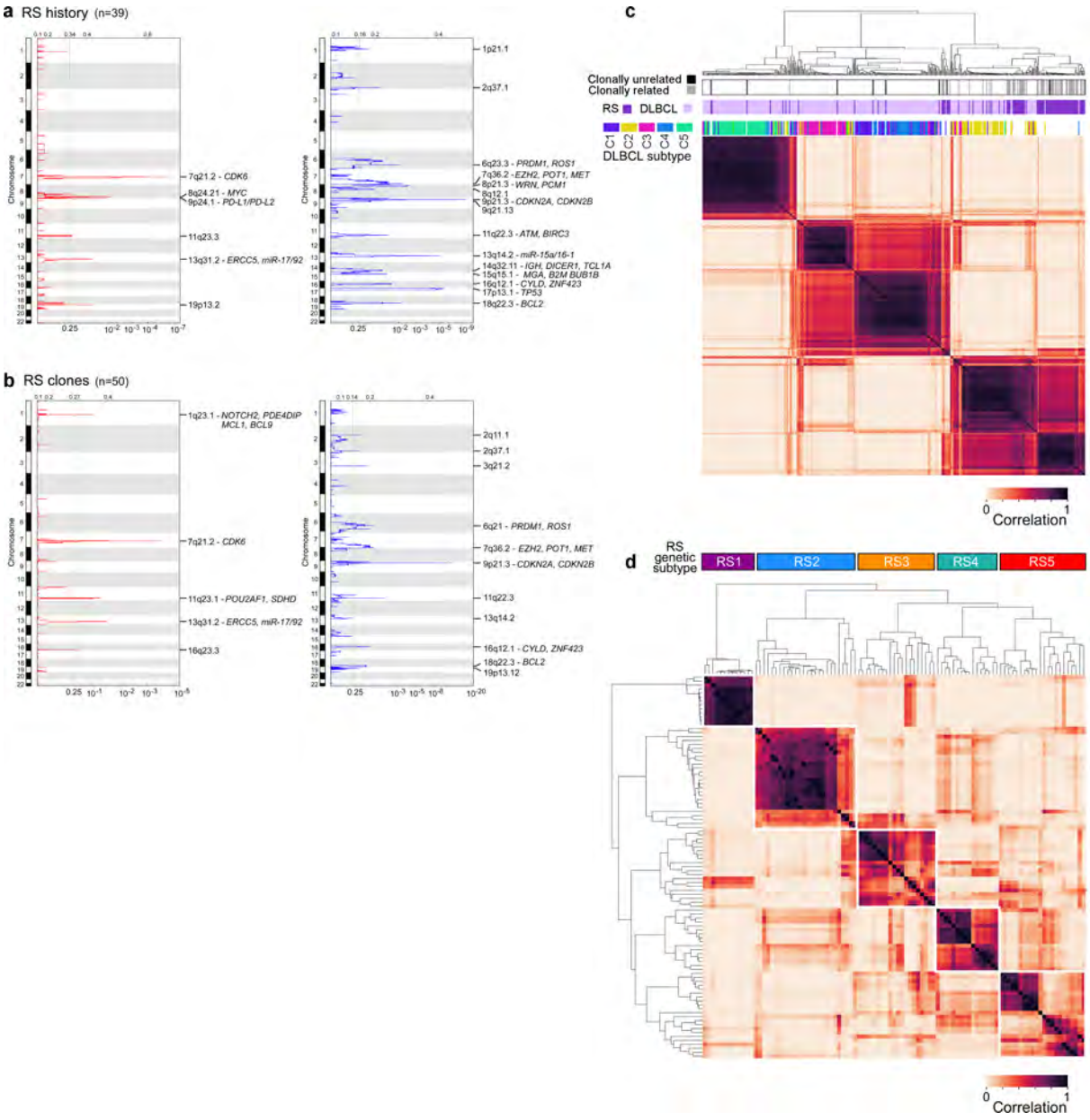
Extended Data 3



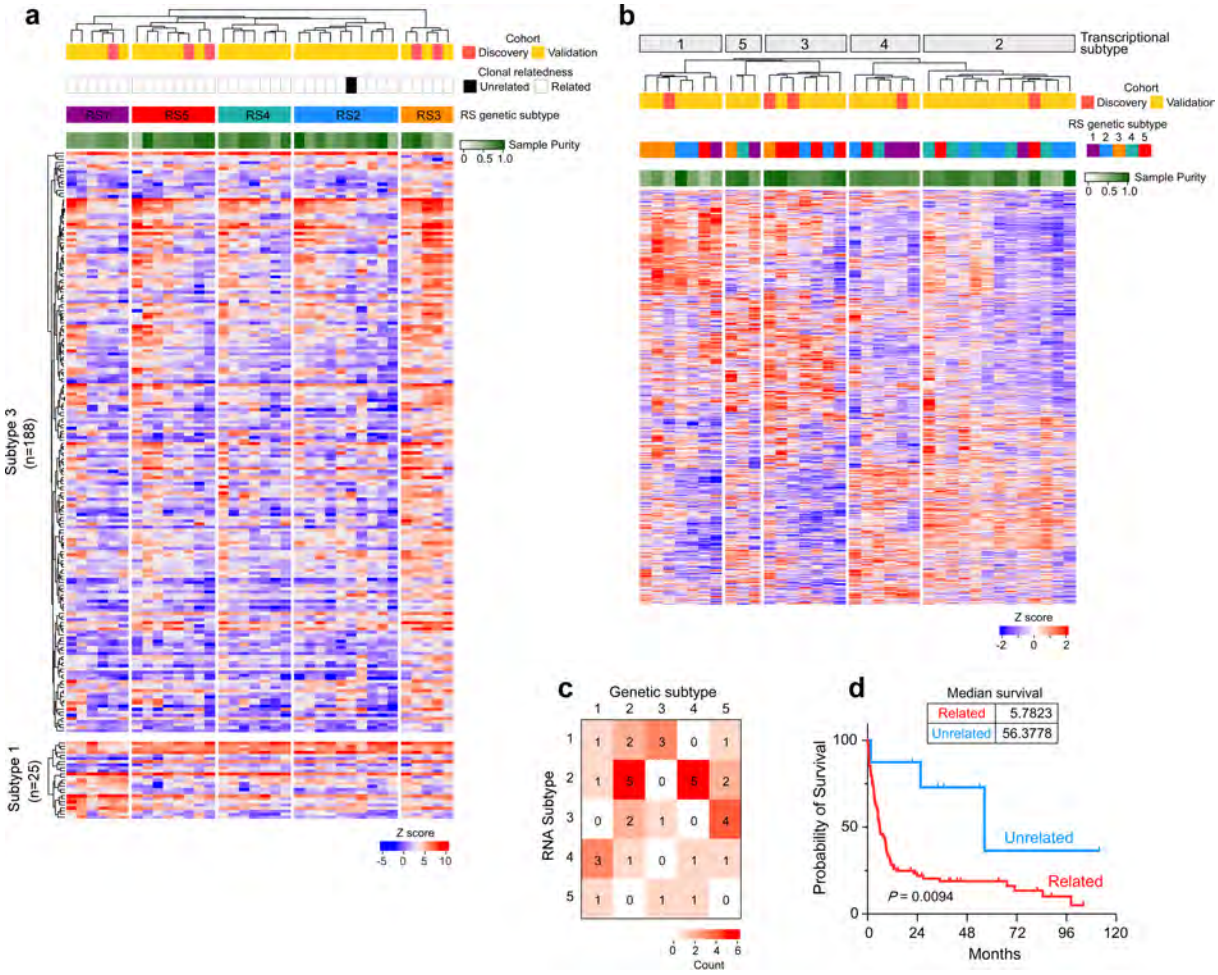
Extended Data 4



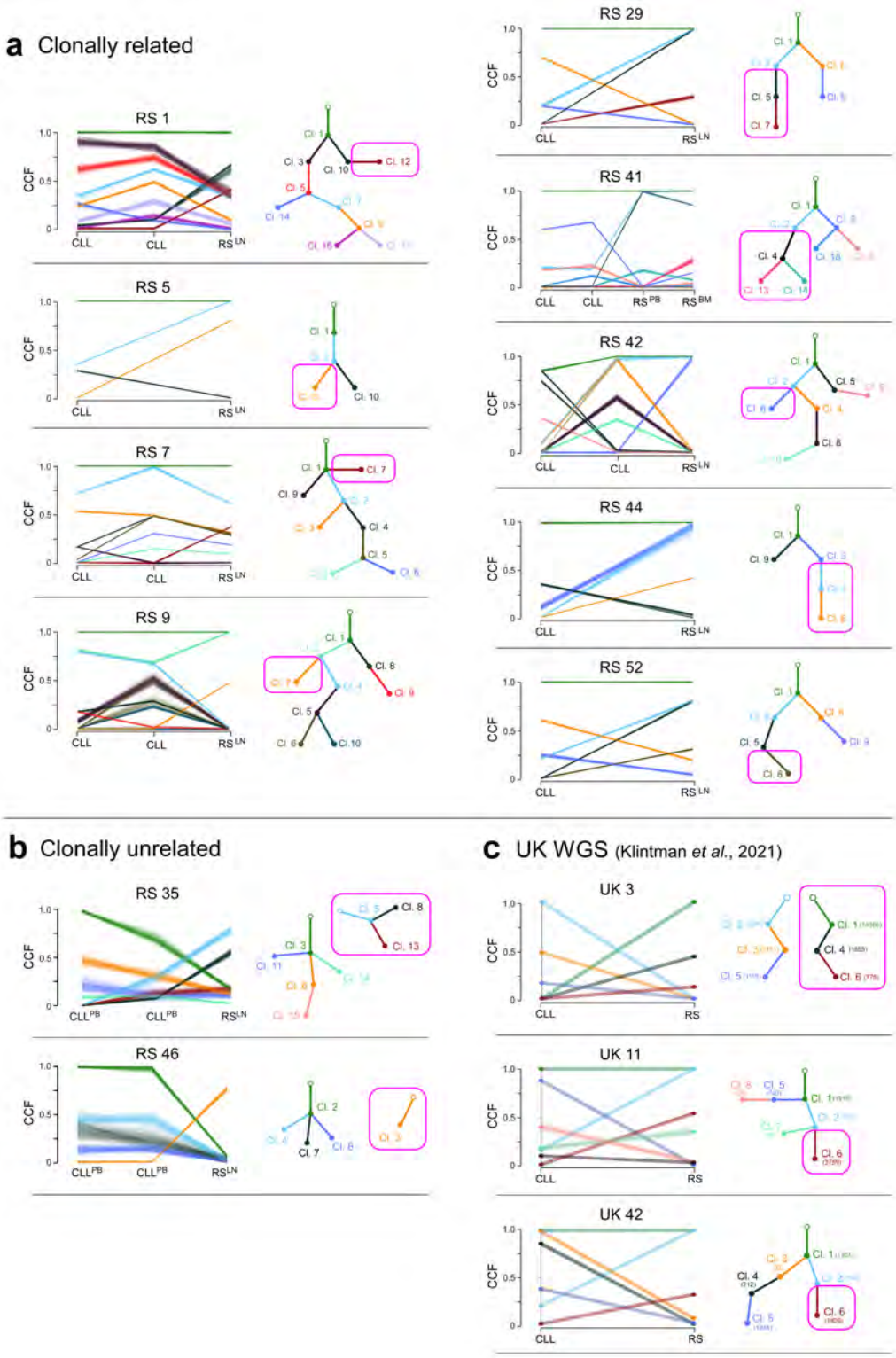
Extended Data 5



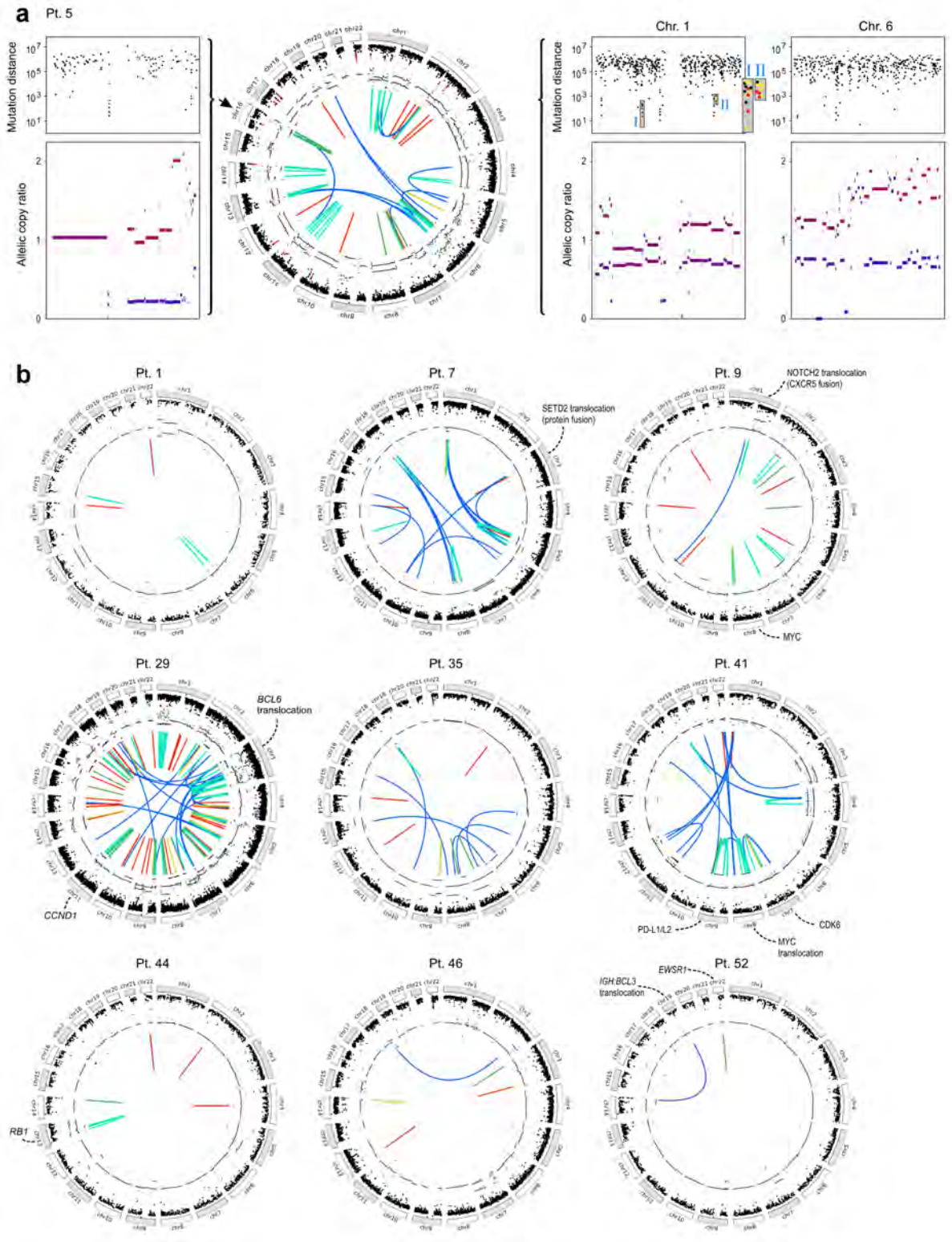
Extended Data 6



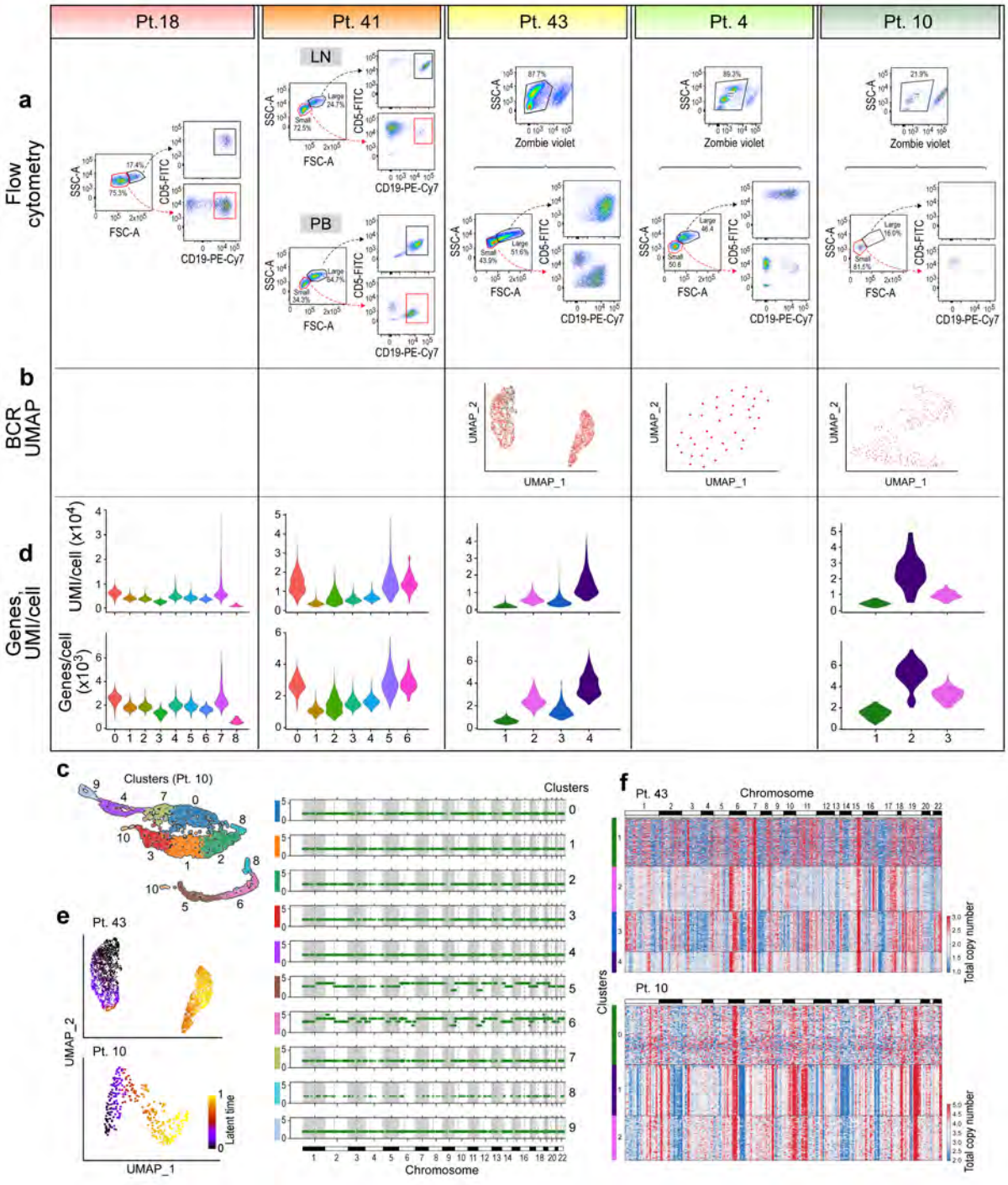
Extended Data 7



Extended Data 8



Extended Data 9



Extended Data 10